Федеральный исследовательский центр
"Информатика и управление"Российской академии наук

На правах рукописи
УДК 519.254

Адуенко Александр Александрович

ВЫБОР МУЛЬТИМОДЕЛЕЙ В ЗАДАЧАХ КЛАССИФИКАЦИИ

01.01.09 — Математическая кибернетика и дискретная математика

Диссертация на соискание ученой степени
кандидата физико-математических наук

Научный руководитель:
д.ф.-м.н., В. В. Стрижов

Москва — 2017

Оглавление

Введение

Актуальность темы.

В данной работе рассматривается задача построения мультимоделей для решения задач двухклассовой классификации [9,23,24,49,78–81]. Задача двухклассовой классификации является базовой в машинном обучении, а задачи многоклассовой классификации могут быть эффективно сведены к решению одной или нескольких задач двухклассовой классификации [95–98]. Задача двухклассовой классификации возникает во многих практических задачах из разных областей. Так задачей двухклассовой классификации является задача определения наличия заболевания у пациента по набору его анализов [92, 93], задача анализа текстов для получения настроения сообщений [94], задача кредитного скоринга [33, 34, 83]. Так задача кредитного скоринга [25, 33] состоит в определении того, будет ли допущен заемщиком неплатеж по кредиту по ответам заемщика на кредитную анкету, включающую информацию о его доходах, семейном положении, собственности, образовании и т.д. [**?**, 43]. Задача становится все более актуальной вместе с распространением и широким использованием разного рода кредитов, особенно потребительских. Так как использование экспертов при приеме решения о выдаче кредитов затратно и не всегда возможно, как, например, в случае с равноправным кредитованием [102], решение о выдаче кредита и ставке принимается с помощью некоторой скоринговой системы [33]. Под скоринговой системой подразумевается автоматизированная система, которая по предоставленным заемщиком данным оценивает вероятность дефолта по кредиту [8, 33]. Отметим, что логистическая регрессия, позволяющая получить интерпретируемую модель, содержащую информацию о важности каждого из признаков, широко используется как метод двухклассовой классификации во многих областях [8, 34, 83, 92, 93], а в области кредитного скоринга является стандартом [8, 33, 83].

Однако одиночная логистическая модель, как и любая обобенно-линейная модель, не позволяет описать неоднородности в данных, поскольку веса признаков одинаковы для всех объектов в выборке, а обучение состоит в определении этих весов [21, 33, 34]. Например, данные могут иметь кластерную структуру и важность признаков, а, значит, и их оптимальный вес, могут зависеть от кластера данных. Для решения проблемы неоднородности данных существует несколько подходов, позволяющих строить композиции классификаторов. В первом подходе каждый объект жестко относится к одной из моделей, причем разбиение признакового пространства на области действия моделей может производиться путем кластеризации [25–27] или разбиения на группы по значениям признака или группы признаков. При этом разбиение на группы по значениям признака можно реализовать путем перекодировки соответствующего признака в рамках одиночной модели [68]. Подход с жестким разбиением объектов по кластерам приводит к построению многоуровневых моделей [49, 78], в которых признаковое пространство разбито на непересекающиеся подмноже-

ства и в каждом из них действует одиночная модель. Такой подход является стандартным в кредитном скоринге [25–28], поскольку позволяет сохранить интерпретируемость построенной мультимодели, если разбиение признакового пространства на части осмысленно и модели, входящие в многоуровневую модель, различимы, и одновременно учесть неоднородности в данных. Вторым подходом является мягкая кластеризация, в которой для каждого объекта есть вероятность отнесения к каждой из моделей, зависящая [9] или не зависящая от объекта [6, 11, 23, 24, 30, 31, 79, 80]. Так бэггинг (англ. bootstrap aggregation) [6, 11] состоит в построении композиции простого голосования одиночных моделей, в бустинге [30, 31] строится композиция путем последовательного добавления классификаторов, в смеси моделей для каждого объекта есть некоторая фиксированая вероятность принадлежать каждой из моделей [79, 80, 82], а в смеси экспертов эти вероятности также зависят от объектов [9]. Сравнение подходов с жестким и мягким разбиением объектов по моделям и кластерам приведено в [75, 99–101].

Предлагаемые методы позволяют учесть неоднородность данных путем построения более сложной модели (мультимодели), содержащей несколько одиночных моделей, однако возникает проблема ее интерпретируемости. Так при жестком разбиении объектов между моделями не происходит учета близости моделей, построенных для разных групп объектов, а потому модели, построенные на разных подвыборках могут совпадать или быть близки. Композиции, построенные с помощью бэггинга, бустинга или смеси моделей также могут содержать в себе множество одинаковых моделей, наличие которых сложно интерпретировать. Более того, даже данные, не имеющие неоднородностей, вместо описания в виде одиночной модели получают описание в виде сложной композиции. Ранее было предложено несколько методов для прореживания таких моделей [10, 12, 13, 15]. В работе [10] предлагается несколько эвристик для прореживания ансамбля моделей из бэггинга. В работах [19, 20] для выбора подмножества моделей в бэггинге предалагается использовать генетические алгоритмы. Другим методом является кластеризация моделей, а затем выбор единственного представителя для каждого кластера [17, 18]. В работах [14, 16] предлагают жадную стратегию постепенного наращивания числа классификаторов в бэггинге с выбором на каждом шаге классификатора, наиболее приближающего композицию к целевому вектору. Для контроля количества моделей в смеси моделей используют введение априорного поощряющего разреженность распределения на веса моделей в смеси [21] и поиск структуры смеси путем максимизации обоснованности [9, 89, 90].

Предлагаемые методы прореживания композиций тем не менее прямо не учитывают близости между моделями в мультимодели, а потому мультимодель по-прежнему может содержать близкие модели, что ведет к неинтерпретируемости и ухудшению качества классификации, так как, например, для малого кластера данных может быть построена отдельная неинформативная модель, оценки параметры которой обладают большой дисперсией. Для получения ста-

тистически различимых моделей в мультимодели можно использовать внешнюю процедуру прореживания, основанную на статистическом сравнении моделей путем подсчета расстояний между апостериорными распределениями параметров для разных моделей, например, с помощью дивергенций Брегмана или f-дивергенций [50, 51, 53, 54]. В данной работе показано, что существующие меры сходства отличают неинформативную модель, построенную, например, для малого кластера данных, от другой модели даже в условиях генерации данных из одиночной модели, а потому не позволяют построить адекватную мультимодель. Для решения этой проблемы введено понятие неинформативности и малоинформативности распределений и предложена функция сходства, позволяющая решать задачу статистического различения моделей. На основании полученных статистических свойств распределения введенной функции сходства в условиях истинности гипотезы о совпадении моделей предложен метод построения адекватных смесей моделей и многоуровневых моделей. Результаты вычислительного эксперимента на синтетических и реальных данных демонстрируют преимущества предлагаемого подхода в терминах качества классификации и интерпретируемости мультимоделей.

Еще одной проблемой является возможное наличие избыточных или мультикоррелированных признаков, что влияет не только на качество классификации построенной модели, но и на ее устойчивость [1, 7]. Для решения задачи отбора признаков используют генетические алгоритмы [35, 36, 40], методы последовательного добавления и удаления признаков [35, 36, 39], методы основанные на анализе матрицы взаимной информации [41], а также методы отбора признаков с помощью решения задачи квадратичной оптимизации [37]. В данной работе в рамках байесовского подхода используется принцип максимума обоснованности для определения структуры моделей [9, 21, 89, 90]. Отметим, что аналитическое выражение для обоснованности для логистической модели и для смеси логистических моделей получить не удается, а для аппроксимации обоснованности используется аппроксимация Лапласа [21] и вариационные нижние оценки [84, 85]. Отметим, однако, что кроме избыточных признаков, которые требуется удалить из рассмотрения, признаковое описание может содержать мультиколлинеаные признаки, например, зашумленные копии одного признака. Общим подходом является построение набора немультиколлинеарных признаков по исходному набору признаков путем оптимизации некоторого критерия качества [7, 38]. В данной работе показано, что подход, связанный с отбором признаков, является неоптимальным и для оптимального учета информации от мультиколлинеарных признаков предлагается их комбинировать. При этом показано, что метод максимума обоснованности не позволяет учесть зависимости между признаками, поскольку оценка максимума обоснованности для ковариационной матрицы весов признаков является асимптотически вырожденной.

Цели работы.

1. Исследовать свойства существующих расстояний между распределениями

и проверить их применимость для решения задачи сравнения моделей.

2. Предложить функцию сходства, которая позволяет решить задачу статистического сравнения моделей.
3. Исследовать статистические свойства распределения предлагаемой функции сходства.
4. Предложить метод учета мультиколлинеарности между признаками.
5. Разработать алгоритм построения адекватных оптимальных обученных мультимоделей и провести вычислительный эксперимент для проверки улучшения качества и интерпретируемости построенных мультимоделей.

Методы исследования. Для достижения поставленных целей используются методы построения мультимоделей для двухклассовой классификации [9, 49, 78–80]. Для обучения многоуровневых моделей используются методы выпуклой оптимизации [21, 87] для независимого нахождения параметров каждой из моделей, входящих в многоуровневую модель. Для обучения смесей моделей используется вариационный EM-алгоритм [29, 103, 104], а для учета многоэкстремальности используется мультистарт [91]. Для построения оптимальных многоуровневых моделей используются методы аппроксимации обоснованности [89, 90] с помощью аппроксимации Лапласа [21] и вариационных нижних оценок [84, 85]. Построение оптимальных смесей моделей производится с помощью методов вариационного байесовского вывода [29, 103], а для аппроксимации обоснованности используются аппроксимация Лапласа [21] и построение вариационных нижних оценок [84, 85]. Для исследования статистических свойств распределений используются результаты теории вероятностей [88] и статистики [2, 3, 58, 59]. Для комбинирования признаков используются результаты статистики для оценки ковариационной матрицы по выборке [60, 61].

Основные положения, выносимые на защиту.

1. Разработана теория выбора адекватных оптимальных обученных мультимоделей.
2. Предложена функция сходства распределений, позволяющая решать задачу статистического различения моделей в мультимодели. Показано, что существующие функции сходства не удовлетворяют требованиям к функции сходства.
3. Предложен метод совместного обучения и отбора признаков для смеси моделей.
4. Доказана асимптотическая вырожденность недиагональной оценки максимума обоснованности ковариационной матрицы параметров логистической модели.
5. Предложен метод комбинирования мультиколлинеарных признаков.

6. Разработан программный комплекс для построения адекватных оптимальных обученных мультимоделей в задачах двуклассовой классификации и комбинирования признаков.

Научная новизна. Разработана теория построения адекватных мультимоделей, все модели в которых являются попарно статистически различимыми. Предложен метод статистического сравнения моделей в мультимодели на основании предложенной функции сходства апостериорных распределений параметров моделей. Показано, что существующие функции сходства между распределениями, порожденные дивергенциями Брегмана и f-дивергенциями, а также информативностью на основании дивергенции Кульбака-Лейблера, не являются корректными, то есть не удовлетворяют требованиям к функции сходства, а потому не позволяют решить задачу статистического сравнения моделей. Показано, что предлагаемая фунция сходства является корректной. Исследованы статистические свойства распределения предлагаемой функции сходства в условиях истинности гипотезы о совпадении моделей. Предложен метод совместного обучения и отбора признаков для смесей моделей. Показана асимптотическая вырожденности недиагональной оценки максимума обоснованности для ковариационной матрицы весов признаков. Предложен метод комбинирования мультиколлинеарных признаков на основании оценки ковариационной матрицы для повышения качества классификации. Получены верхняя и нижняя оценки на максимальное число попарно различимых моделе в мультимодели.

Теоретическая значимость. В данной диссертационной работе показано, что предложенные ранее функции сходства между распределениями не позволяют решить задачу сравнения моделей. Построена функция сходства, позволяющая решить задачу статистического сравнения моделей. Исследованы асимптотические свойства распределения предложенной функции сходства в условиях истинности гипотезы о совпадении моделей. На основании этих статистических свойств построена теория выбора $(s,\ \alpha)$ – адекватных мультимоделей. Получены верхняя и нижняя оценка на максимальное число моделей в адекватной мультимодели. Для исключения избыточных признаков используется метод отбора признаков, основанный на максимизации обоснованности мультимодели. Предложен также алгоритм совместного обучения смеси моделей и отбора признаков. Показано, что недиагональная оценка максимума обоснованности для ковариационной матрицы весов признаков является асимптотически вырожденной, а потому для учета зависимостей между признаками предложен метод их комбинирования.

Практическая значимость. Предложенные в работе методы предназначены для построения адекватных оптимальных обученных мультимоделей, позволяющих учесть статистическую неоднородност выборки, для решения задачи двухклассовой классификации; сравнения моделей и устранения избыточных

моделей из мультиодели для повышения интерпретируемости и качества классификации; выявления мультиколлинеарных признаков и их комбинирования для повышения качества классификации.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных методов на реальных задачах двухклассовой классификации на данных по немецким и австралийским потребительским кредитам, по качеству белого вина, по локализации белков в клетках; публикациями результатов исследования в рецензируемых научных изданиях, в том числе рекомендованных ВАК. Результаты работы докладывались и обсуждались на следующих научных конференциях.

1. Международная конференция «20th Conference of the International Federation of Operational Research Societies», 2014, [64].
2. Всероссийская конференция «57я научная конференция МФТИ», 2014.
3. Международная конференция «27th European Conference for Operational Research», 2015, [65].
4. Всероссийская конференция «Математические методы распознавания образов» ММРО-17, 2015, [66].
5. Международная конференция «Интеллектуализация обработки информации», 2016, [67].

Работа поддержана грантами Российского фонда фундаментальных исследований.

1. 14-07-31205, Российский фонд фундаментальных исследований в рамках гранта “Развитие теории выбора мультимоделей в задачах прогнозирования и классификации”.
2. 13-07-13136, Российский фонд фундаментальных исследований в рамках гранта “Математические методы и средства решения задач прогнозирования состояния железнодорожных объектов и инженерных сооружений по спутниковым снимкам”.

Публикации по теме диссертации. Основные результаты по теме диссертации изложены в 14 печатных изданиях, 9 из которых изданы в журналах, рекомендованных ВАК.

1. Адуенко А.А. Выбор признаков и шаговая логистическая регрессия для задачи кредитного скоринга // Машинное обучение и анализ данных, 2012. № 3. С. 279-291, [68].
2. А. А. Адуенко, А. А. Кузьмин, В. В. Стрижов Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Известия ТулГУ, 2012. № 3. С. 119-131 [69].
3. А. А. Адуенко, В. В. Стрижов Алгоритм оптимального расположения названий коллекции документов // Программная инженерия, 2013. № 3. С. 21–25 [70].

4. А. В. Иванова, А. А. Адуенко, В. В. Стрижов Алгоритм построения логических правил при разметке текстов // Программная инженерия, 2013. № 6. С. 41–47 [71].
5. А. А. Адуенко, Н. И. Амелькин О предельных движениях волчка с внутренней диссипацией в однородном поле тяжести // Труды МФТИ, 2013. № 18(2). С. 126-133 [72].
6. А. А. Кузьмин, А. А. Адуенко, В. В. Стрижов Тематическая классификация тезисов крупной конференции с использованием экспертной модели // Информационные технологии, 2014. № 6. С. 22-26 [73].
7. А. А. Адуенко, Н. И. Амелькин Асимптотические свойства движений тяжелого волчка с внутренней диссипацией // ПММ, 2014. Т. 78. Вып. 1. С. 13-28 [74].
8. A. A. Aduenko , V. V. Strijov Multimodelling and Object Selection for Banking Credit Scoring // 20th Conference of the International Federation of Operational Research Societies. — Barcelona: 2014.— P. 136, [64].
9. А. А. Адуенко, В. В. Стрижов Совместный выбор объектов и признаков в задачах многоклассовой классификации коллекции документов // Инфокоммуникационные технологии, 2014. № 1. С. 47–53 [75].
10. A. A. Aduenko , V. V. Strijov Multimodelling and Model Selection in Bank Credit Scoring // 27th European Conference for Operational Research. — Glasgow: 2015. — P. 273, [65].
11. А. А. Адуенко, В. В. Стрижов Анализ пространства параметров в задачах выбора мультимоделей // Математические методы распознавания образов ММРО-17. Тезисы докладов 17-й Всероссийской конференции с международным участием. — г. Светлогорск, Калининградская область: Торус пресс, 2015. С. 10–11, [66].
12. А. А. Адуенко, А. С. Василейский, А. И. Карелов, И. А. Рейер, К. В. Рудаков, В. В. Стрижов Алгоритмы выделения и совмещения устойчивых отражателей на спутниковых снимках // Компьютерная оптика, 2015. Т. 39. Вып. 4. С. 622–630 [76].
13. А. А. Адуенко, Н. И. Амелькин О резонансных вращениях маятника с вибрирующим подвесом // ПММ. 2015. Т. 79. Вып. 6. С. 756–767 [77].
14. А. А. Адуенко, В. В. Стрижов Анализ пространства параметров в задачах выбора мультимоделей // Интеллектуализация обработки информации ИОИ-2016. Тезисы докладов 11-й Международной конференции. — Москва, Россия-Барселона, Испания: Торус пресс, 2016. С. 10–11, [67].

Личный вклад. Все приведенные результаты, кроме отдельно оговоренных случаев, получены диссертантом лично при научном руководстве д.ф.-м.н. В. В. Стрижова.

Структура и объем работы. Диссертация состоит из оглавления, введения, пяти разделов, заключения, списка иллюстраций, списка таблиц, перечня

основных обозначений и списка литературы из 105 наименований. Основной текст занимает 120 страниц.

Краткое содержание работы по главам. В первой главе вводятся основные понятия и определения. Рассматривается задача двухклассовой классификации, ее решение в общем виде, а также понятие оптимальности и обучения вероятностной модели двухклассовой классификации. Приводится определение модели логистической регрессии, являющейся стандартным метод решения задачи двухклассовой классификации. Рассматриваются многоуровневые модели и смеси моделей, а также априорные распределения на параметры моделей и веса моделей в мультимодели.

Во второй главе рассматривается задача построения оптимальной мультимодели. Приведены методы приближенной оптимизации обоснованности, основанные на аппроксимации Лапласа и вариационных нижних оценках. Показана асимптотическая вырожденность недиагональной оценки максимума обоснованности для ковариационной матрицы параметров логистической модели. Рассмотрена задача комбинирования признаков для учета взаимосвязей между ними и предложена схема оптимального по дисперсии шума комбинирования. Предложены алгоритмы детектирования и учета наличия копий одного признака в данных, а также мультиколлинеарности общего вида.

В третьей главе рассматривается задача обучения мультимоделей. Для обучения смеси моделей используется вариационный EM-алгоритм, который позволяет обучить смесь моделей при известных гиперпараметрах смеси. Предложен также алгоритм совместного обучения и оптимизации смеси моделей, основанный на аппроксимации Лапласа и вариационном EM-алгоритме.

В четвертой главе рассматривается понятие адекватной мультимодели. Вводится понятие статистической различимости моделей с помощью расчета функции сходства между апострериорными распределениями параметров моделей. Рассматриваются требования к корректным функциям сходства между распределениями. Показано, что существующие расстояния между распределениями, включая дивергенции Брегмана и f-дивергенции, не порождают корректного сходства, которое применимо для решения задачи сравнения моделей. Рассмотрена также мера сходства, основанная на информативности, построенной по дивергенции Кульбака-Лейблера и показано, что она также не является корректной. Для решения задачи сравнения моделей предложена функция сходства s-score и показано, что она удовлетворяет всем требованиям к функции сходства, включая характеристическое для решаемой задачи требование неотличимости малоинформативного распределения от других. Показано, что для предлагаемая функция сходства корректно определена для распределений с несовпадающими носителями. Получены асимптотические свойства распределения функции сходства в условиях истинности гипотезы о совпадении пары моделей, которые позволяют решать рассматриваемую задачу статистического сравнения моделей. Предложены алгоритмы построения адекватных мульти-

моделей по уже построенной оптимальной обученной мультимодели для смесей моделей и многоуровневых моделей. Показано наличие свойства монотонности у предлагаемой функции сходства, а также получены верхняя и нижняя оценки на число попарно различимых моделей в мультимодели.

В пятой главе на базе предложенных методов разрабатывается программный комплекс, решающий задачу двухклассовой классификации путем построения адекватной оптимальной обученной мультимодели. Работа программного комплекса анализируется на нескольких наборах синтетических и реальных данных по потребительским кредитам, качеству белого вина, локализации белков в клетках. Результаты, полученные с помощью предложенных методов, сравниваются с результатами известных алгоритмов построения мультимоделей и отбора признаков. Приводится иллюстрация вырожденности недиагональной оценки максимума обоснованности для ковариационной матрицы вектора параметров.

Глава 1
Постановка задачи

Задача двухклассовой классификации является одной из базовых задач в области интеллектуального анализа данных. Многие практические задачи, например, задача определения наличия заболевания по анализам [92, 93], задача определения настроения текстовых сообщений [94], задача кредитного скоринга [33, 34, 83], сводятся к решению задачи двухклассовой классификации. Задачи определения релевантности документа [45, 75], категоризации текстов [69, 73, 105], будучи задачами многоклассовой классификации, могут быть эффективно сведены к решению одной или серии задач двухклассовой классификации [95–98]. Таким образом, теоретические результаты в области решения задачи двухклассовой классификации имеют прямое применение на практике.

Определение 1. *Объектом* называется пара $(\mathbf{x},\ y)$, где $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$ есть вектор признакового описания объекта, а $y \in \pm 1$ есть метка класса.

Определение 2. *Признаковой матрицей* для выборки $D = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I} = \{1,\ \ldots,\ m\}$ размера m называется матрица $\mathbf{X} = [\mathbf{x}_1,\ \ldots,\ \mathbf{x}_m]^\intercal \in \mathbb{R}^{m \times n}$.

Определение 3. *Вектором ответов (вектором значений целевой переменной)* для выборки $D = \{(\mathbf{x}_i, y_i)\}$, $i \in \mathcal{I} = \{1,\ \ldots,\ m\}$ размера m называется вектор $\mathbf{y} = [y_1,\ \ldots,\ y_m]^\intercal \in \{-1,\ 1\}^m$.

Определение 4. *Жесткой моделью двухклассовой классификации* называется параметрическое семейство функций $\mathcal{F}$, отображающих декартово произведение множества значений признакового описания объектов $\mathcal{X}$ и множества значений параметров $\mathcal{W}$ в множество значений целевой переменной $\mathcal{Y} = \{-1,\ 1\}$.

$$f :\ \mathcal{X} \times \mathcal{W} \to \{-1,\ 1\}.$$

Определение 5. *Жестким алгоритмом двухклассовой классификации* называется отображение $f :\ \mathcal{X} \to \mathcal{Y} = \{-1,\ 1\}$, сопоставляющее признаковому описанию объекта метку класса.

Определение 6. *Вероятностной моделью двухклассовой классификации* называется совместное распределение вида

$$p(y,\ \mathbf{w}|\mathbf{x},\ \boldsymbol{\alpha}),$$

где $\mathbf{w} \in \mathcal{W}$ есть набор параметров модели, а $\boldsymbol{\alpha} \in Q_{\boldsymbol{\alpha}}$ есть набор гиперпараметров.

Определение 7. *Вероятностным алгоритмом двухклассовой классификации* называется условное распределение вида

$$q(y|\mathbf{x}) = p(y,\ \mathbf{w}^*|\mathbf{x},\ \boldsymbol{\alpha}^*),$$

полученное из вероятностной модели двухклассовой классификации путем фиксирования значений параметров и гиперпараметров модели.

Определение 8. Вероятностная модель двухклассовой классификации называется *оптимальной* для простой выборки $D = \{\mathbf{x}_i,\ y_i\}$, $i \in \{1,\ \ldots,\ m\}$, если гиперпараметры модели выбраны из условия максимума обоснованности, то есть

$$\boldsymbol{\alpha} = \arg\max_{\tilde{\boldsymbol{\alpha}} \in Q_{\boldsymbol{\alpha}}} \prod_{i=1}^{m} p(y_i|\mathbf{x}_i,\ \tilde{\boldsymbol{\alpha}}).$$

Определение 9. *Обучением* вероятностной модели двухклассовой классификации по простой выборке $D = \{\mathbf{x}_i,\ y_i\}$, $i \in \{1,\ \ldots,\ m\}$ называется получение оценок максимума апостериорной веротности для параметров модели, то есть

$$\mathbf{w} = \arg\max_{\tilde{\mathbf{w}} \in \mathcal{W}} \prod_{i=1}^{m} p(y_i,\ \mathbf{w}|\mathbf{x}_i,\ \boldsymbol{\alpha}),$$

где учтено, что $p(\mathbf{w}|\mathbf{X},\ \mathbf{y},\ \boldsymbol{\alpha}) \propto p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \boldsymbol{\alpha})$.

Определение 10. *Оптимальной обученной вероятностной моделью двухклассовой классификации* для простой выборки $D = \{\mathbf{x}_i,\ y_i\}$, $i \in \{1,\ \ldots,\ m\}$ называется оптимальная вероятностная модель двухклассовой классификации для D, для которой произведено обучение по D для оптимального значения гиперпараметров, то есть

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha} \in Q_{\boldsymbol{\alpha}}} \prod_{i=1}^{m} p(y_i|\mathbf{x}_i,\ \boldsymbol{\alpha}),$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w} \in \mathcal{W}} \prod_{i=1}^{m} p(y_i,\ \mathbf{w}|\mathbf{x}_i,\ \boldsymbol{\alpha}^*).$$

Определение 11. *Задачей двухклассовой классификации* называется задача построения жесткого алгоритма двухклассовой классификации, то есть отображения $f : \mathcal{X} \to \mathcal{Y} = \{-1,\ 1\}$, сопоставляющего признаковому описанию объекта метку класса.

Решением задачи двухклассовой классификации является, например, алгоритм k ближайших соседей с фиксированной метрикой и числом соседей k, полученный в рамках жесткой модели двухклассовой классификации, в которой гиперпараметром выступает число соседей k. Однако, вероятностный алгоритм двухклассовой классификации тоже можно использовать для решения этой задачи, путем добавления решающего правила, которое по $q(y|\mathbf{x})$ выдает метку класса. Стандартным решающим правилом, которое используется в данной работе является следующее.

Определение 12. *Стандартным* жестким алгоритмом двухклассовой классификации, порожденным вероятностным, заданным условным распределением $q(y|\mathbf{x})$, называется отображение $f : \mathcal{X} \to \mathcal{Y} = \{-1,\ 1\}$ вида

$$f(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} q(y|\mathbf{x}).$$

Опишем теперь модель логистической регрессии, которая является стандартной моделью для решения задач двухклассовой классификации во многих прикладных областях [8, 83, 92, 93].

Определение 13. *Моделью логистической регрессии (или одиночной моделью логистической регрессии)* называется вероятностная модель двухклассовой классификации, для которой

$$p(y,\ \mathbf{w}|\mathbf{x},\ \boldsymbol{\alpha}) = p(\mathbf{w}|\boldsymbol{\alpha})\sigma(y\mathbf{w}^{\intercal}\mathbf{x}),$$

где $\mathbf{w} \in \mathcal{W} = \mathbb{R}^n$, $p(\mathbf{w}|\boldsymbol{\alpha})$ есть априорное распределение на вектор параметров $\mathbf{w}$, а $\sigma(x) = 1/(1 + \exp(-x))$.

Приведенная выше модель в определении 13 описывает стандартную модель логистической регрессии, в которой предполагается, что важность признаков не зависит от точки в признаковом пространстве $\mathcal{X}$. При наличии неоднородностей в данных таких, как, например, разная важная признаков в зависимости от кластера данных, требуется от одиночной модели логистической регрессии перейти к моделям и методам, которые позволяют учесть неоднородности. Далее рассмотрим мультимоделирование для учета неоднородностей в данных.

1.1. Понятие мультимодели. Смеси моделей и многоуровневые модели

Зачастую данные могут быть статистически неоднородны, то есть могут содержать несколько совокупностей объектов, для которых взаимосвязь целевой переменной с признаковым описанием не описывается единственной моделью в рамках рассматриваемого семейства моделей. Так при определении вероятности неплатежа заемщика можно предположить, что важность признака доход будет зависеть от размера этого дохода. Также его важность и важность, например, размера семьи может зависеть от географических и социальных различий между разными регионами. Одиночная модель логистической регрессии, являющаяся стандартом в банковском скоринге [8, 83], не может описать такую неоднородность.

Проблему неоднородности в данных позволяют решить мультимодели. Например, для учета возможной неоднородности данных производят разбиения признакового пространства на непересекающиеся множества и для объектов из каждого подмножества строится своя модель [**?**, 49, 78]. Такой подход де-факто является стандартным в кредитном скоринге [**?**, **?**], а само разбиение признакового пространства на части может производиться путем кластеризации объектов [**?**, 25, 26] или путем деления объектов по значению одного или нескольких признаков [**?**, **?**]. Так, если таким признаком выступает возраст, то для клиентов, попавших в каждую из групп возрастов строится отдельная модель. Такой подход с жестким отнесением объекта к одной из моделей на основании его признакового описания порождает многоуровневые модели.

Другим подходом для учета неоднородностей в данных являются смеси моделей [79, 80]. В отличие от многоуровневых моделей, где производится жесткое прикрепление каждого объекта к своей модели на основании признакового описания объекта, в случае мультимоделей для каждого объекта есть некоторая вероятность $\pi_k \in [0,\, 1]$, $k = \overline{1,\, K}$ принадлежности каждой из моделей и прогнозирование класса объекта производится в виде смеси прогнозов в рамках каждой из моделей.

Отметим, что модели, входящие в мультимодель могут быть близки или даже совпадать. В случае многоуровневых моделей это фактически означает сокращение размера выборки для каждой из копий модели, что ведет к большей неточности в определении параметров моделей, то есть к ухудшению качества прогноза (см., например, табл. 5.5, 5.6 и описание в вычислительном эксперименте). Тот же эффект может наблюдаться и для смесей моделей. Более того, если многоуровневая модель или смесь моделей имеет большое число схожих моделей, это ведет к их неинтерпретируемости, что влияет не только на качество прогноза, но и на применимость, например, в кредитном скоринге. Такие мультимодели являются неадекватными имеющимся данным, а потому для учета неоднородностей в данных возникает задача построения адекватных мультимоделей, для которых модели, их составляющие, будут различимыми. Опишем далее, что такое смесь моделей и многоуровневая модель и их обучение, а также требование оптимальности к ним.

Определение 14. *Смесью моделей* называется вероятностная модель, совместное распределение которой для выборки $(\mathbf{X},\, \mathbf{y})$ размера m имеет вид

$$p(\mathbf{y},\, \boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K|\mathbf{X},\, \alpha,\, \mathbf{A}_1,\, \ldots,\, \mathbf{A}_k) = p(\boldsymbol{\pi}|\alpha) \prod_{k=1}^{K} p_k(\mathbf{w}_k|\mathbf{A}_k) \prod_{i=1}^{m} \left(\sum_{l=1}^{K} \pi_l f_l(\mathbf{w}_l,\, \mathbf{x}_i,\, y_i) \right), \quad (1.1)$$

где K – число моделей, входящих в мультимодель, $f_l(\mathbf{w}_l,\, \mathbf{x}_i,\, y_i)$ есть правдоподобие для объекта $(\mathbf{x}_i,\, y_i)$ в модели l, $\boldsymbol{\pi} = [\pi_1,\, \ldots,\, \pi_K]^\intercal$ есть веса моделей, входящих в мультимодель, $\mathbf{w}_1,\, \ldots,\, \mathbf{w}_K$ есть параметры моделей, входящих в мультимодель, а $\alpha \in Q_\alpha$, $\mathbf{A}_1 \in Q_{\mathbf{A}_1}$, $\ldots$, $\mathbf{A}_K \in Q_{\mathbf{A}_K}$ есть гиперпараметры, определяющие априорные распределения вектора весов $\boldsymbol{\pi}$ и векторов параметров моделей $\mathbf{w}_1,\, \ldots,\, \mathbf{w}_K$ соответственно, а Q_α, $Q_{\mathbf{A}_1}$, $\ldots$, $Q_{\mathbf{A}_K}$ есть множества допустимых значений гиперпараметров.

Определение 15. *Многоуровневой моделью* называется вероятностная модель, совместное распределение которой для выборки $(\mathbf{X},\, \mathbf{y})$ размера m имеет вид

$$p(\mathbf{y}, \mathbf{w}_1, \ldots, \mathbf{w}_K|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K) = \prod_{k=1}^{K} p_k(\mathbf{w}_k|\mathbf{A}_k) \prod_{i=1}^{m} \prod_{l=1}^{K} f_l(\mathbf{w}_l, \mathbf{x}_i, y_i)^{[\mathbf{x}_i \in \Omega_l]}, \quad (1.2)$$

где K – число моделей, входящих в многоуровневую модель, $f_l(\mathbf{w}_l,\ \mathbf{x}_i,\ y_i)$ есть правдоподобие для объекта $(\mathbf{x}_i,\ y_i)$ в модели l, $\mathbb{R}^n = \Omega_1 \sqcup \ldots \sqcup \Omega_K$ есть разбиения пространства на области действия моделей, $\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K$ есть параметры моделей, входящих в мультимодель, $\mathbf{A}_1 \in Q_{\mathbf{A}_1},\ \ldots,\ \mathbf{A}_K \in Q_{\mathbf{A}_K}$ есть гиперпараметры, определяющие априорные распределения векторов параметров моделей $\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K$, а $Q_{\mathbf{A}_1},\ \ldots,\ Q_{\mathbf{A}_K}$ есть множества допустимых значений гиперпараметров.

Замечание 1. Отметим, что распространенным подходом в банковском скоринге является разбиение множества клиентских записей на непересекающиеся подмножества и построение отдельной модели для каждого подмножества. Разбиение при этом может быть сделано по значениям некоторого признака (например, возраста или дохода), либо путем кластеризации. В таком случае итоговая предсказательная модель как раз и является многоуровневой моделью.

Определение 16. Смесь моделей называется *оптимальной*, если она обладает наибольшей обоснованностью, то есть совместное правдоподобие мультимодели имеет вид $p(\mathbf{y},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{X},\ \alpha^*,\ \mathbf{A}_1^*,\ \ldots,\ \mathbf{A}_k^*)$, где

$$[\alpha^*,\ \mathbf{A}_1^*,\ \ldots,\ \mathbf{A}_K^*] = \underset{\alpha\in Q_\alpha,\ \mathbf{A}_1\in Q_{\mathbf{A}_1},\ \ldots,\ \mathbf{A}_K\in Q_{\mathbf{A}_K}}{\arg\max} p(\mathbf{y}|\mathbf{X},\ \alpha,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K). \qquad (1.3)$$

Определение 17. Многоуровневая модель называется *оптимальной*, если она обладает наибольшей обоснованностью, то есть совместное правдоподобие многоуровневой модели имеет вид $p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{X},\ \mathbf{A}_1^*,\ \ldots,\ \mathbf{A}_k^*)$, где

$$[\mathbf{A}_1^*,\ \ldots,\ \mathbf{A}_K^*] = \underset{\mathbf{A}_1\in Q_{\mathbf{A}_1},\ \ldots,\ \mathbf{A}_K\in Q_{\mathbf{A}_K}}{\arg\max} p(\mathbf{y}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K). \qquad (1.4)$$

Отметим, что в определении оптимальности максимизация производится по набору априорных распределений. Потому, если в качестве допустимого множества априорных распределений для вектора весов моделей или векторов их параметров взять достаточно широкое семейство распределений, результат получится тривиальным. Поэтому в данной работе в качестве априорных распределений на вектор весов моделей, входящих в мультимодель, и на вектора параметров моделей в мультимодели и многоуровневой модели используются достаточно узкие классы параметрических распределений, конкретный вид которых будет указан в дальнейшем.

Определение 18. *Обучением* смеси моделей, заданной совместным правдоподобием $p(\mathbf{y},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{X},\ \alpha,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_k)$, называется получение оценок максимума апостериорной вероятности на веса моделей, входящих в смесь, и на векторы их параметров, то есть

$$[\boldsymbol{\pi}^*,\ \mathbf{w}_1^*,\ \ldots,\ \mathbf{w}_K^*] = \underset{\boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K}{\arg\max}\ p(\boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{y},\ \mathbf{X},\ \alpha,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K). \quad (1.5)$$

Определение 19. *Обучением* многоуровневой модели, заданной совместным правдоподобием $p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_k)$, называется получение оценок максимума апостериорной вероятности на векторы параметров моделей, входящих в многоуровневую модель, то есть

$$[\mathbf{w}_1^*,\ \ldots,\ \mathbf{w}_K^*] = \arg\max_{\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K} p(\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{y},\ \mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K). \tag{1.6}$$

Опишем далее конкретный вид априорных распределений на вектор весов моделей в смеси моделей и на вектора параметров моделей в смеси моделей и многоуровневой модели, а также вид моделей f_k, $k = \overline{1,\ K}$, используемых в данной работе.

Вид мультимоделей, использующихся в данной работе. В качестве моделей, составляющих мультимодель, используются одиночные модели логистической регрессии. В качестве априорных распределений на веса моделей в смеси моделей используется симметричное распределение Дирихле, а на вектора параметров моделей, входящих в мультимодель накладываем априорное нормальное распределение с нулевым средним. Тогда совместное правдоподобие для смеси моделей и для многоуровневых модели приобретает следующий вид.

Для смеси моделей совместное правдоподобие имеет вид

$$p(\mathbf{y},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{X},\ \alpha,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K) = p(\boldsymbol{\pi}|\alpha)\prod_{k=1}^{K} p_k(\mathbf{w}_k|\mathbf{A}_k)\prod_{i=1}^{m}\left(\sum_{l=1}^{K}\pi_l\sigma(y_i\mathbf{w}_l^{\intercal}\mathbf{x}_i)\right),\ \text{где} \tag{1.7}$$

$$p(\boldsymbol{\pi}|\alpha) = \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)}\prod_{k=1}^{K}\pi_k^{\alpha-1},\ p(\mathbf{w}_k|\mathbf{A}_k) = \frac{\sqrt{\det\mathbf{A}_l}}{(2\pi)^{n/2}}\exp\left(-\tfrac{1}{2}\mathbf{w}_l^{\intercal}\mathbf{A}_l\mathbf{w}_l\right),\ k = 1,\ \ldots,\ K.$$

Для многоуровневой модели совместное правдоподобие имеет вид

$$p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K) = \prod_{k=1}^{K} p_k(\mathbf{w}_k|\mathbf{A}_k)\prod_{i=1}^{m}\prod_{l=1}^{K}\sigma(y_i\mathbf{w}_l^{\intercal}\mathbf{x}_i)^{[\mathbf{x}_i\in\Omega_l]},\ \text{где} \tag{1.8}$$

$\mathbb{R}^n = \Omega_1 \sqcup \ldots \sqcup \Omega_K$ есть разбиения пространства на области действия моделей, а априорные распределения на векторы параметров моделей есть

$$p(\mathbf{w}_k|\mathbf{A}_k) = \frac{\sqrt{\det\mathbf{A}_l}}{(2\pi)^{n/2}}\exp\left(-\tfrac{1}{2}\mathbf{w}_l^{\intercal}\mathbf{A}_l\mathbf{w}_l\right),\ k = 1,\ \ldots,\ K.$$

Глава 2
Построение оптимальной мультимодели. Отбор и комбинирование признаков

Оптимальной мультимоделью в соответствии с (1.3), (1.4) является мультимодель, обладающая наибольшей обоснованностью. При этом максимизация обоснованности производится по априорным распределениям весов моделей, входящим в смесь моделей $p(\boldsymbol{\pi}|\alpha)$, а также по априорным распределениям $p(\mathbf{w}_k|\mathbf{A}_k)$ векторов параметров моделей $\mathbf{w}_k$, $k = 1, \ldots, K$, входящих в смесь моделей или многоуровневую модель.

При использовании широкого класса априорных распределений, в котором проводится оптимизация обоснованности, результат является тривиальным, а потому в данной работе используются достаточно узкие классы параметрических распределений. Так $p(\boldsymbol{\pi}|\alpha) = \mathrm{Dir}(\boldsymbol{\pi}|\alpha)$, $p(\mathbf{w}_k|\mathbf{A}_k) = N(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1})$, $k = 1, \ldots, K$. При этом $\mathbf{A}_k \in Q_{\mathbf{A}_k} = \mathcal{M}$, $k = 1, \ldots, K$, где $\mathcal{M}$ есть некоторое подмножество множества неотрицательно определенных симметричных матриц размера $n \times n$. Предполагая далее параметр α распределения Дирихле фиксированным, опишем метод получения оптимальной мультимодели, который приведет к отбору признаков. Рассмотрим сначала случай $K = 1$, то есть случай одиночной модели.

2.1. Отбор признаков с помощью максимизации обоснованности для случая одиночной модели

В случае одиночной модели, то есть когда $K = 1$ имеем следующее совместное правдоподобие

$$p_{\boldsymbol{\gamma}}(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) = p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w}|\mathbf{A}) = \prod_{i=1}^{m} \sigma^{\gamma_i}(y_i \mathbf{w}^{\intercal}\mathbf{x}_i) N(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1}), \qquad (2.1)$$

где весовой вектор $\boldsymbol{\gamma} = [1, \ldots, 1]^{\intercal} \in \mathbb{R}^m$ введен для общности (полученные формулы с нетривиальным $\boldsymbol{\gamma}$ будут использованы для оптимизации смесей моделей) и считается, что распределение $\mathbf{w}$ от $\mathbf{X}$ не зависит, метки классов y_i есть ± 1, а $\sigma(x)$ есть сигмоидная функция. Тогда принцип максимума обоснованности для данной модели приобретает вид

$$\mathbf{A}^* = \arg\max_{\mathbf{A} \in \mathcal{M}} p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{A}),$$

где $\mathcal{M}$ – некоторое подмножество симметричных неотрицательно определенных матриц. В качестве $\mathcal{M}$ можно рассматривать, например, множество диагональных матриц с неотрицательными элементами на диагонали $\mathcal{M}_{diag}$ или множество всех симметричных неотрицательно определенных матриц $\mathcal{M}_{full}$.

Отметим, что полученная матрица $\mathbf{A}^*$ позволяет установить относительную важность признаков. Так при $\mathbf{A}^*_{jj} = \infty$ получаем, что соответствующий вес признака с номером j w_j априори равен 0, то есть признак является избыточным. Оценка недиагональной матрицы могла бы позволить учесть не только важность каждого признака в отдельности, но и установить взаимосвязи между ними. Проиллюстрируем это рассмотрением следующей модельной ситуации.

Иллюстрация важности учета взаимосвязи между признаками. Задача алгоритмов отбора признаков состоит в выборе подмножества признаков $\mathcal{A}$ исходного множества признаков $\{1, \ldots, n\}$ такого, что максимизируется некоторый критерий качества, который и определяет алгоритм отбора. При этом при наличии мультиколлинеарных признаков мультиколлинеарность устраняют [7] путем удаления одного или нескольких мультиколлинеарных признаков из каждой группы мультиколлинеарных признаков, чтобы полученная матрица признаков $\mathbf{X}_\mathcal{A}$ не была плохо обусловленной и оценки параметров модели были потому устойчивыми. Покажем, что такой подход не является оптимальным и исключение мультиколлинеарных признаков может приводить к ухудшению качества прогноза.

Теорема 1 (Адуенко, 2016)**.** Пусть имеется l независимых в совокупности факторов, а модель генерации данных имеет вид

$$y_i \sim \mathrm{Be}(\sigma(\mathbf{v}^\intercal \mathbf{f}_i)),$$

где $\mathbf{v}$ – вектор параметров модели, а $\mathbf{f}_i$ – вектор значений факторов для i–го объекта. Пусть для каждого объекта вместо пары $(\mathbf{f}_i,\ y_i)$ наблюдается пара $(\mathbf{x}_i,\ y_i)$, где $\mathbf{x}_i = \mathbf{G}\mathbf{f}_i + \boldsymbol{\varepsilon}_i$, где $\mathbf{G}$ – матрица размера $n \times l$, $n \geq l$ полного ранга, а $\boldsymbol{\varepsilon}_i$ – центрированный шум с невырожденной ковариационной матрицей $\boldsymbol{\Sigma}$, причем случайные вектора $\{\boldsymbol{\varepsilon}_i\}$, $i = 1, \ldots$ для разных объектов независимы в совокупности.

Тогда оптимальной в терминах дисперсии шума $\mathbb{E}(\mathbf{w}^\intercal \boldsymbol{\varepsilon}_i)^2$ оценкой $\mathbf{w}$, дающей несмещенную оценку $\mathbf{w}^\intercal \mathbf{x}_i$, является оценка вида

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}\mathbf{G}\left(\mathbf{G}^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{G}\right)^{-1}\mathbf{v}. \tag{2.2}$$

Доказательство. Рассмотрим сначала требование несмещенности $\mathbf{w}^\intercal \mathbf{x}_i$. Так как матрица $\mathbf{G}$ полного ранга, то значения всех факторов для любого объекта $\mathbf{f}_i$ при отсутствии шума можно точно восстановить по $\mathbf{x}_i$, причем при $n > l$ с избытком. При этом требование несмещенности $\mathbf{w}^\intercal \mathbf{x}_i$ лишь означает, что $\mathbf{w}^\intercal \mathbf{x}_i = \mathbf{v}^\intercal \mathbf{f}_i$. При этом в силу отсутствия шума любой вектор $\mathbf{w}$, удовлетворяющий этому требованию, подходит и дает одинаковый результат. В частности можно оставить произвольные l линейно независимых признаков и далее работать с ними. Из-за наличия шума ситуация изменяется и вектора $\mathbf{w}$, удовлетворяющие условию несмещенности $\mathbb{E}\mathbf{w}^\intercal \mathbf{x}_i = \mathbf{v}^\intercal \mathbf{f}_i$, перестают быть равноценными, так как для разных $\mathbf{w}$ $\mathbf{w}^\intercal \mathbf{x}_i$ обладает разной дисперсией.

Рассмотрим сначала условие несмещенности. Из него получаем

$$\mathbb{E}\mathbf{w}^\intercal\mathbf{x}_i = \mathbf{w}^\intercal\mathbf{G}\mathbf{f}_i = \mathbf{v}^\intercal\mathbf{f}_i,$$

откуда допустимые $\mathbf{w}$ удовлетворяют соотношению $\mathbf{G}^\intercal\mathbf{w} = \mathbf{v}$. При этом

$$\mathbf{w}^\intercal\mathbf{x}_i = \mathbf{w}^\intercal(\mathbf{G}\mathbf{f}_i + \boldsymbol{\varepsilon}_i) = \mathbf{v}^\intercal\mathbf{f}_i + \mathbf{w}_i^\intercal\boldsymbol{\varepsilon}_i,$$

откуда условие оптимальности по дисперсии шума линейной комбинации принимает вид

$$\mathbb{E}(\mathbf{w}_i^\intercal\boldsymbol{\varepsilon}_i)^2 = \mathbf{w}^\intercal\mathbb{E}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^\intercal)\mathbf{w} = \mathbf{w}^\intercal\boldsymbol{\Sigma}\mathbf{w} \to \min_{\mathbf{w}},$$

откуда с учетом условия допустимости $\mathbf{w}$ получаем следующую задачу оптимизации

$$\begin{aligned} \mathbf{w}^\intercal\boldsymbol{\Sigma}\mathbf{w} &\to \min_{\mathbf{w}}, \\ \text{при условии } \mathbf{G}^\intercal\mathbf{w} &= \mathbf{v}. \end{aligned} \tag{2.3}$$

Решая задачу (2.3) и вводя вектор множителей Лагранжа $\boldsymbol{\lambda} \in \mathbb{R}^l$, получаем условия оптимальности

$$\begin{aligned} 2\boldsymbol{\Sigma}\mathbf{w} + \mathbf{G}\boldsymbol{\lambda} &= \mathbf{0}, \\ \mathbf{G}^\intercal\mathbf{w} &= \mathbf{v}, \end{aligned}$$

откуда

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}\mathbf{G}\left(\mathbf{G}^\intercal\boldsymbol{\Sigma}^{-1}\mathbf{G}\right)^{-1}\mathbf{v}.$$

□

Проиллюстрируем доказанную теорему несколькими примерами.

- Случай совокупности n мультиколлинеарных признаков, то есть $l = 1$, $\mathbf{G} = \mathbf{e} = [1, \ldots, 1]^\intercal$, $\boldsymbol{\Sigma}^{-1} = \operatorname{diag}(1/\sigma_1^2, \ldots, 1/\sigma_n^2)$.
  В рассматриваемом случае имеется только один фактор, а признаковое описание представляет собой n его зашумленных копий. Тогда все признаки являются мультиколлинеарными и с точки зрения отбора признаков из них стоит оставит только один. Однако в соответствии с теоремой такой подход не является оптимальным, а для достижения минимальной дисперсии шума требуется сложить признаки с некоторыми положительными весами. Получим выражение для этих весов в соответствии с (2.2).

$$\mathbf{w} = \begin{pmatrix} 1/\sigma_1^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1/\sigma_n^2 \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \frac{v}{\frac{1}{\sigma_1^2} + \ldots + \frac{1}{\sigma_n^2}} = v \left[\frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \ldots + \frac{1}{\sigma_n^2}}, \ldots, \frac{\frac{1}{\sigma_n^2}}{\frac{1}{\sigma_1^2} + \ldots + \frac{1}{\sigma_n^2}} \right] \tag{2.4}$$

  Таким образом, складывать признаки стоит в весами обратно пропорциональными дисперсии шума соответствующей копии фактора. В частности, если $\sigma_1^2 = \ldots = \sigma_n^2$, то оптимально взять в качестве оценки фактора среднее значение n его зашумленных копий.

- Тройка мультиколлинеарных признаков, то есть
  $l = 2,\ n = 3,\ \mathbf{G} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}^{\intercal},\ \mathbf{\Sigma}^{-1} = \mathrm{diag}(1/\sigma_1^2,\ 1/\sigma_2^2,\ 1/\sigma_3^2).$
  Аналогично предыдущему, используя (2.2), получим

$$\mathbf{w} = \frac{1}{\frac{1}{\sigma_1^2\sigma_2^2} + \frac{1}{\sigma_1^2\sigma_3^2} + \frac{1}{\sigma_2^2\sigma_3^2}} \begin{pmatrix} \frac{v_1}{\sigma_1^2\sigma_2^2} + \frac{v_1 - v_2}{\sigma_1^2\sigma_3^2} \\ \frac{v_2}{\sigma_1^2\sigma_2^2} + \frac{v_2 - v_1}{\sigma_2^2\sigma_3^2} \\ \frac{v_1}{\sigma_2^2\sigma_3^2} + \frac{v_2}{\sigma_1^2\sigma_3^2} \end{pmatrix}.$$

  Так при $\sigma_1^2 \to \infty$, получим, что оптимально использовать только второй и третий признаки. Аналогично при $\sigma_2^2 \to \infty$ оптимально использовать только первый и третий признак, а при $\sigma_3^2 \to \infty$ – только первый и второй. При этом даже, если, например, $\sigma_1,\ \sigma_2^2 \gg \sigma_3^2$ при $v_1 \neq v_2$ приходится использовать один из первых двух признаков, несмотря на их сильную зашумленность. Однако при $v_1 = v_2$ в той же ситуации оптимально будет использовать только третий признак.
- Две пары мультиколлинеарных признаков, то есть
  $l = 2,\ n = 4,\ \mathbf{G} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}^{\intercal},\ \mathbf{\Sigma}^{-1} = \mathrm{diag}(1/\sigma_1^2,\ 1/\sigma_2^2,\ 1/\sigma_3^2,\ 1/\sigma_4^2)$
  Этот случай сводится к независимому применению результат случая 1 для первой и второй пары мультиколлинеарных признаков.

Таким образом, при наличии мультиколлинеарных признаков устранение мультиколинеарности путем удаления избыточных признаков не является оптимальным, а информация из избыточных признаков может быть использована для улучшения качества прогноза. Поэтому наряду с оценкой важности отдельных признаков, будем оценивать и взаимосвязь между ними.

Оптимизация обоснованности для случая одной модели. Для случая одной модели с совместным правдоподобием

$$p_{\boldsymbol{\gamma}}(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) = p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w}|\mathbf{A}) = \prod_{i=1}^{m} \sigma^{\gamma_i}(y_i \mathbf{w}^{\intercal}\mathbf{x}_i) N(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1}),$$

где $\boldsymbol{\gamma} = [1,\ \ldots,\ 1]^{\intercal} \in \mathbb{R}^m$, оценка максимума обоснованности для матрицы $\mathbf{A}$ имеет вид

$$\mathbf{A}^* = \arg\max_{\mathbf{A} \in \mathcal{M}} p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{A}) = \arg\max_{\mathbf{A} \in \mathcal{M}} \int p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{w}) p(\mathbf{w}|\mathbf{A}) d\mathbf{w}, \tag{2.5}$$

где $\mathcal{M}$ – некоторое подмножество симметричных неотрицательно определенных матриц. Как уже указывалось в качестве $\mathcal{M}$ можно рассматривать, например, множество $\mathcal{M}_{diag}$ всех диагональных матриц размера $n \times n$ с неотрицательными элементами на диагонали. Тогда если $\mathbf{A}^*_{jj} = \infty$, то признак с номером j является незначимым. Для учета зависимостей между признаками можно использовать в качестве $\mathcal{M}$, например, множество всех неотрицательно определенных

матриц $\mathcal{M}_{full}$ размера $n \times n$. Если исходно известны возможные взаимосвязи между признаками, то можно использовать множество матриц $\mathcal{M}$, учитывающее эту информацию. Так, если, например, известно, что признаки упорядочены и связаны могут быть только соседние, в качестве множества матриц $\mathcal{M}$ можно использовать множество трехдиагональных матриц. Однако для реальных данных структуры зависимостей между признаками обычно неизвестна, а потому далее используем в качестве множеств матриц $\mathcal{M}$ $\mathcal{M}_{diag}$ и $\mathcal{M}_{full}$. Отметим, однако, что получение соответствующих формул для других множеств матриц полностью аналогично. Кроме того, как будет показано в дальнейшем, с помощью максимизации обоснованности не удается получить нетривиальную оценку недиагональной части матрицы $\mathbf{A}$ даже асимптотически, а потому оценка максимума обоснованности для недиагональной матрицы $\mathbf{A}$ не представляет интереса, так как никак не указывает на взаимосвязи между признаками.

Отметим, что интеграл в выражении (2.5) аналитически не считается, поскольку правдоподобие $p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X}, \mathbf{w})$ и априорное распределение $p(\mathbf{w}|\mathbf{A})$ не являются сопряженными распределениями, так как априорное распределение нормальное, а правдоподобие не является нормальным относительно $\mathbf{w}$, откуда апостериорное распределение также не будет нормальным. Существует два основных подхода к аппроксимации интеграла в выражении (2.5) и нахождения матрицы $\mathbf{A}$, метод, основанный на аппроксимации Лапласа и метод, основанный на вариационной нижней оценки для сигмоидной функции [21].

Оценка ковариационной матрицы с помощью аппроксимации Лапласа

Перепишем интеграл в выражении (2.5) в следующем виде

$$I(\mathbf{A}) = \int p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w}|\mathbf{A})d\mathbf{w} = \int e^{\log Q_{\boldsymbol{\gamma}}(\mathbf{w})}d\mathbf{w},$$

где $Q_{\boldsymbol{\gamma}}(\mathbf{w}) = (p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{A}))$. Обозначим $q_{\boldsymbol{\gamma}}(\mathbf{w}) = \log Q_{\boldsymbol{\gamma}}(\mathbf{w})$. Пусть $\mathbf{w}_{MP}$ – наиболее вероятное апостериори значение параметров $\mathbf{w}$, то есть

$$\mathbf{w}_{MP} = \arg\max_{\mathbf{w}} p_{\boldsymbol{\gamma}}(\mathbf{w}|\mathbf{X},\ \mathbf{y},\ \mathbf{A}) = \arg\max_{\mathbf{w}} \frac{p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w}|\mathbf{A})}{\int p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{w}')p(\mathbf{w}'|\mathbf{A})d\mathbf{w}'}. \qquad (2.6)$$

Тогда из (2.6) и определения $Q(\mathbf{w})$ имеем

$$\mathbf{w}_{MP} = \arg\max_{\mathbf{w}} q(\mathbf{w}).$$

С учетом определения $\mathbf{w}_{MP}$ воспользуемся разложением Тейлора для $q(\mathbf{w})$ в окрестности $\mathbf{w}_{MP}$ до второго порядка включительно. Линейный член отсутствует, так как $\mathbf{w}_{MP}$ есть точка максимума $q(\mathbf{w})$.

$$q(\mathbf{w}) \approx q(\mathbf{w}_{MP}) - \tfrac{1}{2}(\mathbf{w} - \mathbf{w}_{MP})^{\intercal}\mathbf{H}_{\boldsymbol{\gamma}}^{-1}(\mathbf{w} - \mathbf{w}_{MP}), \text{ где } \mathbf{H}_{\boldsymbol{\gamma}}^{-1} = -\nabla^2 q_{\boldsymbol{\gamma}}(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{MP}}. \qquad (2.7)$$

Используем (2.7) для приближения $I(\mathbf{A})$, получим

$$I(\mathbf{A}) \approx Q(\mathbf{w}_{MP}) \int \exp\left(-\tfrac{1}{2}(\mathbf{w}-\mathbf{w}_{MP})^{\intercal}\mathbf{H}_{\boldsymbol{\gamma}}^{-1}(\mathbf{w}-\mathbf{w}_{MP})\right) d\mathbf{w} = Q_{\boldsymbol{\gamma}}(\mathbf{w}_{MP})(2\pi)^{n/2}\sqrt{\det$$

где при взятии интеграла использовано выражение для константы многомерного нормального распределения с математическим ожиданием $\mathbf{w}_{MP}$ и ковариационной матрицей $\mathbf{H}_{\boldsymbol{\gamma}}$. При этом $\mathbf{H}_{\boldsymbol{\gamma}}$ положительно определена, так как $\mathbf{H}_{\boldsymbol{\gamma}}^{-1}$ есть гессиан функции $-q_{\boldsymbol{\gamma}}(\mathbf{w})$ в точке $\mathbf{w}_{MP}$, а функция $-q_{\boldsymbol{\gamma}}(\mathbf{w})$ выпуклая как сумма двух выпуклых: выпуклого отрицательного логарифма правдоподобия $-\log p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{w})$ и квадратичной функции из $-\log p(\mathbf{w}|\mathbf{A})$. Раскроем теперь $Q_{\boldsymbol{\gamma}}(\mathbf{w}_{MP})$ и получим итоговое выражение для $\log I(\mathbf{A})$.

$$q_{\boldsymbol{\gamma}}(\mathbf{w}_{MP}) = \log p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{w}_{MP}) - \frac{n}{2}\log(2\pi) + \tfrac{1}{2}\log\det\mathbf{A} - \tfrac{1}{2}\mathbf{w}_{MP}^{\intercal}\mathbf{A}\mathbf{w}_{MP},\ \text{откуда}$$

$$\log I(\mathbf{A}) = \log p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{w}_{MP}) + \tfrac{1}{2}\log\det\mathbf{A} - \tfrac{1}{2}\mathbf{w}_{MP}^{\intercal}\mathbf{A}\mathbf{w}_{MP} - \tfrac{1}{2}\log\det\mathbf{H}^{-1}. \quad (2.8)$$

Выпишем теперь выражение для $\mathbf{H}_{\boldsymbol{\gamma}}$ и перейдем затем к описанию оптимизационной задачи и ее решения.

$$\mathbf{H}_{\boldsymbol{\gamma}}^{-1} = -\frac{\partial^2}{\partial\mathbf{w}^2}q_{\boldsymbol{\gamma}}(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{MP}} = -\frac{\partial^2}{\partial\mathbf{w}^2}(\log p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\mathbf{w})+\log p(\mathbf{w}|\mathbf{A}))|_{\mathbf{w}=\mathbf{w}_{MP}} = \mathbf{X}^{\intercal}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{X}+A$$

$\mathbf{R}_{\boldsymbol{\gamma}} = \mathrm{diag}(\gamma_i\sigma(\mathbf{w}_{MP}^{\intercal}\mathbf{x}_i)\sigma(-\mathbf{w}_{MP}^{\intercal}\mathbf{x}_i),\ i=\overline{1,\ m})$. Отсюда наибольший вес получают объекты, близкие к границе между классами в терминах вероятности принадлежности классам. Заменяя знак в $\log I(\mathbf{A})$ для получения задачи минимизации вместо задачи максимизации, имеем

$$\mathbf{A} = \arg\min_{\mathbf{A}\in\mathcal{M}}\left[-\log p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{w}_{MP}) - \tfrac{1}{2}\log\det\mathbf{A} + \tfrac{1}{2}\mathbf{w}_{MP}^{\intercal}\mathbf{A}\mathbf{w}_{MP} + \tfrac{1}{2}\log\det(\mathbf{X}^{\intercal}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{X}+\right. \quad (2.9)$$

Отметим, что в (2.9) оценка максимума апостериорной вероятности для параметров $\mathbf{w}_{MP}$ зависит от матрицы $\mathbf{A}$, поэтому получение полной производной по $\mathbf{A}$ затруднено. Потому для решения задачи (2.9) используем итеративный алгоритм с поочередным пересчетом $\mathbf{w}_{MP}$ и $\mathbf{A}$. При этом в качестве множества матриц $\mathcal{M}$, в котором ищем решение для матрицы $\mathbf{A}$ рассматриваем $\mathcal{M}_{diag}$ и $\mathcal{M}_{full}$.

Опишем теперь итерационный алгоритм решения задачи (2.9). Отметим, что по определению $\mathbf{w}_{MP}$ существует однозначная связь между $\mathbf{A}$ и $\mathbf{w}_{MP}$, а потому итерация по $\mathbf{w}_{MP}$ представляет собой поиск единственного $\mathbf{w}_{MP}$ по найденному значению $\mathbf{A}$. Выбираем начальное приближение для матрицы $\mathbf{w}_{MP}$. В качестве такового можно взять, например, оценку максимума правдоподобия, что соответствует $\mathbf{A} = \boldsymbol{O}$. Далее по очереди выполняем итерации по $\mathbf{A}$ при фиксированном $\mathbf{w}_{MP}$ с предыдущей итерации и по $\mathbf{w}_{MP}$ при фиксированной $\mathbf{A}$ с предыдущей итерации. Отметим, что если наблюдается сходимост итераций, то в силу оптимальности найденной $\mathbf{A}$ при фиксированном $\mathbf{w} = \mathbf{w}_{MP}$ и

в силу того, что $\mathbf{w}_{MP}$ по построению есть оценка максимума апостериорной вероятности, найденное значение $\mathbf{A}$ есть решение (2.9).

Итерация по A (фиксированный $\mathbf{w}_{MP}$)

При фиксированном $\mathbf{w}_{MP}$ задача (2.9) эквивалентна следующей задаче нахождения $\mathbf{A}$

$$l_{\boldsymbol{\gamma}}(\mathbf{A}) = -\tfrac{1}{2}\log\det\mathbf{A} + \tfrac{1}{2}\mathbf{w}_{MP}^{\intercal}\mathbf{A}\mathbf{w}_{MP} + \tfrac{1}{2}\log\det(\mathbf{X}^{\intercal}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{X} + \mathbf{A}) \to \min_{\mathbf{A}\in\mathcal{M}}. \quad (2.10)$$

Далее рассмотрим два случая $\mathcal{M} = \mathcal{M}_{diag}$, $\mathcal{M} = \mathcal{M}_{full}$.

Случай $\mathcal{M} = \mathcal{M}_{diag}$.

В рассматриваемом случае матрица $\mathbf{A}$ имеет вид $\mathbf{A} = \operatorname{diag}(\alpha_1, \ldots, \alpha_n)$. Тогда имеем

$$\frac{\partial l_{\boldsymbol{\gamma}}(\mathbf{A})}{\partial\alpha_j} = \frac{1}{2\alpha_j} - \tfrac{1}{2}w_j^2 - \tfrac{1}{2}H_{jj} = 0.$$

Отсюда получаем выражение для α_j (домножение на α_j произведено для ускорения сходимости к ∞ для избыточных признаков)

$$\alpha_j^{new} = \frac{1 - \alpha_j^{old}H_{jj}^{old}}{w_j^2}. \quad (2.11)$$

Отметим, что при фиксированном $\mathbf{w}_{MP}$ для второй производной $\partial^2 l_{\boldsymbol{\gamma}}(\mathbf{A})/\partial\alpha_j^2$ следующее выражение.

$$\frac{\partial^2 l_{\boldsymbol{\gamma}}(\mathbf{A})}{\partial\alpha_j^2} = \tfrac{1}{2}\left((\mathbf{A}^{-1})_{jj}^2 - ((\mathbf{A} + \mathbf{X}^{\intercal}\mathbf{R}_{\gamma}\mathbf{X})^{-1})_{jj}^2\right) \geq 0,$$

откуда оптимизируемая функция является выпуклой и найденный локальный минимум является одновременно и глобальным.

Случай $\mathcal{M} = \mathcal{M}_{full}$

$$\frac{\partial l_{\boldsymbol{\gamma}}(\mathbf{A})}{\partial\mathbf{A}} = \tfrac{1}{2}\mathbf{A}^{-1} - \tfrac{1}{2}\mathbf{w}_{MP}\mathbf{w}_{MP}^{\intercal} - \tfrac{1}{2}(\mathbf{X}^{\intercal}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{X} + \mathbf{A})^{-1} = \mathbf{0}.$$

Отсюда получаем (с учетом требования симметричности и положительной определенности матрицы $\mathbf{A}$)

$$\mathbf{A}^{new} = \left(\mathbf{w}_{MP}\mathbf{w}_{MP}^{\intercal} + (\mathbf{X}^{\intercal}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{X} + \mathbf{A}^{old})^{-1}\right)^{-1}. \quad (2.12)$$

Итерация по $\mathbf{w}_{MP}$ (фиксированная A).

При фиксированной матрице $\mathbf{A}$ производим пересчет оценки максимума апостериорной вероятности в соответствии с его определением, то есть решаем следующую задачу.

$$\tilde{l}_{\boldsymbol{\gamma}}(\mathbf{w}) = -\log_{\boldsymbol{\gamma}} p(\mathbf{y}|\mathbf{X}, \mathbf{w}) + \tfrac{1}{2}\mathbf{w}^{\intercal}\mathbf{A}\mathbf{w} \to \min_{\mathbf{w}}.$$

Минимизируемая функция в данной задаче является выпуклой, а потому имеет единственный минимум. Для его нахождения можно воспользоваться, например, методом Ньютона-Рафсона [21] или его демпфированной версией [**?**]. Приведем далее выражения для градиента и гессиана минимизируемой функции.

$$\frac{\partial \tilde{l}_{\boldsymbol{\gamma}}(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{A}\mathbf{w} - \sum_{i=1}^{m} \gamma_i \sigma(-y_i \mathbf{w}^{\intercal}\mathbf{x}_i) y_i \mathbf{x}_i,$$

$$\frac{\partial^2 \tilde{l}(\mathbf{w})}{\partial \mathbf{w}^2} = \mathbf{X}^{\intercal}\mathbf{R}_{\boldsymbol{\gamma}}\mathbf{X} + \mathbf{A}.$$

Вариационная оценка матрицы ковариаций

Рассмотрим второй метод нахождения оценки матрицы $\mathbf{A}$, то есть приближенного решения задачи (2.5). Этот метод основан на построении вариационной нижней оценки к функции правдоподобия [21]. Далее приведем определение.

Определение 20. *Вариационной оценкой функции* $f(x) : \mathbb{R} \to \mathbb{R}$ называется функция $g(x, \xi) : \mathbb{R}^2 \to \mathbb{R}$, такая что:

1. $f(x) \geq g(x, \xi) \quad \forall\, x, \xi$
2. $f(\xi) = g(\xi, \xi) \quad \forall\, \xi.$

Для сигмоиды существует вариационная нижняя оценка вида [21]

$$\sigma(x) \geq \sigma(\xi) \exp\{(x - \xi)/2 - \lambda(\xi)(x^2 - \xi^2)\}, \text{ где} \tag{2.13}$$

$$\lambda(\xi) = \frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right]. \tag{2.14}$$

Воспользуемся ей для оценки интеграла в (2.5). Введем, как и ранее, весовой вектор объектов $\boldsymbol{\gamma} = [1, \ldots, 1]^{\intercal} \in \mathbb{R}^m$. Правдоподобие $p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X}, \mathbf{w})$ имеет вид

$$p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{m} p(y_i|\mathbf{x}_i, \mathbf{w})^{\gamma_i} = \prod_{i=1}^{m} \sigma(y_i \mathbf{w}^{\intercal}\mathbf{x}_i)^{\gamma_i}. \tag{2.15}$$

Воспользуемся вариационной нижней оценкой (2.13) для сигмоиды и получим вариационную нижнюю оценку для функции правдоподобия (2.15)

$$p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X}, \mathbf{w}) \geq \prod_{i=1}^{m} \sigma^{\gamma_i}(\xi_i) \exp\left[\gamma_i \left(\frac{y_i \mathbf{w}^{\intercal}\mathbf{x}_i - \xi_i}{2} - \lambda(\xi_i)((\mathbf{w}^{\intercal}\mathbf{x}_i)^2 - \xi_i^2)\right)\right]. \tag{2.16}$$

В (2.16) введен вектор вариационных параметров $\boldsymbol{\xi}$ размера $n \times 1$. Пользуясь вариационной нижней оценкой для правдоподобия (2.16), получим нижнюю оценку (уже не вариационную) для обоснованности модели $p(\mathbf{y}|\mathbf{X}, \mathbf{A})$

$$p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X}, \mathbf{A}) = \int p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\mathbf{A}) d\mathbf{w} \geq$$

$$\int \prod_{i=1}^{m} \sigma^{\gamma_i}(\xi_i) \exp\left[\gamma_i \left(\frac{y_i \mathbf{w}^{\intercal}\mathbf{x}_i - \xi_i}{2} - \lambda(\xi_i)((\mathbf{w}^{\intercal}\mathbf{x}_i)^2 - \xi_i^2)\right)\right] \frac{\sqrt{\det \mathbf{A}}}{(2\pi)^{m/2}} \exp(-1/2 \mathbf{w}^{\intercal}\mathbf{A}\mathbf{w}) d$$

(2

Выпишем выражение для логарифма подинтегрального выражения $\log I(\mathbf{w})$ в (2.17)

$$\log I_{\boldsymbol{\gamma}}(\mathbf{w}) = -\tfrac{1}{2}\mathbf{w}^{\intercal}\mathbf{A}\mathbf{w}+\tfrac{1}{2}\log\det\mathbf{A}+\sum_{i=1}^{m}\gamma_i\log\sigma(\xi_i)+\sum_{i=1}^{m}\gamma_i\frac{y_i\mathbf{w}^{\intercal}\mathbf{x}_i-\xi_i}{2}-\gamma_i\lambda(\xi_i)\left[(\mathbf{w}^{\intercal}\mathbf{x}\right.$$
$$-\tfrac{1}{2}(\mathbf{w}-\mathbf{w}_0)^{\intercal}\mathbf{A}'(\mathbf{w}-\mathbf{w}_0)+C.$$

Введем обозначение $\mathbf{v} = 1/2\sum_{i=1}^{m}\gamma_i y_i\mathbf{x}_i$, тогда имеем

$$\mathbf{A}' = \mathbf{A} + 2\sum_{i=1}^{m}\gamma_i\lambda(\xi_i)\mathbf{x}_i\mathbf{x}_i^{\intercal},\ \mathbf{w}_0 = \mathbf{A}'^{-1}\mathbf{v},$$

$$C = \tfrac{1}{2}\log\det\mathbf{A}+\sum_{i=1}^{m}\gamma_i\log\sigma(\xi_i)-\tfrac{1}{2}\sum_{i=1}^{m}\gamma_i\xi_i+\sum_{i=1}^{m}\gamma_i\lambda(\xi_i)\xi_i^2+\tfrac{1}{2}\mathbf{v}^{\intercal}\mathbf{A}'^{-1}\mathbf{v}-\tfrac{n}{2}\log(2\pi).$$

Пользуясь выражением для константы многомерного нормального распределения имеем следующую нижнюю оценку для логарифма обоснованности

$$\log p(\mathbf{y}|\mathbf{X},\mathbf{A}) \geq \tfrac{1}{2}\log\det\mathbf{A}-\tfrac{1}{2}\log\det\mathbf{A}'+\sum_{i=1}^{m}\gamma_i\log\sigma(\xi_i)-\tfrac{1}{2}\sum_{i=1}^{m}\gamma_i\xi_i+\sum_{i=1}^{m}\gamma_i\lambda(\xi_i)\xi_i^2+$$
$$\tfrac{1}{2}\mathbf{v}^{\intercal}\mathbf{A}'^{-1}\mathbf{v} = b_{\boldsymbol{\gamma}}(\mathbf{A},\,\boldsymbol{\xi}). \quad (2.18)$$

Заменим теперь задачу максимизации обоснованности (2.5) задачей максимизации нижней оценки на ее логарифма

$$[\mathbf{A},\boldsymbol{\xi}] = \arg\max_{\boldsymbol{\xi},\,\mathbf{A}\in\mathcal{M}}\left[\tfrac{1}{2}\log\det\mathbf{A}-\tfrac{1}{2}\log\det\mathbf{A}'+\sum_{i=1}^{m}\gamma_i\log\sigma(\xi_i)-\tfrac{1}{2}\sum_{i=1}^{m}\gamma_i\xi_i+\sum_{i=1}^{m}\gamma_i\lambda(\xi_i)\xi\right. \quad (2.19)$$

Опишем теперь итерационный алгоритм решения задачи (2.19). Выбираем начальное приближение для матрицы $\mathbf{A}$. Далее по очереди выполняем итерации по $\boldsymbol{\xi}$ при фиксированной $\mathbf{A}$ с предыдущей итерации и по $\mathbf{A}$ при фиксированном $\boldsymbol{\xi}$ с предыдущей итерации.

Итерация по $\boldsymbol{\xi}$ (фиксированная A).

Для получения формулы пересчета $\boldsymbol{\xi}$ выпишем условие оптимальности первого порядка для $b_{\boldsymbol{\gamma}}(\mathbf{A},\,\boldsymbol{\xi})$.

$$\frac{\partial b_{\boldsymbol{\gamma}}}{\partial\xi_i} = -\tfrac{\gamma_i}{2}+\gamma_i\lambda'(\xi_i)\xi_i^2+2\gamma_i\lambda(\xi_i)\xi_i+\tfrac{1}{2}\mathbf{v}^{\intercal}\frac{\partial A'^{-1}}{\partial\xi_i}\mathbf{v}+\gamma_i\frac{\sigma'(\xi_i)}{\sigma(\xi_i)}-\gamma_i\lambda'(\xi_i)\,tr(\mathbf{A}'^{-1}\mathbf{x}_i\mathbf{x}_i^{\intercal}) = 0.$$

Преобразуем полученное условие и получаем

$$-\gamma_i\lambda'(\xi_i)\mathbf{x}_i^\intercal\mathbf{A}'^{-1}\mathbf{x}_i+\gamma_i\lambda'(\xi_i)\xi_i^2-\tfrac{1}{2}\mathbf{v}^\intercal\mathbf{A}'^{-1}\frac{\partial\mathbf{A}'}{\partial\xi_i}\mathbf{A}'^{-1}\mathbf{v}+\gamma_i\left(\sigma(-\xi_i)-\tfrac{1}{2}+2\lambda(\xi_i)\xi_i\right)=0,$$

$$\gamma_i\underbrace{\left(\sigma(-\xi_i)-\tfrac{1}{2}+2\lambda(\xi_i)\xi_i\right)}_{=0}+\gamma_i\lambda'(\xi_i)(\xi_i^2-\mathbf{x}_i^\intercal\mathbf{A}'^{-1}\mathbf{x}_i-\mathbf{v}^\intercal\mathbf{A}'^{-1}\mathbf{x}_i\mathbf{x}_i^\intercal\mathbf{A}'^{-1}\mathbf{v})=0,$$

$$\xi_i^2=\mathbf{x}_i^\intercal\left(\mathbf{A}'^{-1}+(\mathbf{A}'^{-1}\mathbf{v})(\mathbf{A}'^{-1}\mathbf{v})^\intercal\right)_{old}\mathbf{x}_i=\mathbf{x}_i^\intercal\left(\mathbf{A}'^{-1}+\mathbf{w}_0\mathbf{w}_0^\intercal\right)_{old}\mathbf{x}_i.$$

Итерация по A (фиксированный $\boldsymbol{\xi}$)

При фиксированном $\boldsymbol{\xi}$ задача (2.19) эквивалентна следующей задаче нахождения $\mathbf{A}$

$$\tilde{b}_{\boldsymbol{\gamma}}(\mathbf{A})=\tfrac{1}{2}\log\det\mathbf{A}-\tfrac{1}{2}\log\det\mathbf{A}'+\tfrac{1}{2}\mathbf{v}^\intercal\mathbf{A}'^{-1}\mathbf{v}\to\max_{\mathbf{A}\in\mathcal{M}}.$$

Далее рассмотрим два разных множества допустимых матриц $\mathcal{M}$: диагональные с положительными диагональными элементами и симметричные положительно определенные матрицы общего вида.

Случай $\mathcal{M}=\mathcal{M}_{diag}$

В рассматриваемом случае матрица $\mathbf{A}$ имеет вид $\mathbf{A}=\operatorname{diag}(\alpha_1,\ \ldots,\ \alpha_n)$. Тогда имеем

$$\frac{\partial b_{\boldsymbol{\gamma}}(\mathbf{A})}{\partial\alpha_j}=\tfrac{1}{2\alpha_j}-\tfrac{1}{2}\mathbf{A}'^{-1}_{jj}-\tfrac{1}{2}\mathbf{v}^\intercal\mathbf{A}'^{-1}\frac{\partial\mathbf{A}'}{\partial\alpha_j}\mathbf{A}'^{-1}\mathbf{v}=0.$$

Отсюда получаем выражение для α_j

$$\alpha_j^{new}=\frac{1}{(\mathbf{A}'^{-1})_{jj}^{\text{old}}+[(\mathbf{A}'^{-1}\mathbf{v})_j^{\text{old}}]^2}. \tag{2.20}$$

Случай $\mathcal{M}=\mathcal{M}_{full}$

$$\frac{\partial b_{\boldsymbol{\gamma}}(\mathbf{A})}{\partial\mathbf{A}}=\tfrac{1}{2}\mathbf{A}^{-1}-\tfrac{1}{2}\mathbf{A}'^{-1}+\tfrac{1}{2}\underbrace{\frac{\partial}{\partial\mathbf{A}}tr(\mathbf{A}'^{-1}\mathbf{v}\mathbf{v}^\intercal)}_{\mathbf{A}'^{-1}\mathbf{v}\mathbf{v}^\intercal\mathbf{A}'^{-1}}=\mathbf{0}.$$

Отсюда получаем (с учетом требования симметричности и положительной определенности матрицы $\mathbf{A}$)

$$\mathbf{A}^{new}=\left(\mathbf{I}+\mathbf{v}\mathbf{v}^\intercal\mathbf{A}'^{-1}_{\text{old}}\right)^{-1}_{\text{old}}. \tag{2.21}$$

Заметим, что положительная определенность матрицы $\mathbf{A}$, определенной в (2.21), следует из положительной определенности $\mathbf{A}'$ на предыдущей итерации и эквивалентной записи (2.21) в виде

$$\mathbf{A}^{-1}=\mathbf{A}'^{-1}+\mathbf{A}'^{-1}\mathbf{v}\mathbf{v}^\intercal\mathbf{A}'^{-1}.$$

EM-алгоритм для нахождения приближенной оценки максимума обоснованности для матрицы A с помощью вариационной нижней оценки.

Заметим, что решить задачу (2.19) можно было также с помощью EM-алгоритма, что несколько упростило бы выкладки, поскольку не потребовалось бы брать интеграл по $\mathbf{w}$. Для этого рассмотрим $\mathbf{w}$ в качестве скрытой переменной. Имеем

$$\log p_{\boldsymbol{\gamma}}(\mathbf{y}|\mathbf{X},\ \mathbf{A}) \geq L_{\boldsymbol{\gamma}}(q,\ \mathbf{A}) = \mathbb{E}_{q(\mathbf{w})} \log p_{\boldsymbol{\gamma}}(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) - \mathbb{E}_{q(\mathbf{w})} \log q(\mathbf{w}) \geq$$
$$\tilde{L}(q,\ \mathbf{A},\ \boldsymbol{\xi}) = \mathbb{E}_{q(\mathbf{w})} \log p_{LB}(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \boldsymbol{\xi}) - \mathbb{E}_{q(\mathbf{w})} \log q(\mathbf{w}),$$

где для получения $p_{LB}(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}, \boldsymbol{\xi})$ использована вариационная нижняя оценка для сигмоидной функции

$$\log p_{LB}(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}, \boldsymbol{\xi}) = \tfrac{1}{2} \log\det \mathbf{A} - \tfrac{1}{2}\mathbf{w}^\intercal \mathbf{A}\mathbf{w} + \sum_{i=1}^{m} \gamma_i \Big[\log \sigma(\xi_i) + \tfrac{y_i \mathbf{w}^\intercal \mathbf{x}_i - \xi_i}{2} - \lambda(\xi_i)((\mathbf{w}^\intercal \mathbf{x}$$

Далее с помощью EM-алгоритма решаем задачу

$$\tilde{L}_{\boldsymbol{\gamma}}(q,\ \mathbf{A},\ \boldsymbol{\xi}) \to \max_{q,\ \mathbf{A},\ \boldsymbol{\xi}}. \tag{2.22}$$

E-шаг. На данном шаге вычисляем аппроксимацию апостериорного распределения на вектор скрытых переменных $\mathbf{w}$ $q(\mathbf{w})$, решая задачу максимизации (2.22) по q. С учетом того, что $\log p_{LB}(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}, \boldsymbol{\xi})$ квадратична по $\mathbf{w}$, $q(\mathbf{w})$ есть нормальное распределение. Для параметров этого распределения имеем

$$q(\mathbf{w}) = N(\mathbf{w}|\mathbf{w}_0,\ \mathbf{A}'^{-1}),\ \text{где}$$

$$\mathbf{v} = \tfrac{1}{2}\sum_{i=1}^{m} \gamma_i y_i \mathbf{x}_i,\ \mathbf{A}' = \mathbf{A} + 2\sum_{i=1}^{m} \gamma_i \lambda(\xi_i)\mathbf{x}_i\mathbf{x}_i^\intercal,\ \mathbf{w}_0 = \mathbf{A}'^{-1}\mathbf{v}.$$

M-шаг. На данном шаге решаем следующую задачу максимизации.

$$\mathbb{E}_{q(\mathbf{w})} \log p_{LB}(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \boldsymbol{\xi}) \to \max_{\mathbf{A},\ \boldsymbol{\xi}}, \tag{2.23}$$

где $q(\mathbf{w})$ получено на E-шаге и предполагается фиксированным.

Итерация по $\boldsymbol{\xi}$ (фиксированная A).

Для получения формулы пересчета $\boldsymbol{\xi}$ выпишем условие оптимальности первого порядка по ξ_i.

$$\gamma_i \left(\sigma(-\xi_i) - \tfrac{1}{2} - \lambda'(\xi_i)[\mathbf{x}_i^\intercal \mathbb{E}_q \mathbf{w}\mathbf{w}^\intercal \mathbf{x}_i - \xi_i^2] + 2\lambda(\xi_i)\xi_i\right) = 0.$$

Преобразуем полученное условие и получаем

$$(\xi_i^{\text{new}})^2 = \mathbf{x}_i^\intercal (\mathbf{w}_0\mathbf{w}_0^\intercal + \mathbf{A}'^{-1})_{\text{new}} \mathbf{x}_i.$$

Итерация по A (фиксированный $\boldsymbol{\xi}$)

При фиксированном $\boldsymbol{\xi}$ задача (2.23) эквивалентна следующей задаче нахождения $\mathbf{A}$

$$-\tfrac{1}{2}\, tr(\mathbf{A}\mathbb{E}_q \mathbf{w}\mathbf{w}^\intercal) + \tfrac{1}{2}\log\det\mathbf{A} \to \max_{\mathbf{A}\in\mathcal{M}}.$$

Далее рассмотрим два разных множества допустимых матриц $\mathcal{M} = \mathcal{M}_{diag}$ и $\mathcal{M} = \mathcal{M}_{full}$

Случай $\mathcal{M} = \mathcal{M}_{diag}$.

В рассматриваемом случае матрица $\mathbf{A}$ имеет вид $\mathbf{A} = \mathrm{diag}(\alpha_1,\ \ldots,\ \alpha_n)$. Тогда получаем выражение для α_j

$$\alpha_j^{new} = \frac{1}{(\mathbb{E}_q \mathbf{w}\mathbf{w}^\intercal)_{jj}^{\text{old}}} = \frac{1}{(\mathbf{w}_0\mathbf{w}_0^\intercal + \mathbf{A}'^{-1})_{\text{old}}}. \tag{2.24}$$

Случай $\mathcal{M} = \mathcal{M}_{full}$.

Условие оптимальности первого порядка имеет вид

$$-\tfrac{1}{2}\mathbb{E}_q \mathbf{w}\mathbf{w}^\intercal + \tfrac{1}{2}\mathbf{A}^{-1} = \mathbf{0}.$$

Отсюда получаем (с учетом требования симметричности и положительной определенности матрицы $\mathbf{A}$)

$$\mathbf{A}^{new} = (\mathbf{w}_0\mathbf{w}_0^\intercal + \mathbf{A}'^{-1})_{\text{old}}^{-1}. \tag{2.25}$$

Таким образом, были получены формулы для нахождения приближения для оценки максимума обоснованности для матрицы $\mathbf{A}$ двумя способами. Оценки были получены как в множестве диагональных матриц $\mathcal{M}_{diag}$, что позволяет отбирать признаки, так и в множестве $\mathcal{M}_{full}$. Хотя оценка максимума обоснованности $\mathbf{A}^* \in \mathcal{M}_{full}$ имеет недиагональные элементы, оказывается, что их значения тривиальны и не связаны с истинными взаимосвязями между признаками, а соответствующая матрица $\mathbf{A}^* \in \mathcal{M}_{full}$ является асимптотически вырожденной, что устанавливает следующая теорема.

Теорема 2 (Адуенко, 2016)**.** Пусть имеется одиночная логистическая модель, заданная совместным правдоподобием

$$p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) = \prod_{i=1}^{m} \sigma(y_i \mathbf{w}^\intercal \mathbf{x}_i) N(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1}),$$

где $y_i \in \{-1,\ 1\}$, $\mathbf{x}_i \in \mathbb{R}^2$, то есть признаковое пространство имеет размерность $n = 2$. Пусть также $\mathbf{w} = [w_1,\ w_2]$, $w_1,\ w_2 \neq 0$.
Обозначим

$$\boldsymbol{\Sigma} = \mathbf{H}^{-1} = \mathbf{X}^\intercal \mathbf{R}\mathbf{X} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},\ \mathbf{A} = \begin{pmatrix} s_1^2 & \kappa s_1 s_2 \\ \kappa s_1 s_2 & s_2^2 \end{pmatrix},$$

где $\mathbf{H}$ есть гессиан $-\log p(\mathbf{y}|\mathbf{X},\ \mathbf{w})$ в точке максимума апостериорной вероятности $\mathbf{w}_{MP} = \arg\max_{\mathbf{w}} p(\mathbf{y},\ \mathbf{w}|\mathbf{X})$, а $\mathbf{R} = \mathrm{diag}(\sigma(y_i\mathbf{w}_{MP}^{\intercal}\mathbf{x}_i)\sigma(-y_i\mathbf{w}_{MP}^{\intercal}\mathbf{x}_i))$,

Тогда если при $m \to \infty$ выполнено

$$\sigma_1^2,\ \sigma_2^2 \overset{\text{п.н.}}{\to} \infty, \tag{2.26}$$

$$\mathbb{P}(\omega :\ \exists\, c_\omega > 0,\ \exists\, m_\omega :\ \forall\, m \geq m_\omega\ 1 - \rho_m^2 \geq c_\omega) = 1, \tag{2.27}$$

то $s_1^*,\ s_2^* \overset{\text{п.н.}}{\to} \infty$, $\kappa^* \overset{\text{п.н.}}{\to} -\mathrm{sign}(w_1 w_2)$.

Доказательство. Покажем, что оценка матрицы $\mathbf{A}$ с помощью максимизации обоснованности является асимптотически вырожденной и определяемая корреляция между весами признаков при достаточное большой выборке зависит не от истинной корреляции между признаками и их весами, а лишь от совпадения или несовпадения знаков весов. В случае совпадающих знаков корреляция стремится 1, в случае несовпадающих она стремится к -1.

Для доказательства рассмотрим случай, когда в модели есть два признака, то есть $n = 2$, причем истинный вектор весов есть $\mathbf{w} = [1 + \varepsilon_1,\ 1 + \varepsilon_2]$, где $\varepsilon_1,\ \varepsilon_2$ – некоторые постоянные, $|\varepsilon_1|,\ |\varepsilon_2| < 1/2$. Как будет показано в дальнейшем, утверждение для любого другого произвольного вектора весов можно получить путем перенормировки признаков.

Введем обозначения

$$\mathbf{A} = \begin{pmatrix} s_1^2 & \kappa s_1 s_2 \\ \kappa s_1 s_2 & s_2^2 \end{pmatrix},\ \mathbf{\Sigma} = \mathbf{X}^{\intercal}\mathbf{R}\mathbf{X} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Пусть на некоторой итерации оценки истинного вектора весов $\hat{\mathbf{w}}$ и матрицы $\mathbf{A}$ получено некоторое значение $\hat{\mathbf{w}}$. Без ограничения общности считаем, что $\hat{\mathbf{w}} = [1,\ 1]^{\intercal}$ (иначе, как указывалось ранее сделаем перенормировку признаков). Заметим, что при достаточно большом числе объектов в выборке наблюдается сходимость $\hat{\mathbf{w}}$ к $\mathbf{w}$ [58, 59]. Поэтому знаки весов в $\hat{\mathbf{w}}$ и $\mathbf{w}$ совпадают с некоторого размера выборки, что позволяет выполнить перенормировку признаков.

Рассмотрим заключительный шаг итерационного процесса (2.10), заключающийся в оптимизации по $\mathbf{A}$.

$$f(\mathbf{A}) = \hat{\mathbf{w}}^{\intercal}\mathbf{A}\hat{\mathbf{w}} - \log|\mathbf{A}| + \log|\mathbf{\Sigma} + \mathbf{A}| \to \min_{\mathbf{A}}. \tag{2.28}$$

Рассматриваем далее множество элементарных исходов Ω_1, для которых выполнено $\forall\omega \in \Omega_1\ \sigma_1^2(\omega) \to \infty,\ \sigma_2^2(\omega) \to \infty$ при $m \to \infty$ и $\exists\, m_\omega :\ \forall\, m \geq m_\omega\ 1 - \rho_m^2 \geq c > 0$. По условию теоремы $\mathbb{P}(\Omega_1) = 1$. Тогда докажем следующее утверждение.

Утверждение 1. Пусть задача решается в условиях $\mathbf{A}$ – неотрицательно определенная матрица и $s_1,\ s_2 \leq C$, где C – некоторая положительная константа(достаточно большая, определяемая из технических соображений). Тогда если $\sigma_1^2,\ \sigma_2^2 \to \infty$ при $m \to \infty$, а $1 - \rho^2 \geq c > 0$, начиная с некоторого m_0, то,

начиная с некоторого σ_0^2 при выполнении $\sigma_1,\ \sigma_2 \geq \sigma_0$ оптимальным решением рассматриваемого шага оптимизационного процесса будет

$$s_1^* = s_2^* = C,\ \kappa^* \to \frac{1-\sqrt{1+4C^4}}{2C^2} \text{ при } \sigma_1,\ \sigma_2 \to \infty.$$

Доказательство. Существует три разных варианта расположения оптимума в рассматриваемой задаче с ограничениями.

1. Оптимум и по s_1, и по s_2 находится во внутренней точке, то есть $0 < s_1^*,\ s_2^* < C$;

2. Оптимум находится на границе и по s_1, и по s_2, то есть $s_1^* = s_2^* = C$;

3. Оптимум по s_1 находится на границе, по s_2 во внутренней точке, то есть $s_1^* = C,\ 0 < s_2^* < C$ (случай, когда оптимум по s_2 на границе, а по s_1 во внутренней точке, полностью аналогичен).

В вариантах выше учтено, что при $s_1 = 0$ или $s_2 = 0$ $f(\mathbf{A}) = +\infty$, что не является минимально возможным значением.

Рассмотрим сначала случай 1. С учетом неактивности ограничений типа неравенства на s_1 и s_2 имеем следующую систему уравнений.

$$\begin{cases} \dfrac{\partial f}{\partial s_1} = 0, \\ \dfrac{\partial f}{\partial s_2} = 0, \\ \dfrac{\partial f}{\partial \kappa} = 0. \end{cases}$$

Получим выражение для детерминанта $|\mathbf{A}+\mathbf{\Sigma}|$ и запишем условия оптимальности в явном виде.

$$|\mathbf{A}+\mathbf{\Sigma}| = \left| \begin{pmatrix} \sigma_1^2 + s_1^2 & \kappa s_1 s_2 + \rho\sigma_1\sigma_2 \\ \kappa s_1 s_2 + \rho\sigma_1\sigma_2 & \sigma_2^2 + s_2^2 \end{pmatrix} \right| =$$
$$= \sigma_1^2\sigma_2^2(1-\rho^2)\left[1 + \frac{s_1^2}{\sigma_1^2(1-\rho^2)} + \frac{s_2^2}{\sigma_2^2(1-\rho^2)} + \frac{s_1^2 s_2^2(1-\kappa^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)} - \frac{2\kappa\rho s_1 s_2}{\sigma_1\sigma_2(1-\rho^2)}\right].$$

Тогда условия оптимальности имеют вид

$$\frac{1}{2}\frac{\partial f}{\partial s_1} = s_1+s_2\kappa-\frac{1}{s_1}+\frac{1}{1+\dfrac{s_1^2}{\sigma_1^2(1-\rho^2)}+\dfrac{s_2^2}{\sigma_2^2(1-\rho^2)}+\dfrac{s_1^2s_2^2(1-\kappa^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)}-\dfrac{2\kappa\rho s_1s_2}{\sigma_1\sigma_2(1-\rho^2)}}\cdot$$
$$\cdot\left(\frac{s_1}{\sigma_1^2(1-\rho^2)}+\frac{s_1s_2^2(1-\kappa^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)}-\frac{\kappa\rho s_2}{\sigma_1\sigma_2(1-\rho^2)}\right)=0,$$

$$\frac{1}{2}\frac{\partial f}{\partial s_2} = s_2+s_1\kappa-\frac{1}{s_2}+\frac{1}{1+\dfrac{s_1^2}{\sigma_1^2(1-\rho^2)}+\dfrac{s_2^2}{\sigma_2^2(1-\rho^2)}+\dfrac{s_1^2s_2^2(1-\kappa^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)}-\dfrac{2\kappa\rho s_1s_2}{\sigma_1\sigma_2(1-\rho^2)}}\cdot$$
$$\cdot\left(\frac{s_2}{\sigma_2^2(1-\rho^2)}+\frac{s_1^2s_2(1-\kappa^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)}-\frac{\kappa\rho s_1}{\sigma_1\sigma_2(1-\rho^2)}\right)=0,$$

$$\frac{1}{2}\frac{\partial f}{\partial \kappa} = s_1s_2+\frac{\kappa}{1-\kappa^2}-\frac{1}{1+\dfrac{s_1^2}{\sigma_1^2(1-\rho^2)}+\dfrac{s_2^2}{\sigma_2^2(1-\rho^2)}+\dfrac{s_1^2s_2^2(1-\kappa^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)}-\dfrac{2\kappa\rho s_1s_2}{\sigma_1\sigma_2(1-\rho^2)}}\cdot$$
$$\cdot\left(\frac{s_1^2s_2^2\kappa}{\sigma_1^2\sigma_2^2(1-\rho^2)}-\frac{\rho s_1s_2}{\sigma_1\sigma_2(1-\rho^2)}\right).$$

Рассмотрим первые два условия отдельно

$$s_1+s_2\kappa-\frac{1}{s_1}+\alpha(s_1,\ s_2,\ \kappa)=0, \tag{2.29}$$

$$s_2+s_1\kappa-\frac{1}{s_2}+\beta(s_1,\ s_2,\ \kappa)=0. \tag{2.30}$$

Из (2.29) и (2.30) с учетом $s_1,\ s_2>0$ имеем

$$s_1^2-s_2^2+\alpha(s_1,\ s_2,\ \kappa)s_1-\beta(s_1,\ s_2,\ \kappa)s_2=0.$$

Учитывая, что $|\alpha(s_1,\ s_2,\ \kappa)|,\ |\beta(s_1,\ s_2,\ \kappa)|\le\dfrac{2C}{c\sigma^2}+\dfrac{C^3}{c\sigma^3}\to 0$ при $\sigma^2=\min(\sigma_1^2,\ \sigma_2^2)\to\infty$, получаем, что

$$s_2=s_1+\delta=s+\delta(\kappa,\ s_1),\ \text{где}\ \max_{\kappa,\ s_1}|\delta(\kappa,\ s_1)|\to 0\ \text{при}\ \sigma^2\to\infty. \tag{2.31}$$

Подставим (2.31) в условие оптимальности по κ, получим

$$s(s+\delta(\kappa,\ s))+\frac{\kappa}{1-\kappa^2}+\gamma(s,\ s+\delta(\kappa,\ s),\ \kappa)=0. \tag{2.32}$$

С учетом $s\le C$, $\max_{\kappa,\ s_1}|\delta(\kappa,\ s_1)|\to 0 \text{при} \sigma^2\to\infty$ и $|\gamma(s_1,\ s_2,\ \kappa)|\le\dfrac{C^4}{c\sigma^4}+\dfrac{C^2}{c\sigma^2}\to 0$ при $\sigma^2\to\infty$ получаем, что

$$-s^2-\delta_1(\kappa,\ s)\le\frac{\kappa}{1-\kappa^2}\le-s^2+\delta_1(\kappa,\ s),$$

где $\max_{\kappa,\, s} |\delta_1(\kappa,\ s)| \to 0$ при $\sigma^2 \to \infty$. Тогда с учетом строгого возрастания $\dfrac{\kappa}{1-\kappa^2}$ по κ с производной ≥ 1 для любого значения κ имеем единственность решения (2.32) при $\sigma^2 = \sigma_{01}^2$, причем

$$\kappa = \frac{1-\sqrt{1+4s^4}}{2s^2} + \delta_2(s), \text{ где } \max_s |\delta_2(s)| \to 0 \text{ при } s \to \infty. \tag{2.33}$$

Вернемся теперь к условию оптимальности по s_1, получим:

$$s + (s + \delta(\kappa,\ s))\kappa - \frac{1}{s} + \alpha(s,\ s + \delta(\kappa,\ s),\ \kappa) = 0.$$

Отсюда имеем

$$\kappa = -1 + \frac{\delta(\kappa,\ s)}{s + \delta(\kappa,\ s)} + \frac{1}{s(s + \delta(\kappa,\ s))} - \frac{\alpha(s,\ s + \delta(\kappa,\ s),\ \kappa)}{s + \delta(\kappa,\ s)},$$

что, начиная с некоторого $\sigma^2 \geq \sigma_{02}^2$ выполнено

$$\kappa \geq -1 + \frac{1}{s^2} - \frac{1}{10s^2} = -1 + \frac{0.9}{s^2}. \tag{2.34}$$

Аналогично, начиная с некоторого $\sigma^2 \geq \sigma_{03}^2$, из (2.33) имеем

$$\kappa \leq -1 + \frac{1}{2s^2} + \frac{1}{10s^2} = -1 + \frac{0.6}{s^2}.$$

Отсюда с учетом (2.34) при $\sigma^2 \geq \sigma_0^2 = \max(\sigma_{01}^2,\ \sigma_{02}^2,\ \sigma_{03}^2)$ имеем противоречие, так как κ не может удовлетворять двум полученным неравенствам одновременно.

Таким образом, начиная с некоторого $\min(\sigma_1^2,\ \sigma_2^2) = \sigma^2 \geq \sigma_0^2$ случай 1 реализоваться не может.

Пусть теперь реализуется случай 3. Без ограничения общности считаем, что $s_1^* = C$, $0 < s_2^* < C$. Условия оптимальности по s_2 и κ записываются в виде

$$Cs_2 + \frac{\kappa}{1-\kappa^2} + \gamma(C,\ s_2,\ \kappa) = 0, \tag{2.35}$$

$$s_2 + \kappa C - \frac{1}{s_2} + \beta(C,\ s_2,\ \kappa) = 0. \tag{2.36}$$

С учетом $s_2 > 0$ (а, значит, и $\kappa < 0$, начиная с некоторого $\sigma^2 \geq \sigma_{01}^2$), умножая первое из равенств на κ, а второе на s_2, получим

$$s_2^2 = \frac{1}{1-\kappa^2} + \gamma(C,\ s_2,\ \kappa)\kappa - \beta(C,\ s_2,\ \kappa)s_2. \tag{2.37}$$

Из (2.35) имеем

$$C^2 s_2^2 = \frac{\kappa^2}{(1-\kappa^2)^2} + \gamma^2(C,\ s_2,\ \kappa) + 2\gamma(C,\ s_2,\ \kappa)\frac{\kappa}{1-\kappa^2}. \tag{2.38}$$

Из (2.37) и (2.38) имеем

$$\frac{\kappa^2}{1-\kappa^2} = C^2+C^2\gamma(C,s_2,\kappa)\kappa(1-\kappa^2)-\beta(C,s_2,\kappa)s_2C^2(1-\kappa^2)-\gamma^2(C,s_2,\kappa)(1-\kappa^2)-2\kappa\gamma$$
$$= C^2 + \nu(s_2,\ \kappa),\ \text{где}\ \max_{s_2,\,\kappa}|\nu(\kappa,\ s_2)| \to 0\ \text{при}\ \sigma^2 \to \infty.$$

Отсюда, начиная с некоторого $\sigma^2 \geq \sigma_{02}^2$ выполнено

$$\frac{\kappa^2}{1-\kappa^2} \geq C^2 - \frac{1}{4} \Longleftrightarrow \frac{1}{1-\kappa^2} \geq C^2 + \frac{3}{4}.$$

Тогда с учетом $\max_{\kappa,\,s_2}|\gamma(C,\ s_2,\ \kappa)| \to 0$ и $\max_{\kappa,\,s_2}|\beta(C,\ s_2,\ \kappa)| \to 0$ из (2.37) имеем, что, начиная с некоторого $\sigma^2 \geq \sigma_{03}^2$ выполнено

$$s_2^2 \geq C^2 + \tfrac{1}{2} > C^2,$$

что приводит к противоречию. Отсюда случай 3 также не мог реализоваться при

$$\min(\sigma_1^2,\ \sigma_2^2) = \sigma^2 \geq \max(\sigma_{01}^2,\ \sigma_{02}^2,\ \sigma_{03}^2).$$

Таким образом, реализуется случай 2, то есть, начиная с некоторого $\sigma^2 \geq \sigma_{01}^2$ $s_1^* = s_2^* = C$. Условие оптимальности по κ тогда приобретает вид

$$C^2 + \frac{\kappa}{1-\kappa^2} + \gamma(C,\ C,\ \kappa) = 0 \tag{2.39}$$

С учетом $\max_\kappa|\gamma(C,\ C,\ \kappa)| \to 0$ и $\max_\kappa|\gamma'(C,\ C,\ \kappa)| \to 0$ при $\sigma^2 \to \infty$ и

$$\left(\frac{\kappa}{1-\kappa^2}\right)' \geq 1$$

получаем, что, начиная с некоторого $\sigma^2 \geq \sigma_{02}^2$ уравнение (2.39) имеет единственное решение относительно κ, причем

$$\kappa = \frac{1-\sqrt{1+4C^4}}{2C^2} + \delta_4,\ \text{где}\ \delta_4 \to 0\ \text{при}\ \sigma^2 \to \infty.$$

Таким образом, имеем требуемое утверждение при $\sigma^2 \geq \sigma_0^2 = \max(\sigma_{01}^2,\ \sigma_{02}^2)$. □

Следствие 1. В силу условия теоремы утверждение 1 выполнено $\forall\,\omega \in \Omega_1$, так как в качестве m_0 можно взять m_ω, а $\forall\,\omega \in \Omega_1\ \exists m' :\ \forall\,m \geq m'\ \min(\sigma_1^2(\omega),\ \sigma_2^2(\omega)) > \sigma_0^2$.

Рассмотрим теперь задачу оптимизации без ограничений $s_1,\ s_2 \leq C$. Из доказанного выше имеем, что

$$s_1,\ s_2 \to \infty\ \text{при}\ \sigma^2 \to \infty,$$

поскольку независимо от C, начиная с некоторого $\sigma^2 \geq \sigma_0^2$ $s_1^* = s_2^* = C$. Докажем тогда следующее утверждение.

Утверждение 2. Пусть задача решается в условиях **A** – неотрицательно определенная матрица. Тогда если $\sigma_1^2,\ \sigma_2^2 \to \infty$ при $m \to \infty$ и $1 - \rho_m^2 \geq c > 0$ при $m \geq m_0$, то $s_1^*,\ s_2^* \to \infty$, $\kappa^* \to -1$.

Доказательство. Утверждение касательно $s_1^*,\ s_2^* \to \infty$ уже показано. Рассмотрим теперь условие оптимальности по κ для задачи без ограничений.

$$s_1 s_2 + \frac{\kappa}{1-\kappa^2} - \frac{1}{1 + \frac{s_1^2}{\sigma_1^2(1-\rho^2)} + \frac{s_2^2}{\sigma_2^2(1-\rho^2)} + \frac{s_1^2 s_2^2(1-\kappa^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)} - \frac{2\kappa\rho s_1 s_2}{\sigma_1\sigma_2(1-\rho^2)}} \cdot \left(\frac{s_1^2 s_2^2 \kappa}{\sigma_1^2\sigma_2^2(1-\rho^2)} + \frac{\rho s_1 s_2}{\sigma_1\sigma_2(1-\rho^2)} \right) = 0.$$

Покажем, что третье слагаемое в выписанном выше уравнении больше, чем $C_1 - \frac{|\kappa|}{1-\kappa^2}$, где C_1 – некоторая константа, а тогда с учетом $s_1 s_2 \to +\infty$, получим, что

$$\frac{\kappa}{1-\kappa^2}(1+\alpha) \to -\infty, \text{ где } 1 \leq \alpha \geq -1 \Longleftrightarrow \kappa \to -1 \text{ при } \sigma^2 \to \infty.$$

Введем обозначения

$$1 + \frac{s_1^2}{\sigma_1^2(1-\rho^2)} + \frac{s_2^2}{\sigma_2^2(1-\rho^2)} + \frac{s_1^2 s_2^2(1-\kappa^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)} - \frac{2\kappa\rho s_1 s_2}{\sigma_1\sigma_2(1-\rho^2)} = \nu_1(s_1,\ s_2,\ \kappa),$$

$$\frac{s_1^2 s_2^2 \kappa}{\sigma_1^2\sigma_2^2(1-\rho^2)} + \frac{\rho s_1 s_2}{\sigma_1\sigma_2(1-\rho^2)} = \nu_2(s_1,\ s_2,\ \kappa).$$

Найдем условия на κ, при которых

$$\frac{\nu_2}{\nu_1} < \frac{|\kappa|}{1-\kappa^2}. \tag{2.40}$$

Если $\nu_2 < 0$, то (2.40) выполнено. Далее считаем, что $\nu_2 \geq 0$. Тогда или $\kappa > 0$, или $\rho > 0$. При этом максимальное значение $\dfrac{\nu_2}{\nu_1}$ достигается при $\rho > 0$ и $\kappa > 0$, поскольку, если $\kappa\rho < 0$ значение ν_1 возрастает, а $|\nu_2|$, напротив, уменьшается. Поэтому рассматриваем далее $\kappa,\ \rho > 0$ (при $\rho \geq 0$ неравенство (2.40) автоматически будет выполнено по крайней мере при тех же κ).

При $\rho > 0,\ \kappa > 0$ получаем, что неравенство (2.40) эквивалентно следующему

$$1 + \frac{s_1^2}{\sigma_1^2(1-\rho^2)} + \frac{s_2^2}{\sigma_2^2(1-\rho^2)} + \frac{s_1^2 s_2^2(1-\kappa^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)} - \frac{2\kappa\rho s_1 s_2}{\sigma_1\sigma_2(1-\rho^2)} > \frac{s_1^2 s_2^2(1-\kappa^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)} + \frac{\rho s_1 s_2}{\sigma_1\sigma_2(1-\rho^2)}.$$

Преобразуя это неравенство с учетом $1 - \rho^2 \geq c$ получаем

$$1 - \rho^2 + \frac{s_1^2}{\sigma_1^2} + \frac{\sigma_2^2}{\sigma_2^2} > \frac{\rho s_1 s_2}{\sigma_1\sigma_2}\left(\frac{1}{\kappa} + \kappa\right).$$

Полученное неравенство выполнено по крайней мере при

$$\frac{1}{\kappa} + \kappa \leq \frac{2}{\rho},$$

что дает

$$\kappa \geq \kappa_0 = \frac{1}{\rho}(1 - \sqrt{1-\rho^2}) \leq \frac{1+\sqrt{2-c}}{\sqrt{1-c}} < 1.$$

Таким образом, при $|\kappa| \geq \kappa_0$ требуемое выполнено. Покажем, что при $|\kappa| < \kappa_0$ слагаемое ограничено по модулю константой. Имеем

$$\nu_1 \geq (1-|\kappa|)\frac{2s_1s_2}{\sigma_1\sigma_2(1-\rho^2)} + \frac{s_1^2s_2^2(1-\kappa^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)},$$

$$|\nu_2| \leq \frac{s_1^2s_2^2|\kappa|}{\sigma_1^2\sigma_2^2(1-\rho^2)} + \frac{|\rho|s_1s_2}{\sigma_1\sigma_2(1-\rho^2)},$$

откуда

$$\left|\frac{\nu_2}{\nu_1}\right| \leq \max\left(\frac{|\rho|}{2(1-|\kappa|}, \frac{|\kappa|}{1-\kappa^2}\right) \leq \frac{1}{1-\kappa_0^2} < \infty.$$

Таким образом, имеем требуемое.

□

Следствие 2. В силу условия теоремы утверждение 2 выполнено $\forall \omega \in \Omega_1$, так как в качестве m_0 можно взять m_ω и $\forall\, \omega \in \Omega_1 :\ \min(\sigma_1^2(\omega),\, \sigma_2^2(\omega)) \to \infty$ при $m \to \infty$.

Так как в силу следствия 2 утверждение 2 имеет место для любого $\omega \in \Omega_1$, а $\mathbb{P}(\Omega_1) = 1$, получаем требуемое $s_1^*,\ s_2^* \overset{\text{п.н.}}{\to} \infty$, $\kappa^* \overset{\text{п.н.}}{\to} -\text{sign}(w_1w_2)$. □

Границы применимости теоремы.

Покажем теперь, что требования теоремы не являются ограничивающими и выполнены при достаточно широких предположениях. Считаем, что $\mathbf{w}_{MP} \overset{\text{п.н.}}{\to} \mathbf{w}$, где $\mathbf{w}$ есть истинный вектор параметров модели. Условия наличия такой сходимости приведены в (**??**). Рассмотрим множество элементарных исходов Ω_2, для которых выполнена сходимость $\mathbf{w}_{MP} \overset{\text{п.н.}}{\to} \mathbf{w}$, тогда $\mathbb{P}(\Omega_2) = 1$. Считаем далее, что множество значений признаков ограничено, то есть $\exists C :\ \|\mathbf{x}\|_\infty \leq C$.

Рассмотрим произвольный $\omega \in \Omega_2$. В силу сходимости $\mathbf{w}_{MP}^m(\omega) \to \mathbf{w}$ при $m \to \infty$, получаем

$$\forall\, \varepsilon > 0\, \exists\, m_\varepsilon :\ \forall\, m \geq m_\varepsilon\, \|\mathbf{w}_{MP}^m(\omega) - \mathbf{w}\| < \varepsilon. \tag{2.41}$$

Обозначим $\lambda(\mathbf{v},\, \mathbf{x}) = \sigma(\mathbf{v}^\intercal\mathbf{x})\sigma(-\mathbf{v}^\intercal\mathbf{x})$. С учетом ограниченности $\mathbf{x}$ имеем

$$\lambda(\mathbf{w},\, \mathbf{x}) \geq C_0 > 0. \tag{2.42}$$

С учетом непрерывности $\lambda(\mathbf{v},\ \mathbf{x})$ по аргументам, ограниченности множества значений признаков $\|\mathbf{x}\|_\infty \leq C$ и ограниченности $\|\mathbf{w}_{MP}^m(\omega)\|$ в силу наличия сходимости условие (2.41) можно переписать в виде

$$\forall\, \delta > 0\, \exists\, m_\delta :\ \forall\, m \geq m_\delta \left| \frac{\lambda(\mathbf{w}_{MP}^m(\omega),\ \mathbf{x})}{\lambda(\mathbf{w},\ \mathbf{x})} \right| \leq \delta. \tag{2.43}$$

Фиксируем далее произволное $1/2 > \delta > 0$ и рассматриваем $m \geq m_\delta$. Отметим, что для матрицы $\mathbf{\Sigma} = \mathbf{X}^\intercal\mathbf{R}\mathbf{X}$ справедливо следующее представление в силу диагональности матрицы $\mathbf{R}$.

$$\mathbf{X}^\intercal\mathbf{R}\mathbf{X} = \sum_{i=1}^m \lambda(\mathbf{w}_{MP}^m(\omega),\ \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\intercal = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. \tag{2.44}$$

Покажем теперь, что $\sigma_j^2(\omega) \to \infty$, $j = 1,\ 2$ при достаточно общих предположениях о процессе генерации выборки. С учетом (2.44) имеем

$$\sigma_j^2(\omega) = \sum_{i=1}^m \lambda(\mathbf{w}_{MP}^m(\omega),\ \mathbf{x}_i)x_{ij}^2 \geq (1-\delta)\sum_{i=1}^m \lambda^2(\mathbf{w},\ \mathbf{x}_i)x_{ij}^2 \geq C_0(1-\delta)\sum_{i=1}^m x_{ij}^2.$$

Таким образом, для того, чтобы получить $\sigma_j^2(\omega) \to \infty$ при $m \to \infty$ достаточно потребовать

$$\sum_{i=1}^m x_{ij}^2 \to \infty \text{ при } m \to \infty. \tag{2.45}$$

Здесь предполагается, как и ранее, что распределение $p(\mathbf{X})$ неизвестно. Если же считать, что объекты генерируются независимо из некоторого известного распределения, то есть

$$p(\mathbf{X}) = \prod_{i=1}^m p(\mathbf{x}_i),$$

то условие (2.45) в силу усиленного закона больших чисел можно заменить на

$$\mathbb{E}x_{ij}^2 > 0, \tag{2.46}$$

то есть требуется, чтобы признак j не являлся тождественно нулевым. Таким образом, при выполнении (**??**) или (2.46) соответственно получаем $\forall\, \omega \in \Omega_2$: $\sigma_1^2,\ \sigma_2^2 \to \infty$, то есть выполненность условия (2.26) теоремы.

Рассмотрим теперь недиагональный элемент матрицы $\mathbf{\Sigma}$ и покажем, что при достаточно общих предположениях выполнено условие (2.27) теоремы. Для

коэффициента корреляции $\rho^m_{MP}(\omega)$ имеем

$$\rho^2_{MP}(\omega) = \frac{\sum_{i=1}^m \lambda^2(\mathbf{w}^m_{MP}(\omega),\, \mathbf{x}_i)x^2_{i1}x^2_{i2} + \sum_{i,\,l=1,i\neq l}^m \lambda(\mathbf{w}^m_{MP}(\omega),\, \mathbf{x}_i)\lambda(\mathbf{w}^m_{MP}(\omega),\, \mathbf{x}_l)x_{i1}x$$
$$\frac{\sum_{i=1}^m \lambda^2(\mathbf{w}^m_{MP}(\omega),\, \mathbf{x}_i)x^2_{i1}x^2_{i2} + \sum_{i,\,l=1,i\neq l}^m \lambda(\mathbf{w}^m_{MP}(\omega),\, \mathbf{x}_i)\lambda(\mathbf{w}^m_{MP}(\omega),\, \mathbf{x}_l)|x_{i1}||x_{i2}||x_{l1}||x_{l2}|}{\sum_{i=1}^m \lambda(\mathbf{w}^m_{MP}(\omega),\, \mathbf{x}_i)x^2_{i1}\sum_{l=1}^m \lambda(\mathbf{w}^m_{MP}(\omega),\, \mathbf{x}_l)x^2_{l2}}$$
$$\left(\frac{1+\delta}{1-\delta}\right)^2 \rho^2(\omega), \text{ где}$$

$$\rho^2(\omega) = \frac{\sum_{i=1}^m \lambda^2(\mathbf{w},\, \mathbf{x}_i)x^2_{i1}x^2_{i2} + \sum_{i,\,l=1,i\neq l}^m \lambda(\mathbf{w},\, \mathbf{x}_i)\lambda(\mathbf{w},\, \mathbf{x}_l)|x_{i1}||x_{i2}||x_{l1}||x_{l2}|}{\sum_{i=1}^m \lambda(\mathbf{w},\, \mathbf{x}_i)x^2_{i1}\sum_{l=1}^m \lambda(\mathbf{w},\, \mathbf{x}_l)x^2_{l2}}. \tag{2.47}$$

Обозначим $\tilde{\mathbf{X}} = |||x_{ij}|||$, $i = \overline{1,\, m}$, $j = \overline{1,\, 2}$, тогда $\rho(\omega)$ (2.47) есть коэффициент корреляции в матрице $\tilde{\mathbf{\Sigma}}^m = \tilde{\mathbf{X}}^\intercal \mathbf{R}\tilde{\mathbf{X}}$.

Считаем далее, что генерируются независимо из некоторого известного распределения, то есть

$$p(\mathbf{X}) = \prod_{i=1}^m p(\mathbf{x}_i).$$

Тогда в силу

$$\tilde{\mathbf{\Sigma}}^m = \sum_{i=1}^m \lambda(\mathbf{w},\, \mathbf{x}_i)\tilde{\mathbf{x}}_i\tilde{\mathbf{x}}_i^\intercal$$

и усиленного закона больших чисел получаем

$$\tfrac{1}{m}\tilde{\mathbf{\Sigma}}^m \overset{\text{п.н.}}{\to} \mathbb{E}_{\mathbf{x}}\lambda(\mathbf{w},\, \mathbf{x})\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal \text{ при } m \to \infty. \tag{2.48}$$

Считая далее, что сходимость в (2.48) выполнена при $\omega \in \Omega_3$, где $\mathbb{P}(\Omega_3) = 1$, для $\omega \in \Omega_4 = \Omega_2 \cap \Omega_3$ получаем

$$\rho(\omega) \to \frac{\mathbb{E}_{\mathbf{x}}\lambda(\mathbf{w},\, \mathbf{x})|x_1||x_2|}{\sqrt{\mathbb{E}_{\mathbf{x}}\lambda(\mathbf{w},\, \mathbf{x})x_1^2}\sqrt{\mathbb{E}_{\mathbf{x}}\lambda(\mathbf{w},\, \mathbf{x})x_2^2}} = \rho_0.$$

С учетом произвольности $\delta > 0$, если $\rho_0^2 < 1$, то условие (2.27) будет выполнено, так как, взяв δ так, что

$$\rho_0^2\left(\frac{1+\delta}{1-\delta}\right)^2 \leq \frac{3\rho_0^2+1}{4}$$

и взяв m_0 так, что

$$\forall\, m \geq m_0 :\ |\rho^2(\omega) - \rho_0^2| \leq \left(\frac{1-\delta}{1+\delta}\right)^2 \frac{1-\rho_0^2}{4}$$

получим $\forall\, m \geq m_\omega = \max(m_0,\, m_\delta)$

$$\rho^2_{MP}(\omega) \leq \left(\frac{1+\delta}{1-\delta}\right)^2 \rho^2(\omega) \leq \left(\frac{1+\delta}{1-\delta}\right)^2 \left[\rho_0^2 + \left(\frac{1-\delta}{1+\delta}\right)^2 \frac{1-\rho_0^2}{4}\right] = \frac{\rho_0^2+1}{2} < 1.$$

Отсюда условие теоремы (2.27) выполнено с $m_\omega = \max(m_0,\, m_\delta)$ и $c_\omega = \frac{1-\rho_0^2}{2}$.

Таким образом, остается проверить, что при достаточно общих предположениях о распределении $p(\mathbf{x}_i)$ матрица $\mathbb{E}_{\mathbf{x}}\lambda(\mathbf{w},\, \mathbf{x})\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal$ является невырожденной. Определим необходимые условия вырожденности $\mathbb{E}_{\mathbf{x}}\lambda(\mathbf{w},\, \mathbf{x})\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal$. Имеем

$$\mathbb{E}_{\mathbf{x}}\lambda(\mathbf{w},\, \mathbf{x})\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal = \begin{pmatrix} \lambda(\mathbf{w},\, \mathbf{x})x_1^2 & \lambda(\mathbf{w},\, \mathbf{x})|x_1||x_2| \\ \lambda(\mathbf{w},\, \mathbf{x})|x_1||x_2| & \lambda(\mathbf{w},\, \mathbf{x})x_2^2 \end{pmatrix},$$

тогда вводя случайные величины $\xi_1 = \sqrt{\lambda(\mathbf{w},\, \mathbf{x})}|x_1|$, $\xi_2 = \sqrt{\lambda(\mathbf{w},\, \mathbf{x})}|x_2|$ получаем, что условие вырожденности матрицы (с учетом положительности элементов) $\mathbb{E}_{\mathbf{x}}\lambda(\mathbf{w},\, \mathbf{x})\tilde{\mathbf{x}}\tilde{\mathbf{x}}^\intercal$ приобретает вид

$$\mathbb{E}^2\xi_1\xi_2 = \mathbb{E}\xi_1^2\mathbb{E}\xi_2^2.$$

Тогда, пользуясь неравенством Коши-Буняковского-Шварца, получаем, что для случайных величин $\xi_1,\, xi_2$ выполнено

$$\xi_2 \overset{\text{п.н.}}{=} 0 \text{ или } \exists\, C \geq 0 :\ \xi_1 \overset{\text{п.н.}}{=} C\xi_2.$$

Отсюда ситуация вырожденности матрицы наблюдается с учетом $\lambda(\mathbf{w},\, \mathbf{w}) > 0$ либо когда один из признаков тождественно нулевой, либо если $|x_2| = C|x_1|$, то есть когда признаки пропорциональны.

Таким образом, если признаки не являются тождественно нулевыми и коллинеарными, то условия теоремы выполнены и указанная асимптотическая вырожденность оценки ковариационной матрицы метод максимума обоснованности будет иметь место.

2.2. Отбор признаков с помощью максимизации обоснованности для многоуровневой модели

Для многоуровневой модели в соответствии с (1.8) существует разбиение признакового пространства вида $\mathbb{R}^n = \Omega_1 \sqcup \ldots \sqcup \Omega_K$ на K непересекающихся частей, в каждой из которых используется своя модель. Введем соответствующее этому разбиению разбиение множества индексов объектов $\{1,\, \ldots,\, m\} = \mathcal{I} = \mathcal{I}_1 \sqcup \ldots \sqcup \mathcal{I}_K$, где $\forall i \in \mathcal{I}_k\, \mathbf{x}_i \in \Omega_k$, $k = \overline{1,\, K}$. Тогда совместное правдоподобие для многоуровневой модели приобретает вид

$$p(\mathbf{y},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K|\mathbf{X},\, \mathbf{A}_1,\, \ldots,\, \mathbf{A}_K) = \prod_{k=1}^{K}\left[p(\mathbf{w}_k|\mathbf{A}_k)\prod_{i\in\mathcal{I}_k}\sigma(y_i\mathbf{w}_k^\intercal\mathbf{x}_i)\right]. \qquad (2.49)$$

Из (2.49) заключаем, что совместное правдоподобие для многоуровневой модели есть произведение K независимых компонент для одиночных моделей, построенных на непересекающихся множествах объектов, объединение которых составляет все множество объектов. Тогда интегрируя по $\mathbf{w}_1, \ldots, \mathbf{w}_K$, получаем

$$p(\mathbf{y}|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K) = \prod_{k=1}^{K} p(\mathbf{y}_{\mathcal{I}_k}|\mathbf{X}_{\mathcal{I}_k}, \mathbf{A}_k),$$

откуда с учетом (1.4) получаем, что задача (1.4) нахождения оптимальной многоуровневой модели эквивалентна решению K независимых задач вида

$$\mathbf{A}_k^* = \arg\max_{\mathbf{A}_k \in Q_{\mathbf{A}_k}} p(\mathbf{y}_{\mathcal{I}_k}|\mathbf{X}_{\mathcal{I}_k}, \mathbf{A}_k), \; k = \overline{1, K}. \tag{2.50}$$

Здесь $\mathbf{X}_{\mathcal{I}_k}$, $\mathbf{y}_{\mathcal{I}_k}$ обозначено усечение матрицы признаков и вектора ответов на множество индексов $\mathcal{I}_k$, то есть

$$\mathbf{X}_{\mathcal{I}_k} = [\mathbf{x}_i, \; i \in \mathcal{I}_k]^\intercal, \; \mathbf{y}_{\mathcal{I}_k} = [y_i, \; i \in \mathcal{I}_k]^\intercal.$$

Отметим теперь, что каждая из задач (2.50) представляет собой задачу максимизации обоснованности для одиночной модели логистической регрессии на усеченном множестве объектов. Такая задача рассмотрена в предыдущей подсекции, а потому нахождение оптимальной многоуровневой модели сводится к решению K уже рассмотренных задач.

2.3. Отбор признаков с помощью максимизации обоснованности для смеси моделей

Для смеси моделей совместное правдоподобие имеет вид

$$p(\mathbf{y}, \boldsymbol{\pi}, \mathbf{w}_1, \ldots, \mathbf{w}_K|\mathbf{X}, \alpha, \mathbf{A}_1, \ldots, \mathbf{A}_k) = p(\boldsymbol{\pi}|\alpha) \prod_{k=1}^{K} p_k(\mathbf{w}_k|\mathbf{A}_k) \prod_{i=1}^{m} \left(\sum_{l=1}^{K} \pi_l \sigma(y_i \mathbf{w}_l^\intercal \mathbf{x}_i) \right). \tag{2.51}$$

С учетом асимтотической вырожденности недиагональной оценки ковариационной матрицы $\mathbf{A}$ даже для случая одной модели, далее считаем, что $\mathbf{A}_k = \operatorname{diag}(\boldsymbol{\alpha}_k)$, $\boldsymbol{\alpha}_k \in \mathbb{R}^n$. Как уже указывалось, в качестве априорного распределения на веса моделей $\boldsymbol{\pi}$, входящих в смесь моделей, используется симметричное распределение Дирихле с параметром α, а априорные распределения на $\mathbf{w}_k$ есть нормальные, то есть

$$p(\boldsymbol{\pi}|\alpha) = \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha-1}, \; p(\mathbf{w}_k|\mathbf{A}_k) = \frac{\sqrt{\det \mathbf{A}_l}}{(2\pi)^{n/2}} \exp\left(-\tfrac{1}{2}\mathbf{w}_l^\intercal \mathbf{A}_l \mathbf{w}_l\right), \; k = 1, \ldots, K.$$

Введем матрицу скрытых переменных $\mathbf{Z}$ размера $m \times K$, где $z_{ik} \in \{0, 1\}$ и $z_{ik} = 1$ тогда и только тогда, когда объект $(\mathbf{x}_i,\, y_i)$ отнесен к модели с номером k. Тогда совместное правдоподобие для модели со скрытыми переменными примет вид

$$p(\mathbf{y}, \boldsymbol{\pi}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{Z} | \mathbf{X}, \alpha, \mathbf{A}_1, \ldots, \mathbf{A}_k) = p(\boldsymbol{\pi}|\alpha) \prod_{k=1}^{K} p_k(\mathbf{w}_k|\mathbf{A}_k) \prod_{i=1}^{m} \prod_{l=1}^{K} (\pi_l \sigma(y_i \mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i))^{z_{il}}$$

Считая параметр α распределения Дирихле фиксированным, получаем следующую задачу максимизации обоснованности для оценки неизвестных ковариационных матриц $\mathbf{A}_1,\ \mathbf{A}_K$

$$\log p(\mathbf{y}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K) \max_{\mathbf{A}_1,\, \ldots,\, \mathbf{A}_K \in \mathcal{M}_{diag}}.$$

Оценим эту величину снизу, вводя распределение $q(\mathbf{w}_1, \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi},\ \mathbf{Z})$ и используя вариационную нижнюю оценку для сигмоидной функции, получим

$$\log p(\mathbf{y}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K) \geq L(q,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K) =$$
$$\mathbb{E}_q \log p(\mathbf{y}, \boldsymbol{\pi}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{Z}|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_k) - \mathbb{E}_q \log q(\mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z}) \geq \tilde{L}(q, \mathbf{A}_1, \ldots$$
$$\mathbb{E}_q \log p(\mathbf{y},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \mathbf{Z}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_k,\ \boldsymbol{\xi}) - \mathbb{E}_q \log q(\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi},\ \mathbf{Z}),\ \text{где}$$

$$\log p(\mathbf{y}, \boldsymbol{\pi}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{Z}|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_k, \boldsymbol{\xi}) = \log \Gamma(K\alpha) - K \log \Gamma(\alpha) + \sum_{k=1}^{K} (\alpha - 1) \log \pi_k$$
$$\sum_{i=1}^{m} \sum_{k=1}^{K} z_{ik} \left[\log \sigma(\xi_{ik}) + \frac{y_i \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i - \xi_{ik}}{2} - \lambda(\xi_{ik})((\mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i)^2 - \xi_{ik}^2) \right] +$$
$$\frac{1}{2} \sum_{k=1}^{K} \log \det \mathbf{A}_k - \frac{1}{2} \sum_{k=1}^{K} \mathbf{w}_k^{\mathsf{T}} \mathbf{A}_k \mathbf{w} + \sum_{k=1}^{K} \log \pi_k \sum_{i=1}^{m} z_{ik} - K \frac{n}{2} \log(2\pi).$$

Будем теперь решать задачу

$$\tilde{L}(q,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \boldsymbol{\xi}) \to \max_{q,\, \mathbf{A}_1,\, \ldots,\, \mathbf{A}_K,\, \boldsymbol{\xi}}. \tag{2.52}$$

Для этого воспользуемся вариационным EM-алгоритмом, то есть решение для q будем искать в классе распределений, которые имеют следующую факторизацию.

$$Q = \left\{ q :\ q(\mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi},\ \mathbf{Z}) = \prod_{k=1}^{K} q_{\mathbf{w}_k}(\mathbf{w}_k) q_{\boldsymbol{\pi}}(\boldsymbol{\pi}) \prod_{i=1}^{m} q_{\mathbf{z}_i}(\mathbf{z}_i) \right\}.$$

E-шаг

На E-шаге проводим максимизацию $\tilde{L}(q, \mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\xi})$ по $q(\mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z})$

в классе Q. Тогда получаем следующие формулы для распределений $q(\boldsymbol{\pi})$, $q_{\mathbf{w}_k}(\mathbf{w}_k)$, $k = \overline{1,\ K}$, $q_{\mathbf{z}_i}(\mathbf{z}_i)$, $i = \overline{1,\ m}$ [29].

$$\log q_{\boldsymbol{\pi}}(\boldsymbol{\pi}) \propto \mathbb{E}_{q_{\setminus \boldsymbol{\pi}}} \log p(\mathbf{y},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \mathbf{Z}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_k,\ \boldsymbol{\xi}),$$

$$\log q_{\mathbf{w}_k}(\mathbf{w}_k) \propto \mathbb{E}_{q_{\setminus \mathbf{w}_k}} \log p(\mathbf{y},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \mathbf{Z}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_k,\ \boldsymbol{\xi}),$$

$$\log q_{\mathbf{z}_i}(\mathbf{z}_i) \propto \mathbb{E}_{q_{\setminus \mathbf{z}_i}} \log p(\mathbf{y},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \mathbf{Z}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_k,\ \boldsymbol{\xi}),$$

где пропорциональность приведена с точностью до аддитивной постоянной. Получим теперь явный вид распределений $q(\boldsymbol{\pi})$, $q_{\mathbf{w}_k}(\mathbf{w}_k)$, $k = \overline{1,\ K}$, $q_{\mathbf{z}_i}(\mathbf{z}_i)$, $i = \overline{1,\ m}$.

$$\log q_{\boldsymbol{\pi}}(\boldsymbol{\pi}) = \sum_{k=1}^{K} \log \pi_k \left(\alpha - 1 + \sum_{i=1}^{m} \mathbb{E} z_{ik} \right) + C_0,$$

где C_0 есть некоторая постоянная. Отсюда заключаем, что распределение на $\boldsymbol{\pi}$ есть распредение Дирихле, вообще говоря несимметричное, с параметрами $\alpha + \sum_{i=1}^{m} \mathbb{E} z_{ik}$, $k = \overline{1,\ K}$, то есть

$$q_{\boldsymbol{\pi}}(\boldsymbol{\pi}) = \frac{\Gamma(K\alpha + m)}{\prod_{k=1}^{K} \Gamma\left(\alpha + \sum_{i=1}^{m} \mathbb{E} z_{ik}\right)} \prod_{k=1}^{K} \pi_k^{\alpha - 1 + \sum_{i=1}^{m} \mathbb{E} z_{ik}} [\min_l \pi_l > 0].$$

Аналогично для $q_{\mathbf{z}_i}(\mathbf{z}_i)$ имеем

$$\log q_{\mathbf{z}_i}(\mathbf{z}_i) \propto C_1^i + \sum_{k=1}^{K} z_{ik} \mathbb{E} \log \pi_k + \sum_{k=1}^{K} z_{ik} \left[\log \sigma(\xi_{ik}) + \frac{y_i (\mathbb{E}\mathbf{w}_k)^\intercal \mathbf{x}_i - \xi_{ik}}{2} - \lambda(\xi_{ik}) \left(\mathbf{x}_i^\intercal \mathbb{E}(\mathbf{w}_k \mathbf{w}_k^\intercal) \right.\right.$$

Отсюда получаем, что

$$\mathbb{P}(z_{ik} = 1) = \frac{\exp\left[\mathbb{E} \log \pi_k + \log \sigma(\xi_{ik}) + \frac{y_i (\mathbb{E}\mathbf{w}_k)^\intercal \mathbf{x}_i - \xi_{ik}}{2} - \lambda(\xi_{ik}) \left(\mathbf{x}_i^\intercal \mathbb{E}(\mathbf{w}_k \mathbf{w}_k^\intercal) \mathbf{x}_i - \xi_{ik}^2 \right) \right.}{\sum_{l=1}^{K} \exp\left[\left(\mathbb{E} \log \pi_l + \log \sigma(\xi_{il}) + \frac{y_i (\mathbb{E}\mathbf{w}_l)^\intercal \mathbf{x}_i - \xi_{il}}{2} - \lambda(\xi_{il}) \left(\mathbf{x}_i^\intercal \mathbb{E}(\mathbf{w}_l \mathbf{w}_l^\intercal) \mathbf{x}_i - \xi_{il}^2 \right.\right.\right.}$$

Аналогично для $q_{\mathbf{w}_k}(\mathbf{w}_k)$ получаем

$$\log q_{\mathbf{w}_k}(\mathbf{w}_k) \propto C_2^k - \tfrac{1}{2} \mathbf{w}_k^\intercal \mathbf{A}_k \mathbf{w}_k^\intercal + \sum_{i=1}^{m} \mathbb{E} z_{ik} \left(\frac{y_i \mathbf{w} - k^\intercal \mathbf{x}_i}{2} - \lambda(\xi_{ik}) (\mathbf{w}_k^\intercal \mathbf{x}_i)^2 \right) =$$

$$C_2^k - \tfrac{1}{2} \mathbf{w}_k^\intercal \mathbf{A}_k \mathbf{w}_k + \mathbf{w}_k^\intercal \mathbf{m}_k - \mathbf{w}_k^\intercal \left[\sum_{i=1}^{m} \mathbb{E} z_{ik} \lambda(\xi_{ik}) \mathbf{x}_i \mathbf{x}_i^\intercal \right] \mathbf{w}_k,$$

где $\mathbf{m}_k = \frac{1}{2} \sum_{i=1}^{m} y_i \mathbb{E} z_{ik} \mathbf{x}_i$. Отсюда получаем, что $q_{\mathbf{w}_k}(\mathbf{w}_k)$ есть нормалное распределение, то есть

$$q_{\mathbf{w}_k}(\mathbf{w}_k) = N(\mathbf{w}_k | \tilde{\mathbf{A}}_k^{-1} \mathbf{m}_k,\ \tilde{\mathbf{A}}_k^{-1}),\ \tilde{\mathbf{A}}_k = \mathbf{A}_k + 2 \sum_{i=1}^{m} \mathbb{E} z_{ik} \lambda(\xi_{ik}) \mathbf{x}_i \mathbf{x}_i^\intercal.$$

M-шаг
На M-шаге решается следующая задача максимизации

$$\mathbb{E}_q \log p(\mathbf{y},\, \boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_k,\, \mathbf{Z}|\mathbf{X},\, \mathbf{A}_1,\, \ldots,\, \mathbf{A}_k,\, \boldsymbol{\xi}) \to \max_{\boldsymbol{\xi},\, \mathbf{A}_1,\, \mathbf{A}_k}. \tag{2.53}$$

Для решения задачи (2.53) воспользуемся итеративным алгоритмом. Сначала будем производить максимизацию по $\boldsymbol{\xi}$ при фиксированных $\mathbf{A}_k$, а затем по $\mathbf{A}_k$, $k = 1, \ldots, K$ при фиксированном $\boldsymbol{\xi}$. Выпишем сначала выражения для максимизируемой функции в (**??**).

$$\mathbb{E}_q \log p(\mathbf{y},\, \boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_k,\, \mathbf{Z}|\mathbf{X},\, \mathbf{A}_1,\, \ldots,\, \mathbf{A}_k,\, \boldsymbol{\xi}) = C_3 + \tfrac{1}{2}\sum_{k=1}^{K} \log\det \mathbf{A}_k -$$

$$\tfrac{1}{2}\sum_{k=1}^{K} tr(\mathbf{A}_k \mathbb{E}(\mathbf{w}_k \mathbf{w}_k^\intercal)) + \sum_{i=1}^{m}\sum_{k=1}^{K} \mathbb{E} z_{ik} \left[\log\sigma(\xi_{ik}) + \frac{y_i(\mathbb{E}\mathbf{w}_k)^\intercal \mathbf{x}_i - \xi_{ik}}{2} - \lambda(\xi_{ik}) \left(\mathbf{x}_i^\intercal \mathbb{E}(\mathbf{w}_k \mathbf{w}_k^\intercal)\right.\right.$$

Итерация по $\boldsymbol{\xi}$ при фиксированных $\mathbf{A}_1, \ldots, \mathbf{A}_K$.

С учетом того, что максимизируемая функция распадается в сумму mK независимых компонент по элементам ξ_{ik} матрицы $\boldsymbol{\xi}$, максимизацию производим по каждой из переменных ξ_{ik} независимо. Условие оптимальности первого порядка по ξ_{ik} имеет вид

$$\sigma(-\xi_{ik}) - \tfrac{1}{2} - \lambda'(\xi_{ik})(\mathbf{x}_i^\intercal \mathbb{E}\mathbf{w}_k \mathbf{w}_k^\intercal \mathbf{x}_i - \xi_{ik}^2) + 2\lambda(\xi_{ik})\xi_{ik} = 0.$$

С учетом $\lambda(\xi) = \frac{1}{2\xi}(\sigma(\xi) - \frac{1}{2})$ и $\sigma(\xi) + \sigma(-\xi) = 1$, получаем, что условие оптимальности по ξ_{ik} приобретает вид

$$\sigma(-\xi_{ik}) - \tfrac{1}{2} - \lambda'(\xi_{ik})\mathbf{x}_i^\intercal \mathbb{E}\mathbf{w}_k \mathbf{w}_k^\intercal \mathbf{x}_i + \sigma(\xi_{ik}) - \tfrac{1}{2} + \lambda'(\xi_{ik})\xi_{ik}^2 = 0,$$

откуда

$$\xi_{ik}^2 = \mathbf{x}_i^\intercal (\mathbb{E}\mathbf{w}_k \mathbf{w}_k^\intercal)\mathbf{x}_i.$$

Итерация по $\mathbf{A}_1, \ldots, \mathbf{A}_K$ при фиксированном $\boldsymbol{\xi}$.

С учетом того, что максимизируемая функция распадается в сумму K независимых компонент по матрицам $\mathbf{A}_k$, $k = \overline{1, K}$, максимизацию производим по каждой из переменных $\mathbf{A}_k$ независимо. Тогда для элемента a_k^{pq} матрицы $\mathbf{A}_k$, находящегося на месте (p, q) условие оптимальности первого порядка приобретает вид

$$(\mathbf{A}_k^{-1})^{pq} = \mathbb{E}(\mathbf{w}_k \mathbf{w}_k^\intercal)^{pq},$$

где было учтено, что матрица $\mathbf{A}_k$ симметрична и

$$\frac{\partial \log\det \mathbf{A}}{\partial \mathbf{A}} = \mathbf{A}^{-\intercal},\ \frac{\partial\, tr(\mathbf{A}\mathbf{B})}{\partial \mathbf{A}} = \mathbf{B}^\intercal.$$

Таким образом, если искать решение для $\mathbf{A}_k$ в множестве матриц $\mathcal{M}_{full}$, то $\mathbf{A}_k = (\mathbb{E}\mathbf{w}_k\mathbf{w}_k^\intercal)^{-1}$. Однако, как было показано а теореме **??** оценка недиагональной ковариационной матрицы путем максимизации обоснованности является асимптотически вырожденной, а потому далее рассматриваем оценку $\mathbf{A}_k$ в множестве диагональных матриц $\mathbf{A}_{diag}$. Обозначим $\mathbf{A}_k = \mathrm{diag}(\alpha_{kj})$, $j = \overline{1,\, n}$. Тогда условие оптимальности первого порядка приобретает вид

$$\alpha_{kj} = \frac{1}{\mathbb{E}w_{kj}^2},\ j = \overline{1,\, n}.$$

Таким образом, повторяя итерации EM-алгоритма, получим оценки $\mathbf{A}_1, \ldots, \mathbf{A}_K$, которые позволят отобрать признаки в каждой из моделей.

2.4. Комбинирование признаков для учета взаимосвязей между ними

Отметим, что, как доказано ранее, оценка максимума обоснованности для ковариационной матрицы $\mathbf{A}$ является асимптотически вырожденной, а недиагональный элемент определяется не наличием связи между признаками, а лишь знаками весов. По этой причине поиск оценок максимума обоснованности для $\mathbf{A}_k$, $k = \overline{1,\, K}$ производится в множестве неотрицательно определенных диагональных матриц $\mathcal{M}_{diag}$. Однако, как показано в (2.2), при наличии зависимостей между признаками комбинирование признаков оказывается предпочтительным по отношению к отбору признаков в терминах дисперсии шума в признаковом описании. Отметим, однако, что истинная корреляционная структура признаков неизвестна, а потому оптимальное комбинирование признаков, дающееся (2.2), не реализуемо.

Рассмотрим далее несколько методов учета мультиколлинеарности признаков. В качестве исходных данных эти методы используют оценку ковариационной матрицы признаков $\mathbf{\Sigma}$. В работе в качестве таких оценок используются следующие.

- $\mathbf{\Sigma}^0 = \frac{1}{m}\tilde{\mathbf{X}}^\intercal\tilde{\mathbf{X}}$, где $\tilde{\mathbf{X}}$ есть центрированная матрица признаков, $\tilde{\mathbf{X}} = \mathbf{X} - \frac{1}{m}\mathbf{X}^\intercal\mathbf{e}_m\mathbf{e}_n^\intercal$, где $\mathbf{e}_m$, $\mathbf{e}_n$ есть единичные векторы размерности m и n соответственно.
- $\mathbf{\Sigma}^1 = \lambda_1\mathbf{\Sigma}^0 + (1-\lambda_1)\mathrm{diag}(\sigma_j^2)$, где $\sigma_j^2 = \mathbf{\Sigma}^0_{jj}$, $\lambda_1 \in [0,\, 1]$.
- $\mathbf{\Sigma}^2 = \lambda_2\mathbf{\Sigma}^0 + (1-\lambda_2)\sigma_0^2\mathbf{I}$, где $\sigma_0^2 = \frac{1}{n}\sum_{j=1}^n \mathbf{\Sigma}^0_{jj}$.

Отметим, что существует множество оценок ковариационной матрицы, которые обладают лучшими по отношению к выборочной ковариационной матрице свойствами [60–63]. Две такие оценки со стягиванием к диагональной и стягиванием к единичной матрице рассмотрены в данной работе. Однако существуют оценки ковариационной матрицы, основанные на стягивании к постоянным корреляциям, к матрицам заданной структуры [62], например, с помощью байесовских моделей генерации данных [63]. Однако оценки ковариационной матрицы, построенные с помощью этих методов, не используются в данной работе,

поскольку основной целью является демонстрация преимущества комбинирования признаков, а не исследования методов получения оценок ковариационной матрицы признаков. Отметим, что тем не менее другие оценки ковариационной матрицы признаков также могут быть использованы для комбинирования признаков с помощью предлагаемых далее методов.

Прежде, чем перейти к изложению предлагаемых методов комбинирования признаков, укажем оптимальные значения λ_1 и λ_2, которые используются в данной работе.

Утверждение 3. Пусть $\mathbb{E}\mathbf{x}_i = \mathbf{0}$, $\mathbb{E}\mathbf{x}_i\mathbf{x}_i^\intercal = \mathbf{\Sigma}$, а в качестве оценки ковариационной матрицы используется

$$\mathbf{\Sigma}^1 = \lambda\mathbf{\Sigma}^0 + (1-\lambda)\mathrm{diag}(\sigma_j^2), \text{ где } \mathbf{\Sigma}^0 = \tfrac{1}{m}\mathbf{X}^\intercal\mathbf{X}, \sigma_j^2 = \mathbf{\Sigma}^0_{jj}, \lambda \in [0,\ 1].$$

Тогда решением задачи $f(\lambda) = \mathbb{E}\|\mathbf{\Sigma}^1 - \mathbf{\Sigma}\|^2 \to \min_\lambda$ является

$$\lambda^* = \frac{\sum_{j_1\neq j_2=1}^n \mathbb{E}^2\zeta_{j_1}\zeta_{j_2}}{\frac{m-1}{m}\sum_{j_1\neq j_2=1}^n \mathbb{E}^2\zeta_{j_1}\zeta_{j_2} + \frac{1}{m}\sum_{j_1\neq j_2=1}^n \mathbb{E}\zeta_{j_1}^2\zeta_{j_2}^2},$$

где ζ_j обозначает случайную величину, соответствующую признаку с номером j.

Доказательство. Вычислим сначала $\mathbb{E}\left(\sum_{i=1}^m x_{ij_1}x_{ij_2}\right)^2$.

$$\mathbb{E}\left(\sum_{i=1}^m x_{ij_1}x_{ij_2}\right)^2 = \mathbb{E}\left(\sum_{i=1}^m x_{ij_1}^2x_{ij_2}^2\right) + \sum_{i_1\neq i_2=1}^m \mathbb{E}x_{i_1j_1}x_{i_1j_2}x_{i_2j_1}x_{i_2j_2} =$$
$$m\mathbb{E}\zeta_{j_1}^2\zeta_{j_2}^2 + m(m-1)\mathbb{E}^2\zeta_{j_1}\zeta_{j_2}.$$

Отметим, что диагональ матрицы $\mathbf{\Sigma}^1$ от λ не зависит, а потому имеем

$$m^2 f(\lambda) = \sum_{j_1\neq j_2=1}^n \mathbb{E}\left(\lambda\sum_{i=1}^m x_{ij_1}x_{ij_2} - m\mathbb{E}\zeta_{j_1}\zeta_{j_2}\right)^2 =$$
$$\sum_{j_1\neq j_2=1}^n \left[\lambda^2\mathbb{E}\left(\sum_{i=1}^m x_{ij_1}x_{ij_2}\right)^2 + m^2\mathbb{E}^2\zeta_{j_1}\zeta_{j_2} - 2m^2\lambda\mathbb{E}^2\zeta_{j_1}\zeta_{j_2}\right] =$$
$$m\sum_{j_1\neq j_2=1}^n \left(\lambda^2\mathbb{E}\zeta_{j_1}^2\zeta_{j_2}^2 + \mathbb{E}^2\zeta_{j_1}\zeta_{j_2}\left[\lambda^2(m-1) + m - 2\lambda m\right]\right).$$

Отсюда в силу квадратичногости по λ $f(\lambda)$ с положительным коэффициентом при квадратичном по λ члене, получаем выражение для оптимального λ.

$$\lambda^* = \frac{\sum_{j_1\neq j_2=1}^n \mathbb{E}^2\zeta_{j_1}\zeta_{j_2}}{\frac{m-1}{m}\sum_{j_1\neq j_2=1}^n \mathbb{E}^2\zeta_{j_1}\zeta_{j_2} + \frac{1}{m}\sum_{j_1\neq j_2=1}^n \mathbb{E}\zeta_{j_1}^2\zeta_{j_2}^2},$$

□

Утверждение 4 ([62]). Пусть $\mathbb{E}\mathbf{x}_i = \mathbf{0}$, $\mathbb{E}\mathbf{x}_i\mathbf{x}_i^\intercal = \mathbf{\Sigma}$, а в качестве оценки ковариационной матрицы используется

$$\mathbf{\Sigma}^2 = \lambda\mathbf{\Sigma}^0 + (1-\lambda)\sigma_0^2\mathbf{I}, \text{ где}\mathbf{\Sigma}^0 = \tfrac{1}{m}\mathbf{X}^\intercal\mathbf{X}, \ \sigma_0^2 = \tfrac{1}{n}\sum_{j=1}^{n}\mathbf{\Sigma}_{jj}^0.$$

Тогда решением задачи $f(\lambda) = \mathbb{E}\|\mathbf{\Sigma}^2 - \mathbf{\Sigma}\|^2 \to \min_\lambda$ является

$$\lambda^* = \frac{\nu^2}{\nu^2 + \beta^2}, \text{ где}$$

$$\nu^2 = \sum_{j_1 \neq j_2 = 1}^{n} \mathbb{E}^2\zeta_{j_1}\zeta_{j_2} + \mathbf{d}^\intercal(\mathbf{I} - \tfrac{1}{n}\mathbf{E})\mathbf{d}, \text{ где } \mathbf{d} = [\mathbb{E}\zeta_j^2, \ j = \overline{1, \ n}]^\intercal, \ \mathbf{E} = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}$$

$$\beta^2 = \tfrac{1}{m}\sum_{j_1,\, j_2 = 1}^{n} \left(\mathbb{E}\zeta_{j_1}^2\zeta_{j_2}^2 - \mathbb{E}^2\zeta_{j_1}\zeta_{j_2}\right),$$

где ζ_j обозначает случайную величину, соответствующую признаку с номером j.

Отметим, что выражения для оптимального λ^* в обоих утверждениях предполагают исходную центрированность признаков и знание набора истинных математических ожиданий для некоторых произведений признаков. Так как эти значения неизвестны, соотвествующие метематические ожидания заменяются на средние по выборке. Опишем далее методы комбинирования признаков, основанные на анализе полученной ковариационной матрицы $\hat{\mathbf{\Sigma}}$.

Детектирование копий одного признака. Пусть признаки в признаковой матрице $\mathbf{X}$ центрированы по выборке, а также шкалированы к дисперсии 1. Пусть для них имеется некоторая оценка ковариационной матрицы признаков $\hat{\mathbf{\Sigma}}$. Считаем далее, что мультиколилинеарность между признакам есть только одного типа, а именно в выборке могут содержаться зашумленные копии одного и того же фактора, то есть имеется матрица значений факторов $\mathbf{F} \in \mathbb{R}^{m \times n_0}$, а $\mathbf{X} = \mathbf{F}\mathbf{V} + \boldsymbol{\varepsilon}$, где $\mathbf{V} \in \{0, 1\}^{n_0 \times n}$ и $\mathbf{e}_{n_0}^\intercal\mathbf{V} = \mathbf{e}_n^\intercal$, то есть в каждом столбце матрицы $\mathbf{V}$ ровно одна единица, а $\boldsymbol{\varepsilon}$ есть некоторая шумовая матрица.

Отметим, что если дан набор $\mathcal{J} \in 2^{\{1, \dots, n\}}$, $|\mathcal{J}| = n_1$ зашумленных копий некоторого фактора, то при известных дисперсиях шума в каждой из копий, схема оптимального комбинирования дается формулой (2.4). Так как дисперсии шумов неизвестны, то для их оценки воспользуемся следующим итеративным алгоритмом.

1. Инициализация: весовой вектор $\mathbf{v} = \mathbf{e}$, оценка фактора $\hat{\mathbf{f}} = \nu\sum_{j\in\mathcal{J}} v_j\mathbf{f}_j$, где ν константа, шкалирующая $\hat{\mathbf{f}}$ к дисперсии 1;
2. Подсчет оценки дисперсии признаков относительно оценки фактора

$$d_j^2 = \tfrac{1}{m}\|\mathbf{f}_j - \hat{\mathbf{f}}\|^2$$

и подсчет оценки дисперсии шума

$$\hat{\sigma}_j^2 = \max\left(0,\ \frac{d_j^2}{2}\left(2 - \frac{d_j^2}{2}\right)\right).$$

3. Пересчет весового вектора $\mathbf{v}$

$$v_j = \frac{\frac{1}{\hat{\sigma}_j^2}}{\sum_{l\in\mathcal{J}} \frac{1}{\hat{\sigma}_l^2}};$$

4. Пересчет оценки фактора

$$\hat{\mathbf{f}}^{\text{new}} = \nu^{\text{new}} \sum_{j\in\mathcal{J}} v_j \mathbf{f}_j,$$

где ν^{new} константа, шкалирующая $\hat{\mathbf{f}}^{\text{new}}$ к дисперсии 1. Останавливаемся, если $\|\hat{\mathbf{f}}^{\text{new}} - \hat{\mathbf{f}}\| < \delta$, где $\delta \geq 0$ – заданная постоянная.

Заметим, что данный алгоритм при $n_1 = 2$ вырождается в использование шкалированной полусуммы признаков в качестве оценки фактора. При $n_1 > 2$ результат уже является нетривиальным. Обозначим результат работы этого алгоритма как $\hat{f}(\mathcal{J})$.

Опишем далее алгоритм комбинирования признаков. Пусть задана некоторая граница по корреляции $\rho_0 > 0$. Считаем, что пары признаков с оценкой корреляции в матрице $\hat{\mathbf{\Sigma}}$ больше ρ_0 являются связанными. Построим разбиение набора признаков на группы так, что оценки факторов, построенные по группам имеют корреляцию не больше ρ.

Алгоритм последовательного парного объединения по наибольшей корреляции. Будем последовательно объединять пары признаков с наибольшей на данном шаге корреляцией, пока для любой пары корреляция не станет не больше ρ_0. Такая идея приводит к следующему алгоритму.

1. Инициализация: считаем каждый из признаков отдельным фактором. Для каждого фактора j сохраняем набор C_j исходных признаков, которые ему соответствуют. Исходно $\forall\, j\ C_j = \{j\}$.
2. Находим пару наиболее коррелированных признаков.

$$[j_1^*,\ j_2^*] = \arg\max_{j_1\neq j_2} \frac{\hat{\mathbf{\Sigma}}_{j_1 j_2}}{\sqrt{\hat{\mathbf{\Sigma}}_{j_1 j_1}\hat{\mathbf{\Sigma}}_{j_2 j_2}}} = \arg\max_{j_1\neq j_2} \rho_{j_1 j_2}.$$

3. Если

$$\frac{\hat{\mathbf{\Sigma}}_{j_1^* j_2^*}}{\sqrt{\hat{\mathbf{\Sigma}}_{j_1^* j_1^*}\hat{\mathbf{\Sigma}}_{j_2^* j_2^*}}} = \rho_{j_1^* j_2^*} \leq \rho_0,$$

останавливаемся, так как построена требуемая комбинация признаков. Иначе переходим на шаг 4.

4. Пересчитываем наборы признаков, соответствующие факторам.

$$C_{j_1^*} \cup C_{j_2^*} \to C_{j_1^*}, \varnothing \to C_{j_2^*},$$

удаляем признак с номером j_2^* из матрицы признаков и пересчитываем признак с номером j_1^* в соответствии с описанным ранее алгоритмом объединения признаков

$$\hat{\mathbf{f}}(C_{j_1^*}) \to \mathbf{f}_{j_1^*}.$$

Переходим на шаг 2.

Алгоритм последовательного поиска наибольших клик по корреляции. Будем последовательно искать максимальный набор признаков такой, что все корреляции внутри набора больше ρ_0. При наличии нескольких наборов одинакового размера выбираем тот, суммарная корреляции признаков в котором выше. Такая идея приводит к следующему алгоритму.

1. Инициализация: считаем каждый из признаков отдельным фактором. Для каждого фактора j сохраняем набор C_j исходных признаков, которые ему соответствуют. Исходно $\forall\, j\ C_j = \{j\}$.
2. Находим наибольшую клику коррелированных признаков

$$\tilde{\mathcal{J}} = \operatorname*{Arg\,max}_{\mathcal{J} \in 2^{\{1,\ldots,n\}}} |\mathcal{J}| [\min_{j_1,\, j_2 \in \mathcal{J}} \rho_{j_1 j_2} > \rho_0].$$

Среди всех наибольших клик находим клику с наибольшей суммарной внутренней корреляцией.

$$\mathcal{J}^* = \operatorname*{arg\,max}_{\mathcal{J} \in \tilde{\mathcal{J}}} \sum_{j_1,\, j_2 \in \mathcal{J}} \rho_{j_1 j_2}.$$

3. Если для $\mathcal{J}^*$ $\min_{j_1,\, j_2 \in \mathcal{J}^*} \rho_{j_1 j_2} \leq \rho_0$, останавливаемся, поскольку не осталось пар коррелированных признаков. Иначе переходим на шаг 4.
4. Пересчитываем наборы признаков, соответствующие факторам. Пусть $j^* \in \mathcal{J}^*$.

$$\cup_{j \in \mathcal{J}^*} C_j \to C_{j^*}, \varnothing \to C_j,\ j \in \mathcal{J}^* \setminus \{j^*\}.$$

Удаляем признаки с номерами $j \in \mathcal{J}^* \setminus \{j^*\}$ из матрицы признаков и пересчитываем признак с номером j^* в соответствии с описанным ранее алгоритмом объединения признаков

$$\hat{\mathbf{f}}(C_{j^*}) \to \mathbf{f}_{j^*}.$$

Переходим на шаг 2.

Результатом работы обоих приведенных здесь алгоритмов является новая матрица признаков, содержащая вообще говоря меньше признаков, при этом корреляции между любыми двумя из них не превышают ρ_0. Отметим, что ρ_0 является параметром обоих алгоритмов. Выбор значения ρ_0 предлагается делать с помощью кросс-валидации на обучающей выборке.

Таким образом, построен метод, позволяющих детектировать группу мультиколлинеарных признаков, которые представляют собой копии одноо и того же признака. Отметим, что метод легко модифицируется на случай, когда признак может представлять собой противоположный фактору. В этом случае критерием добавления в набор является $|\rho| > \rho_0$, а для применения формулы (2.4) признак требуется заменить на противоположный. Рассмотрим далее общую задачу детектирования линейной зависимости между признаками и метод учета мультиколлинеарности.

Детектирование мультиколлинеарности между признаками. Пусть признаки в признаковой матрице $\mathbf{X}$ центрированы по выборке, а также шкалированы к дисперсии 1. Пусть для них имеется некоторая оценка ковариационной матрицы признаков $\hat{\mathbf{\Sigma}}$.

Определение 21. Набор признаков с индексами $j \in \mathcal{J}$ называется δ – *мультиколлинеарным* для $\delta > 0$ с вектором весов $\boldsymbol{\theta}$, $\theta_j \neq 0 \Longleftrightarrow j \in \mathcal{J}$, если $\|\hat{\mathbf{\Sigma}}\boldsymbol{\theta}\|_1 < \delta$.

Отметим, что даже, если коэффициенты линейной зависимости априори известны, оптимальное комбинирование признаков в соответствии с (2.3) зависит от истинных весов признаков, которые неизвестны. Более того, шумовые составляющие признаков также неизвестны и как в случае с парой признаков, являющихся копией одного признака, установить, который из них содержит больше шума, по признаковой матрице невозможно. Поэтому предлагается следующий алгоритм учета мультиколлинеарности.

- Находим наиболее мультиколлинеарный набор признаков, решая следующую задачу.
$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \left[\|\hat{\mathbf{\Sigma}}\boldsymbol{\theta}\|_1 + \tau\|\boldsymbol{\theta}\|_1 \right],$$
где $\tau > 0$ – коэффициент регуляризации.
- Если $\|\hat{\mathbf{\Sigma}}\boldsymbol{\theta}^*\|_1 \geq \delta$, то останавливаемся, поскольку найденный набор не является δ-мультиколлинеарным. Иначе переходим на шаг 3.
- Вычисляем невязку $\boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\theta}^*$ и набор активных признаков $\mathcal{J} = \{j : \theta_j^* \neq 0\}$.
- Пусть $j_0 \in \mathcal{J}$. Поправим признаки с номерами из $\mathcal{J}$ так, что $\mathbf{X}^{\text{new}}\boldsymbol{\theta}^* = \mathbf{0}$,
$$\mathbf{f}_j - \frac{1}{\mathbf{e}^\intercal\boldsymbol{\theta}^*}\boldsymbol{\varepsilon} \to \mathbf{f}_j,\ j \in \mathcal{J}.$$
- Удаляем признак с номером j_0 и шкалируем признаки с номерами $j \in \mathcal{J} \setminus \{j_0\}$ до дисперсии 1.
- Пересчитываем матрицу ковариации $\hat{\mathbf{\Sigma}}$. Возвращаемся на шаг 1.

Предлагаемый алгоритм имеет два параметра τ и δ, которые предлагается определять с помощью кросс-валидации на обучающей выборке.

Глава 3
Обучение мультимоделей

Ранее было показано, как получить оценки максимума обоснованности для гиперпараметров априорных распределений, то есть получить оптимальную мультимодель. Для того, чтобы предсказывать класс вновь поступивших объектов опишем далее процесс обучения мультимоделей. Получение обученной $(s,\, \alpha)$ – адекватной оптимальной мультимодели и является целью данной работы. Учет требования адекватности будет описан в следующем разделе.

3.1. Обучение одиночной модели

В соответствии с (2.1) совместное правдоподобие для случая одиночной логистической модели имеет вид

$$p(\mathbf{y},\, \mathbf{w}|\mathbf{X},\, \mathbf{A}) = p(\mathbf{y}|\mathbf{X},\, \mathbf{w})p(\mathbf{w}|\mathbf{A}) = \prod_{i=1}^{m} \sigma(y_i \mathbf{w}^\intercal \mathbf{x}_i) N(\mathbf{w}|\mathbf{0},\, \mathbf{A}^{-1}).$$

Тогда при $K = 1$ из (1.6) получаем, что обучение одиночной модели состоит в получении оценки максимума апостериорной вероятности для вектора параметров модели $\mathbf{w}$, то есть

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{y},\, \mathbf{X},\, \mathbf{A}) = \arg\min_{\mathbf{w}} -\log p(\mathbf{w}|\mathbf{y},\, \mathbf{X},\, \mathbf{A}).$$

Для апостериорного распределения на $\mathbf{w}$ имеем

$$p(\mathbf{w}|\mathbf{y},\, \mathbf{X},\, \mathbf{A}) = \frac{p(\mathbf{y},\, \mathbf{w}|\mathbf{X},\, \mathbf{A})}{p(\mathbf{y}|\mathbf{X},\, \mathbf{A})},$$

где знаменатель есть обоснованность одиночной модели и от $\mathbf{w}$ не зависит. Тогда получаем, что

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} -\log p(\mathbf{y},\, \mathbf{w}|\mathbf{X},\, \mathbf{A}). \tag{3.1}$$

Для отрицательного логарифма совместного распределения имеем

$$l(\mathbf{w}) = -\log p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}) = \tfrac{1}{2}\log\det \mathbf{A} + \tfrac{n}{2}\log(2\pi) + \tfrac{1}{2}\mathbf{w}^\intercal \mathbf{A}\mathbf{w} - \sum_{i=1}^{m} \log \sigma(y_i \mathbf{w}^\intercal \mathbf{x}_i).$$

Для градиента и гессиана $l(\mathbf{w})$ имеем

$$\frac{\partial l(\mathbf{w})}{\partial \mathbf{w}} = A\mathbf{w} - \sum_{i=1}^{m} \frac{1}{\sigma(y_i \mathbf{w}^\intercal \mathbf{x}_i)} \sigma(y_i \mathbf{w}^\intercal \mathbf{x}_i)\sigma(-y_i \mathbf{w}^\intercal \mathbf{x}_i) y_i \mathbf{x}_i = A\mathbf{w} - \sum_{i=1}^{m} \sigma(-y_i \mathbf{w}^\intercal \mathbf{x}_i) y_i \mathbf{x}_i, \tag{3.2}$$

$$\frac{\partial^2 l(\mathbf{w})}{\partial \mathbf{w}^2} = \mathbf{A} - \sum_{i=1}^{m} y_i \mathbf{x}_i \sigma(y_i \mathbf{w}^\intercal \mathbf{x}_i)\sigma(-y_i \mathbf{w}^\intercal \mathbf{x}_i)(-y_i \mathbf{x}_i^\intercal) = \mathbf{A} + \sum_{i=1}^{m} \sigma(\mathbf{w}^\intercal \mathbf{x}_i)\sigma(-\mathbf{w}^\intercal \mathbf{x}_i)\mathbf{x}_i \mathbf{x}_i^\intercal. \tag{3.3}$$

Введем обозначение $\mathbf{R} = \operatorname{diag}(\sigma(\mathbf{w}^\intercal \mathbf{x}_i)\sigma(-\mathbf{w}^\intercal \mathbf{x}_i))$, $i = \overline{1,\ m}$, тогда для гессиана отрицательного логарифма правдоподобия получаем следующую формулу

$$\frac{\partial^2 l(\mathbf{w})}{\partial \mathbf{w}^2} = \mathbf{A} + \mathbf{X}^\intercal \mathbf{R} \mathbf{X},$$

откуда в силу неотрицательной определенности матрицы $\mathbf{A}$ получаем, что функция $l(\mathbf{w})$ является выпуклой, а потому имеет единственный глобальный минимум. Для его нахождения можно воспользоваться, например, методом Ньютона-Рафсона [21] или его демпфированной версией [**?**].

3.2. Обучение многоуровневой модели

В соответствии с (2.49) совместное правдоподобие для случая многоуровневой модели из логистических моделей имеет вид

$$p(\mathbf{y}, \mathbf{w}_1, \ldots, \mathbf{w}_K | \mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K) = \prod_{k=1}^{K} \left[\frac{\sqrt{\det \mathbf{A}_k}}{(2\pi)^{n/2}} \exp\left(-\tfrac{1}{2}\mathbf{w}_k^\intercal \mathbf{A}_k \mathbf{w}_k\right) \prod_{i \in \mathcal{I}_k} \sigma(y_i \mathbf{w}_k^\intercal \mathbf{x}_i) \right], \text{ гд}$$

$\{1,\ \ldots,\ m\} = \mathcal{I} = \mathcal{I}_1 \sqcup \ldots \sqcup \mathcal{I}_K$ есть разбиение множества индексов объектов выборки на непересекающиеся множества по их принадлежности областям действия Ω_k, $k = \overline{1,\ K}$ каждой из моделей, $\mathbb{R}^n = \Omega_1 \sqcup \ldots \sqcup \Omega_K$.

Тогда из (1.6) в силу представления совместного правдоподобия многоуровневой модели в виде произведения K компонент, каждая из которых содержит только один из $\mathbf{w}_k$, получаем, что обучение многоуровневой модели состоит в решении K независимых задач вида

$$\mathbf{w}_k^* = \arg\max_{\mathbf{w}_k} \frac{\sqrt{\det \mathbf{A}_k}}{(2\pi)^{n/2}} \exp\left(-\tfrac{1}{2}\mathbf{w}_k^\intercal \mathbf{A}_k \mathbf{w}_k\right) \prod_{i \in \mathcal{I}_k} \sigma(y_i \mathbf{w}_k^\intercal \mathbf{x}_i),\ k = \overline{1,\ K},$$

что эквивалентно задаче (3.1) на усеченном множестве объектов

$$\mathbf{w}_k^* = \arg\max_{\mathbf{w}_k} p(\mathbf{y}_{\mathcal{I}_k},\ \mathbf{w}_k | \mathbf{X}_{\mathcal{I}_k},\ \mathbf{A}_k) = \arg\min_{\mathbf{w}_k} -\log p(\mathbf{y}_{\mathcal{I}_k},\ \mathbf{w}_k | \mathbf{X}_{\mathcal{I}_k},\ \mathbf{A}_k),\ k = \overline{1,\ K}. \tag{3.4}$$

Таким образом, обучение многоуровневой модели сводится к обучению K одиночных моделей на непересекающихся множествах объектов.

3.3. Обучение смеси моделей

В соответствии с (1.7) совместное правдоподобие для случая смеси моделей из логистических моделей имеет вид

$$p(\mathbf{y},\,\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K|\mathbf{X},\,\alpha,\,\mathbf{A}_1,\,\ldots,\,\mathbf{A}_K) = \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)}\prod_{k=1}^{K}\pi_k^{\alpha-1}\prod_{k=1}^{K}\frac{\sqrt{\det\mathbf{A}_l}}{(2\pi)^{n/2}}\exp\left(-\tfrac{1}{2}\mathbf{w}_l^{\intercal}\mathbf{A}_l\mathbf{w}_l\right)\prod_{i=1}^{m}\left(\sum_{l=1}^{K}\pi_l\sigma(y_i\mathbf{w}_l^{\intercal}\mathbf{x}_i)\right).$$

Тогда из (1.5) обучение смеси моделей состоит в поиске оценок максимума апостериорной вероятности на вектор весов $\boldsymbol{\pi}$ моделей, входящих в смесь, и на векторы параметров этих моделей $\mathbf{w}_k$, $k=\overline{1,\,K}$, то есть

$$[\boldsymbol{\pi}^*,\,\mathbf{w}_1^*,\,\ldots,\,\mathbf{w}_K^*] = \underset{\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K}{\arg\max}\; p(\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K|\mathbf{y},\,\mathbf{X},\,\alpha,\,\mathbf{A}_1,\,\ldots,\,\mathbf{A}_K). \quad (3.5)$$

Для апостериорного распределения $p(\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K|\mathbf{y},\,\mathbf{X},\,\alpha,\,\mathbf{A}_1,\,\ldots,\,\mathbf{A}_K)$ имеем

$$p(\boldsymbol{\pi},\mathbf{w}_1,\ldots,\mathbf{w}_K|\mathbf{y},\mathbf{X},\alpha,\mathbf{A}_1,\ldots,\mathbf{A}_K) = \frac{p(\mathbf{y},\,\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K|\mathbf{X},\,\alpha,\,\mathbf{A}_1,\,\ldots,\,\mathbf{A}_K)}{p(\mathbf{y}|\mathbf{X},\,\alpha,\,\mathbf{A}_1,\,\ldots,\,\mathbf{A}_K)},$$

где знаменатель $p(\mathbf{y}|\mathbf{X},\,\alpha,\,\mathbf{A}_1,\,\ldots,\,\mathbf{A}_K)$ есть обоснованность смеси моделей и от $\boldsymbol{\pi}$, $\mathbf{w}_1$, $\ldots$, $\mathbf{w}_K$ не зависит. Отсюда задача (3.5) эквивалентна следующей

$$[\boldsymbol{\pi}^*,\,\mathbf{w}_1^*,\,\ldots,\,\mathbf{w}_K^*] = \underset{\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K}{\arg\max}\;\log p(\mathbf{y},\,\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K|\mathbf{X},\,\alpha,\,\mathbf{A}_1,\,\ldots,\,\mathbf{A}_K).$$

Введем матрицу скрытых переменных $\mathbf{Z}$ размера $m\times K$, где $z_{ik}\in\{0,1\}$ и $z_{ik}=1$ тогда и только тогда, когда объект $(\mathbf{x}_i,\,y_i)$ отнесен к модели с номером k. Тогда совместное правдоподобие для модели со скрытыми переменными примет вид

$$p(\mathbf{y},\boldsymbol{\pi},\mathbf{w}_1,\ldots,\mathbf{w}_K,\mathbf{Z}|\mathbf{X},\alpha,\mathbf{A}_1,\ldots,\mathbf{A}_k) = p(\boldsymbol{\pi}|\alpha)\prod_{k=1}^{K}p_k(\mathbf{w}_k|\mathbf{A}_k)\prod_{i=1}^{m}\prod_{l=1}^{K}(\pi_l\sigma(y_i\mathbf{w}_l^{\intercal}\mathbf{x}_i))^{z_{il}}$$

Оценим $\log p(\mathbf{y},\,\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K|\mathbf{X},\,\alpha,\,\mathbf{A}_1,\,\ldots,\,\mathbf{A}_K)$ снизу, вводя распределение $q(\mathbf{Z})$, получим

$$\log p(\mathbf{y},\,\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K|\mathbf{X},\,\alpha,\,\mathbf{A}_1,\,\ldots,\,\mathbf{A}_k) \geq L(q,\,\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K) = \mathbb{E}_q\log p(\mathbf{y},\,\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K,\,\mathbf{Z}|\mathbf{X},\,\alpha,\,\mathbf{A}_1,\,\ldots,\,\mathbf{A}_k) - \mathbb{E}_q\log q(\mathbf{Z}),\ \text{где}$$

$$\log p(\mathbf{y},\boldsymbol{\pi},\mathbf{w}_1,\ldots,\mathbf{w}_K,\mathbf{Z}|\mathbf{X},\alpha,\mathbf{A}_1,\ldots,\mathbf{A}_k) = \log(K\alpha) - K\log\alpha + \sum_{k=1}^{K}(\alpha-1)\log\pi_k +$$

$$\tfrac{1}{2}\sum_{k=1}^{K}\log\det\mathbf{A}_k - \tfrac{Kn}{2}\log(2\pi) - \tfrac{1}{2}\sum_{k=1}^{K}\mathbf{w}_k^{\intercal}\mathbf{A}_k\mathbf{w}_k + \sum_{l=1}^{K}\log\pi_l\sum_{i=1}^{m}z_{il} + \sum_{l=1}^{K}\sum_{i=1}^{m}z_{il}\log\sigma(y_i\mathbf{w}_l^{\intercal}\mathbf{x}_i)$$

Будем теперь решать задачу

$$L(q,\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K) \to \max_{q,\,\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K}. \tag{3.6}$$

Для этого воспользуемся вариационным EM-алгоритмом, то есть решение для q будем искать в классе распределений, которые имеют следующую факторизацию.

$$Q = \left\{ q :\ q(\mathbf{Z}) = \prod_{i=1}^{m} q_{\mathbf{z}_i}(\mathbf{z}_i) \right\}.$$

E-шаг.
На E-шаге производится максимизация $L(q,\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K)$ по $q(\mathbf{Z})$ в классе Q. Тогда получаем для $p_{\mathbf{z}_i}(\mathbf{z}_i)$ следующие выражения.

$$\log p_{\mathbf{z}_i}(\mathbf{z}_i) = \mathbb{E}_{q\setminus\mathbf{z}_i} \log p(\mathbf{y}, \boldsymbol{\pi}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{Z}|\mathbf{X}, \alpha, \mathbf{A}_1, \ldots, \mathbf{A}_k) \propto \sum_{l=1}^{K} z_{il}(\log \pi_l + \log \sigma(y_i \mathbf{w}$$

где пропорциональность приведена с точностью до аддитивной постоянной. Отсюда получаем, что

$$\mathbb{P}(z_{il} = 1) = \frac{\pi_l \sigma(y_i \mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i)}{\sum_{k=1}^{K} \pi_k \sigma(y_i \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i)}.$$

M-шаг.
На M-шаге EM-алгоритма производим максимизацию $L(q,\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K)$ по $\boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K$, что приводит к задаче

$$\mathbb{E}_q \log p(\mathbf{y},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \mathbf{Z}|\mathbf{X},\ \alpha,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_k) \to \max_{\boldsymbol{\pi},\,\mathbf{w}_1,\,\ldots,\,\mathbf{w}_K},$$

что эквивалентно следующей задаче максимизации

$$\sum_{k=1}^{K}(\alpha-1)\log\pi_k - \tfrac{1}{2}\sum_{k=1}^{K}\mathbf{w}_k^{\mathsf{T}}\mathbf{A}_k\mathbf{w}_k + \sum_{l=1}^{K}\log\pi_l \sum_{i=1}^{m}\mathbb{E}z_{il} + \sum_{l=1}^{K}\sum_{i=1}^{m}\mathbb{E}z_{il}\log\sigma(y_i\mathbf{w}_l^{\mathsf{T}}\mathbf{x}_i) \to \max_{\boldsymbol{\pi},\,\mathbf{w}_1,} \tag{3.7}$$

Так как максимизируемая функция представляет собой сумму функций, каждая из которых зависит только от одной из переменных $\boldsymbol{\pi}$, $\mathbf{w}_1$, $\ldots$, $\mathbf{w}_K$, данная задача максимизации распадается на $K+1$ независимую подзадачу вида

$$\sum_{k=1}^{K}(\alpha - 1 + \sum_{i=1}^{m}\mathbb{E}z_{ik})\log\pi_k \to \max_{\boldsymbol{\pi}} \text{ при условии } \sum_{l=1}^{K}\pi_l = 1, \tag{3.8}$$

$$l_k(\mathbf{w}_k) = \tfrac{1}{2}\sum_{k=1}^{K}\mathbf{w}_k^{\mathsf{T}}\mathbf{A}_k\mathbf{w}_k - \sum_{i=1}^{m}\mathbb{E}z_{ik}\log\sigma(y_i\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i) \to \min_{\mathbf{w}_k},\ k = \overline{1,\ K}. \tag{3.9}$$

Задача (3.8) имеет аналитическое решение вида

$$\pi_k = \begin{cases} 0, \text{ если } \alpha - 1 + \sum_{i=1}^{m} \mathbb{E} z_{ik} \leq 0, \\ \dfrac{\alpha - 1 + \sum_{i=1}^{m} \mathbb{E} z_{ik}}{\sum_{l=1}^{K} \max(0,\ \alpha - 1 + \sum_{i=1}^{m} \mathbb{E} z_{il})}, \text{ если } \alpha - 1 + \sum_{i=1}^{m} \mathbb{E} z_{ik} > 0, \end{cases} \tag{3.10}$$

то есть на М-шаге происходит прореживание смеси моделей и удаление избыточных компонент из смеси. Задачи (3.9) представляют собой задачу поиска оценки максимума апостериорной вероятности для одиночной логистической модели (3.1), но со взвешенными объектами с неотрицательными весами. В силу неотрицательности весов оптимизируемая функция в (3.9) является выпуклой, а потому имеет единственный минимум, для нахождения которого, как и в случае одиночной модели, можно воспользоваться, например, методом Ньютона-Рафсона [21] или его демпфированной версией [**?**]. Получим выражения для градиента и гессиана в задаче (3.9) аналогичные (3.2) и (3.3).

$$\frac{\partial l_k(\mathbf{w}_k)}{\partial \mathbf{w}_k} = A_k \mathbf{w}_k - \sum_{i=1}^{m} \mathbb{E} z_{ik} \sigma(-y_i \mathbf{w}^\intercal \mathbf{x}_i) y_i \mathbf{x}_i, \tag{3.11}$$

$$\frac{\partial^2 l(\mathbf{w}_k)}{\partial \mathbf{w}_k^2} = \mathbf{A}_k + \sum_{i=1}^{m} \mathbb{E} z_{ik} \sigma(\mathbf{w}^\intercal \mathbf{x}_i) \sigma(-\mathbf{w}^\intercal \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i^\intercal. \tag{3.12}$$

3.4. Алгоритм совместного обучения и оптимизации смеси моделей

Для того, чтобы построить обученную оптимальную мультимодель требуется сначала найти оценки максимума обоснованности для ковариационных матриц $\mathbf{A}_1^*, \ldots, \mathbf{A}_K^*$ (1.3), а затем обучить смесь моделей при фиксированных гиперпараметрах априорных распределений в соответствии с (3.5). Предложим далее алгоритм, который позволяет одновременно с обучением смеси моделей производить оценку гиперпараметров априорных распределений.

Считаем, как и ранее, что параметр α априорного распределения Дирихле фиксирован. Действуем аналогично приведенному выше алгоритму обучения смеси моделей, но считаем, что матрицы $\mathbf{A}_1, \ldots, \mathbf{A}_K$ неизвестны. Введем матрицу скрытых переменных $\mathbf{Z}$ размера $m \times K$, где $z_{ik} \in \{0,\ 1\}$ и $z_{ik} = 1$ тогда и только тогда, когда объект $(\mathbf{x}_i,\ y_i)$ отнесен к модели с номером k. Тогда совместное правдоподобие для модели со скрытыми переменными примет вид

$$p(\mathbf{y}, \boldsymbol{\pi}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{Z} | \mathbf{X}, \alpha, \mathbf{A}_1, \ldots, \mathbf{A}_k) = p(\boldsymbol{\pi}|\alpha) \prod_{k=1}^{K} p_k(\mathbf{w}_k | \mathbf{A}_k) \prod_{i=1}^{m} \prod_{l=1}^{K} \left(\pi_l \sigma(y_i \mathbf{w}_l^\intercal \mathbf{x}_i)\right)^{z_{il}}$$

Оценим $\log p(\mathbf{y},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K | \mathbf{X},\ \alpha,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K)$ снизу, вводя распределение

$q(\mathbf{Z})$, получим

$$\log p(\mathbf{y},\, \boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K|\mathbf{X},\, \alpha,\, \mathbf{A}_1,\, \ldots,\, \mathbf{A}_k) \geq L(q,\, \boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K) = \mathbb{E}_q \log p(\mathbf{y},\, \boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K,\, \mathbf{Z}|\mathbf{X},\, \alpha,\, \mathbf{A}_1,\, \ldots,\, \mathbf{A}_k) - \mathbb{E}_q \log q(\mathbf{Z}),\ \text{где}$$

$$\log p(\mathbf{y}, \boldsymbol{\pi}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{Z}|\mathbf{X}, \alpha, \mathbf{A}_1, \ldots, \mathbf{A}_k) = \log(K\alpha) - K\log\alpha + \sum_{k=1}^{K}(\alpha-1)\log\pi_k +$$

$$\frac{1}{2}\sum_{k=1}^{K}\log\det\mathbf{A}_k - \frac{Kn}{2}\log(2\pi) - \frac{1}{2}\sum_{k=1}^{K}\mathbf{w}_k^{\mathsf{T}}\mathbf{A}_k\mathbf{w}_k + \sum_{l=1}^{K}\log\pi_l\sum_{i=1}^{m}z_{il} + \sum_{l=1}^{K}\sum_{i=1}^{m}z_{il}\log\sigma(y_i\mathbf{w}_l^{\mathsf{T}}\mathbf{x}_i)$$

Будем теперь решать задачу

$$L(q,\, \boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K) \to \max_{q,\, \boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K}. \tag{3.13}$$

Отметим, что в задаче (3.13) в отличие от задачи (3.6) содержатся неизвестные матрицы $\mathbf{A}_1,\, \ldots,\, \mathbf{A}_K$, которые определяются в соответствии с принципом максимума обоснованности (1.3), и от их значений зависят полученные оценки максимума апостериорной вероятности. Для решения задачи (3.13) воспользуемся вариационным EM-алгоритмом, то есть решение для q будем искать в классе распределений, которые имеют следующую факторизацию.

$$Q = \left\{ q:\ q(\mathbf{Z}) = \prod_{i=1}^{m} q_{\mathbf{z}_i}(\mathbf{z}_i) \right\}.$$

E-шаг.
На E-шаге производится максимизация $L(q,\, \boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K)$ по $q(\mathbf{Z})$ в классе Q. Тогда получаем для $p_{\mathbf{z}_i}(\mathbf{z}_i)$ следующие выражения.

$$\log p_{\mathbf{z}_i}(\mathbf{z}_i) = \mathbb{E}_{q\setminus\mathbf{z}_i}\log p(\mathbf{y}, \boldsymbol{\pi}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{Z}|\mathbf{X}, \alpha, \mathbf{A}_1, \ldots, \mathbf{A}_k) \propto \sum_{l=1}^{K} z_{il}(\log\pi_l + \log\sigma(y_i\mathbf{w}$$

где пропорциональность приведена с точностью до аддитивной постоянной. Отсюда получаем, что

$$\mathbb{P}(z_{il} = 1) = \frac{\pi_l\sigma(y_i\mathbf{w}_l^{\mathsf{T}}\mathbf{x}_i)}{\sum_{k=1}^{K}\pi_k\sigma(y_i\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i)}.$$

M-шаг.
На M-шаге EM-алгоритма производим максимизацию $L(q,\, \boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K)$ по $\boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K$, что приводит к задаче

$$\mathbb{E}_q \log p(\mathbf{y},\, \boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K,\, \mathbf{Z}|\mathbf{X},\, \alpha,\, \mathbf{A}_1,\, \ldots,\, \mathbf{A}_k) \to \max_{\boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K}, \tag{3.14}$$

в которой, однако, есть неизвестные $\mathbf{A}_1,\ \ldots,\ \mathbf{A}_K$, являющиеся решением задачи (1.3). Воспользуемся нижней оценкой $\mathbb{E}_q \log p(\mathbf{y}, \boldsymbol{\pi}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \mathbf{Z}|\mathbf{X}, \alpha, \mathbf{A}_1, \ldots, \mathbf{A}_k) - \mathbb{E}_q \log q(\mathbf{Z})$ как аппроксимацией $\log p(\mathbf{y},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K|\mathbf{X},\ \alpha,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_k)$, тогда

$$p(\mathbf{y}, \boldsymbol{\pi}, \mathbf{w}_1, \ldots, \mathbf{w}_K|\mathbf{X}, \alpha, \mathbf{A}_1, \ldots, \mathbf{A}_k) \approx e^{-\mathbb{E}_q \log q(\mathbf{Z})} p(\boldsymbol{\pi}|\alpha) \prod_{k=1}^{K} N(\mathbf{w}_k|\mathbf{0}, \mathbf{A}_k^{-1}) \prod_{k=1}^{K} \pi_k^{\sum_{i=1}^{m} \mathbb{E}}$$
$$\prod_{k=1}^{K} \prod_{i=1}^{m} \sigma(y_i \mathbf{w}_k^{\intercal} \mathbf{x}_i)^{\mathbb{E} z_{ik}}. \quad (3.15)$$

Тогда из (3.15) для обоснованности $p(\mathbf{y}|\mathbf{X},\ \alpha,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K)$ имеем

$$p(\mathbf{y}|\mathbf{X},\ \alpha,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K) \approx e^{-\mathbb{E}_q \log q(\mathbf{Z})} \int p(\boldsymbol{\pi}|\alpha) \prod_{k=1}^{K} \pi_k^{\sum_{i=1}^{m} \mathbb{E} z_{ik}} d\boldsymbol{\pi}$$
$$\prod_{k=1}^{K} \int N(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1}) \prod_{i=1}^{m} \sigma(y_i \mathbf{w}_k^{\intercal} \mathbf{x}_i)^{\mathbb{E} z_{ik}} d\mathbf{w}_k. \quad (3.16)$$

Для интеграла по $\boldsymbol{\pi}$ имеем

$$\int p(\boldsymbol{\pi}|\alpha) \prod_{k=1}^{K} \pi_k^{\sum_{i=1}^{m} \mathbb{E} z_{ik}} d\boldsymbol{\pi} = \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \int \prod_{k=1}^{K} \pi_k^{\alpha - 1 + \sum_{i=1}^{K} \mathbb{E} z_{ik}} = \frac{\Gamma(K\alpha) \prod_{k=1}^{K} \Gamma(\alpha + \sum_{i=1}^{m} \mathbb{E} z}{\Gamma^K(\alpha) \Gamma(K\alpha + m)}$$

С учетом факторизации по $\mathbf{A}_k,\ k = \overline{1,\ K}$ в (3.16) получаем для оценок максимума обоснованности $\mathbf{A}_1 *,\ \ldots,\ \mathbf{A}_K^*$

$$\mathbf{A}_k^* = \arg\max_{\mathbf{A}_k} \int N(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1}) \prod_{i=1}^{m} \sigma(y_i \mathbf{w}_k^{\intercal} \mathbf{x}_i)^{\mathbb{E} z_{ik}} d\mathbf{w}_k,\ k = \overline{1,\ K}. \quad (3.17)$$

Отметим, что (3.17) представляет собой оценку максимума обоснованности для ковариационной матрицы одиночной логистической модели с экспоненциально взвешенными объектами с вектором весов $\boldsymbol{\gamma}_k = [\mathbb{E} z_{ik},\ i = \overline{1,\ m}]^{\intercal}$. Решение задачи (3.17) рассмотрено ранее для одиночной логистической модели, а получение аппроксимации оценки максимума обоснованности $\mathbf{A}_k^*$ может быть выполнено одним из трех описанных способов: с помощью аппроксимации Лапласа, с помощью вариационной нижней оценки и с помощью EM-алгоритма с вариационной нижней оценкой.

После нахождения $\mathbf{A}_1^*,\ \ldots,\ \mathbf{A}_K^*$ задача (3.14) становится эквивалентна ранее рассмотренной задаче M-шага обучения смеси моделей (3.7), что приводит к оценке (3.10) для $\boldsymbol{\pi}$ и к задачам минимизации (3.9) выпуклой функций $l_k(\mathbf{w}_k),\ k = \overline{1,\ K}$ с градиентом и гессианом, дающимися формулами (3.11) и (3.12), где $\mathbf{A}_k,\ k = \overline{1,\ K}$ заменено на оценки максимума обоснованности для ковариационных матриц $\mathbf{A}_k^*,\ k = \overline{1,\ K}$. Таким образом, получаем метод совместного обучения и оптимизации смеси моделей.

Глава 4
Построение $(s,\ \alpha)$ – адекватных мультимоделей

Ранее были рассмотрены задачи нахождения оптимальной мультимодели, а также ее обучения. Отметим, однако, что оптимальная обученная мультимодель может иметь похожие модели. Так для многоуровневой модели обучение производится фактически независимо для каждой из компонент, а потому модели, получающиеся для разных компонент данных, могут быть статистически неразличимыми. Наличие избыточных моделей ведет не только к потере интерпретируемости, но и к понижению качества прогноза новых объектов, поскольку оценки параметров моделей получаются по меньшим выборкам, а потому являются менее точными. В случае смеси моделей, хотя при обучении и происходит прореживание смеси и удаление избыточных моделей, полученная смесь моделей по-прежнему может содержать близкие модели, а потому смесь может не являться адекватной. Потому рассмотрим далее алгоритмы построения адекватной мультимодели по обученной оптимальной.

4.1. Сравнение моделей

Линейная регрессия. Случай независимых моделей

В данной секции будет рассмотрена задача сравнения моделей для линейной регрессии в случае, когда множества объектов, по которым построены модели не пересекаются. Это может иметь место, например, при обучении по огромным выборкам, когда предпочтительнее оказывается обучение по подвыборкам, а затем комбинация результатов, например, с помощью простого голосования.

Подход классической статистики.

Рассмотрим две модели линейной регрессии f_1 и f_2. Пусть $\mathbf{w}_1$ и $\mathbf{w}_2$ есть неизвестные параметры, соответствующие моделям f_1 и f_2, а $(\mathbf{X}_1, \mathbf{y}_1)$ и $(\mathbf{X}_2, \mathbf{y}_2)$ есть выборки, полученные из f_1 и f_2 соответственно. Считаем при этом, что объекты в первой и второй выборках разные. Тогда имеем следующие модели

$$\mathbf{y}_1 = \mathbf{X}_1\mathbf{w}_1 + \boldsymbol{\varepsilon}_1,\ \boldsymbol{\varepsilon}_1 \sim N(0,\ \sigma_1^2\mathbf{I}),$$

$$\mathbf{y}_2 = \mathbf{X}_2\mathbf{w}_2 + \boldsymbol{\varepsilon}_2,\ \boldsymbol{\varepsilon}_2 \sim N(0,\ \sigma_2^2\mathbf{I}).$$

Оценки максимума правдоподобия для параметров моделей имеют вид

$$\hat{\mathbf{w}}_k = (\mathbf{X}_k^\intercal\mathbf{X}_k)^{-1}\mathbf{X}_k^\intercal\mathbf{y}_k. \tag{4.1}$$

Они имеют распределения $\hat{\mathbf{w}}_1 \sim N(\mathbf{w}_1,\ \sigma_1^2(\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1})$ и $\hat{\mathbf{w}}_2 \sim N(\mathbf{w}_2,\ \sigma_2^2(\mathbf{X}_2^\intercal\mathbf{X}_2)^{-1})$. Однако параметры $\mathbf{w}_1$ и $\mathbf{w}_2$ неизвестны. В условиях истинности гипотезы H_0 о совпадении параметров моделей, то есть

$\mathbf{w}_1 = \mathbf{w}_2 = \mathbf{w}$, имеем $\hat{\mathbf{w}}_1 \sim N(\mathbf{w},\, \sigma_1^2(\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1})$ и $\hat{\mathbf{w}}_2 \sim N(\mathbf{w},\, \sigma_2^2(\mathbf{X}_2^\intercal\mathbf{X}_2)^{-1})$. Тогда для разности имеем

$$\Delta\mathbf{w} = \hat{\mathbf{w}}_1 - \hat{\mathbf{w}}_2 \sim N(\mathbf{0},\, \sigma_1^2(\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1} + \sigma_2^2(\mathbf{X}_2^\intercal\mathbf{X}_2)^{-1}).$$

Если σ_1^2 и σ_2^2 известны, то

$$s = (\Delta\mathbf{w})^\intercal(\sigma_1^2(\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1} + \sigma_2^2(\mathbf{X}_2^\intercal\mathbf{X}_2)^{-1})^{-1}(\Delta\mathbf{w}) \sim \chi^2(n).$$

Тогда критической областью Ω_α для гипотезы H_0 при уровне значимости α является область $s > t_\alpha^n$, где t_α^n есть $1-\alpha$ квантиль распределения $\chi^2(n)$. Это же можно записать в виде

$$S = \exp(-\tfrac{s}{2}) < \exp(-\tfrac{t_\alpha^n}{2}),$$

где S – значение s-score для данной пары распределений.

Пусть теперь σ_1^2 и σ_2^2 неизвестны. Тогда заменим σ_1^2 и σ_2^2 на их несмещенные оценки $\hat{\sigma}_1^2$ и $\hat{\sigma}_2^2$

$$\hat{\sigma}_1^2 = \frac{(\mathbf{y}_1 - \mathbf{X}_1\hat{\mathbf{w}}_1)^\intercal(\mathbf{y}_1 - \mathbf{X}_1\hat{\mathbf{w}}_1)}{m_1 - n},\ \hat{\sigma}_2^2 = \frac{(\mathbf{y}_2 - \mathbf{X}_2\hat{\mathbf{w}}_2)^\intercal(\mathbf{y}_2 - \mathbf{X}_1\hat{\mathbf{w}}_2)}{m_2 - n}$$

в формуле для статистики s.

Утверждение 5. Пусть выполнена гипотеза H_0 о совпадении параметров моделей. Пусть также объекты поступают таким образом, что

$$\frac{\lambda_{\max}(\mathbf{X}_1^\intercal\mathbf{X}_1)}{\lambda_{\min}(\mathbf{X}_1^\intercal\mathbf{X}_1)} = o(\sqrt{m_1}) \text{ при } m_1 \to \infty,\ \frac{\lambda_{\max}(\mathbf{X}_2^\intercal\mathbf{X}_2)}{\lambda_{\min}(\mathbf{X}_2^\intercal\mathbf{X}_2)} = o(\sqrt{m_2}) \text{ при } m_2 \to \infty.$$

Тогда при $m_1,\, m_2 = \Theta(m_1) \to \infty$ $s \xrightarrow{d} \chi^2(n)$.

Доказательство.

$$s = (\Delta\mathbf{w})^\intercal(\boldsymbol{\Sigma}+\Delta\boldsymbol{\Sigma})^{-1}(\Delta\mathbf{w}) = (\Delta\mathbf{w})\boldsymbol{\Sigma}^{-1}(\Delta\mathbf{w})+(\Delta\mathbf{w})^\intercal\boldsymbol{\Sigma}^{-1}\left(\sum_{k=1}^{\infty}(\Delta\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1})^k\right)(\Delta\mathbf{w}) =$$

где $\boldsymbol{\Sigma} = \sigma_1^2(\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1}+\sigma_2^2(\mathbf{X}_2^\intercal\mathbf{X}_2)^{-1}$, $\Delta\boldsymbol{\Sigma} = (\hat{\sigma}_1^2-\sigma_1^2)(\mathbf{X}_1^\intercal\mathbf{X}_1)^{-1}+(\hat{\sigma}_2^2-\sigma_2^2)(\mathbf{X}_2^\intercal\mathbf{X}_2)^{-1}$, что получено с учетом

$$\|\Delta\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\| \le \|\Delta\boldsymbol{\Sigma}\|\|\boldsymbol{\Sigma}^{-1}\| \le \frac{\|\Delta\boldsymbol{\Sigma}\|}{\|\boldsymbol{\Sigma}\|}\|\boldsymbol{\Sigma}^{-1}\|\|\boldsymbol{\Sigma}\| \le \frac{\|\Delta\boldsymbol{\Sigma}\|}{\|\boldsymbol{\Sigma}\|}\max\left(\frac{\lambda_{\max}(\mathbf{X}_1^\intercal\mathbf{X}_1)}{\lambda_{\min}(\mathbf{X}_1^\intercal\mathbf{X}_1)},\ \frac{\lambda_{\max}(\mathbf{X}\right.$$

$$\le \max\left(\frac{|\hat{\sigma}_1^2-\sigma_1^2|}{\sigma_1^2},\ \frac{|\hat{\sigma}_2^2-\sigma_2^2|}{\sigma_2^2}\right)\max\left(\frac{\lambda_{\max}(\mathbf{X}_1^\intercal\mathbf{X}_1)}{\lambda_{\min}(\mathbf{X}_1^\intercal\mathbf{X}_1)},\ \frac{\lambda_{\max}(\mathbf{X}_2^\intercal\mathbf{X}_2)}{\lambda_{\min}(\mathbf{X}_2^\intercal\mathbf{X}_2)}\right) \xrightarrow{p} 0 \text{ при } m_1, m_2 \to$$

так как из неравенства Чебышева имеем

$$\mathbb{P}(|\hat{\sigma}_1^2 - \sigma_1^2| \ge \varepsilon\sigma_1^2) \le \frac{\mathbb{D}{\hat{\sigma}_1}^2}{\varepsilon^2\sigma_1^4} = \frac{2}{\varepsilon^2(m_1-n)},$$

$$\mathbb{P}(|\hat{\sigma}_2^2 - \sigma_2^2| \geq \varepsilon\sigma_2^2) \leq \frac{\mathbb{D}{\hat{\sigma}_2}^2}{\varepsilon^2\sigma_2^4} = \frac{2}{\varepsilon^2(m_2 - n)},$$

откуда

$$\mathbb{P}\left(\max\left(\frac{|\hat{\sigma}_1^2 - \sigma_1^2|}{\sigma_1^2},\ \frac{|\hat{\sigma}_2^2 - \sigma_2^2|}{\sigma_2^2}\right) \geq \varepsilon\right) = O\left(\frac{1}{m_1} + \frac{1}{m_2}\right).$$

Тогда для второго слагаемого в выражении для s имеем по вероятности

$$\left|(\Delta\mathbf{w})^\intercal\boldsymbol{\Sigma}^{-1}\left(\sum_{k=1}^{\infty}(\Delta\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1})^k\right)(\Delta\mathbf{w})\right| \leq 2\|\Delta\mathbf{w}\|\|\boldsymbol{\Sigma}^{-1}\|\|\Delta\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\|\|\Delta\mathbf{w}\|$$

Отсюда имеем по вероятности для второго слагаемого

$$\Delta s = \left(\frac{1}{m_1} + \frac{1}{m_2}\right)\max\left(\frac{\lambda_{\max}(\mathbf{X}_1^\intercal\mathbf{X}_1)}{\lambda_{\min}(\mathbf{X}_1^\intercal\mathbf{X}_1)},\ \frac{\lambda_{\max}(\mathbf{X}_2^\intercal\mathbf{X}_2)}{\lambda_{\min}(\mathbf{X}_2^\intercal\mathbf{X}_2)}\right)^2 = o(1)\text{ при } m_1,\ m_2 \to \infty.$$

Таким образом, имеем требуемое. □

Обобщим предыдущее утверждение и сформулируем его безотносительно к конкретной решаемой задаче различения моделей.

Утверждение 6. Пусть имеется последовательность многомерных нормальных случайных векторов $\mathbf{w}_m \sim N(\mathbf{w}_0,\ \boldsymbol{\Sigma}_m)$, $\mathbf{w}_m \in \mathbb{R}^n$. Рассмотрим $s_m = (\mathbf{w}_m - \mathbf{v}_m)^\intercal(\boldsymbol{\Sigma}_m + \Delta\boldsymbol{\Sigma}_m)^{-1}(\mathbf{w}_m - \mathbf{v}_m)$. Пусть $\|\Delta\boldsymbol{\Sigma}_m\|\|\boldsymbol{\Sigma}_m^{-1}\|q_1(m) \xrightarrow{p} 0$ при $m \to \infty$, $\lambda_{\max}(\boldsymbol{\Sigma}_m)/\lambda_{\min}(\boldsymbol{\Sigma}_m) = o(q_1(m))$, причем $q_1(m) \geq 1$ есть возрастающая функция и $q_1(m) \to \infty$ при $m \to \infty$. Пусть $\mathbf{v}_m$ есть последовательность случайных векторов, для которой имеется следующая сходимость

$$\|\mathbf{v}_m - \mathbf{w}_0\|^2\|\boldsymbol{\Sigma}_m^{-1}\|m^\beta \xrightarrow{p} 0.$$

Тогда $s_m \xrightarrow{d} \chi^2(n)$.

Доказательство. Перепишем s_m в следующем виде

$$s_m = (\mathbf{w}_m - \mathbf{v}_m)^\intercal(\boldsymbol{\Sigma}_m + \Delta\boldsymbol{\Sigma}_m)^{-1}(\mathbf{w}_m - \mathbf{v}_m) = (\mathbf{w}_m - \mathbf{w}_0)^\intercal(\boldsymbol{\Sigma}_m + \Delta\boldsymbol{\Sigma}_m)^{-1}(\mathbf{w}_m - \mathbf{w}_0) +$$
$$+ (\mathbf{w}_0 - \mathbf{v}_m)^\intercal(\boldsymbol{\Sigma}_m + \Delta\boldsymbol{\Sigma}_m)^{-1}(\mathbf{w}_0 - \mathbf{v}_m) + 2(\mathbf{w}_0 - \mathbf{v}_m)^\intercal(\boldsymbol{\Sigma}_m + \Delta\boldsymbol{\Sigma}_m)^{-1}(\mathbf{w}_m - \mathbf{w}_0) =$$
$$= s' + s'' + s'''.$$

Покажем, что последние два слагаемых в сумме сходятся по вероятности к нулю.

$$s'' = \|(\mathbf{w}_0 - \mathbf{v}_m)^\intercal(\boldsymbol{\Sigma}_m + \Delta\boldsymbol{\Sigma}_m)^{-1}(\mathbf{w}_0 - \mathbf{v}_m)\| \leq \|\mathbf{w}_0 - \mathbf{v}_m\|^2\|(\boldsymbol{\Sigma}_m + \Delta\boldsymbol{\Sigma}_m)^{-1}\| \leq$$
$$\leq \|\mathbf{w}_0 - \mathbf{v}_m\|^2\|\boldsymbol{\Sigma}_m^{-1}\|\|(\mathbf{I} + \boldsymbol{\Sigma}_m^{-1}\Delta\boldsymbol{\Sigma}_m)^{-1}\|.$$

Рассмотрим только те исходы Ω', для которых $\|\Delta\mathbf{\Sigma}_m\|\|\mathbf{\Sigma}_m^{-1}\| \leq 1/2$. Вероятность таких исходов $\mathbb{P}(\Omega') \to 1$ при $m \to \infty$ в силу $\gamma > 0$. Тогда

$$\|(\mathbf{I} + \mathbf{\Sigma}_m^{-1}\Delta\mathbf{\Sigma}_m)^{-1}\| \leq \frac{1}{1 - \|\mathbf{\Sigma}_m^{-1}\|\|\Delta\mathbf{\Sigma}_m\|} \leq 2.$$

Окончательно в силу $\mathbb{P}(\Omega') \to 1$ при $m \to \infty$ имеем

$$\|(\mathbf{w}_0 - \mathbf{v}_m)^\intercal(\mathbf{\Sigma}_m + \Delta\mathbf{\Sigma}_m)^{-1}(\mathbf{w}_0 - \mathbf{v}_m)\| \xrightarrow{p} 0 \text{ при } m \to \infty.$$

Для последнего слагаемого s''' для тех же Ω' имеем

$$s''' = (\mathbf{w}_0 - \mathbf{v}_m)^\intercal(\mathbf{\Sigma}_m + \Delta\mathbf{\Sigma}_m)^{-1}(\mathbf{w}_m - \mathbf{w}_0) \leq 2\|\mathbf{w}_0 - \mathbf{v}_m\|\|\mathbf{\Sigma}_m^{-1}\|\|\mathbf{w}_m - \mathbf{w}_0\|.$$

Для $\|\mathbf{w}_m - \mathbf{w}_0\|$ имеем

$$\mathbb{P}(\|\mathbf{w}_m^j - \mathbf{w}_0^j\| \geq g(m)\sqrt{\|\mathbf{\Sigma}_m\|}) \to 0 \text{ при } m \to \infty$$
$$\text{для любой положительной функции } g(m) : g(m) \to \infty \text{ при } m \to \infty. \quad (4.2)$$

Выберем те исходы Ω'' из Ω', для которых неравенство $\|\mathbf{w}_m^j - \mathbf{w}_0^j\| \leq g(m)\sqrt{\|\mathbf{\Sigma}_m\|}$ выполнено для всех $j = 1, \ldots, n$. Для таких исходов имеем

$$s''' \leq 2\|\mathbf{\Sigma}_m^{-1}\|\|\mathbf{w}_0 - \mathbf{v}_m\|g(m)\sqrt{n}\sqrt{\|\mathbf{\Sigma}_m\|} = 2g(m)\sqrt{\|\mathbf{w}_0 - \mathbf{v}_m\|^2\|\mathbf{\Sigma}_m^{-1}\|\,(\|\mathbf{\Sigma}_m^{-1}\|\|\mathbf{\Sigma}_m\|)}$$

В силу $\|\mathbf{\Sigma}_m^{-1}\|\|\mathbf{\Sigma}_m\| = O(q_1(m))$ выберем

$$g(m) = \left(\frac{q_1(m)}{\|\mathbf{\Sigma}_m^{-1}\|\|\mathbf{\Sigma}_m\|}\right)^{1/2},$$

тогда

$$s''' \leq 2\sqrt{\|\mathbf{w}_0 - \mathbf{v}_m\|^2\|\mathbf{\Sigma}_m^{-1}\|q_1(m)} \xrightarrow{p} 0,$$

откуда имеем требуемое из условия и $\mathbb{P}(\Omega'') \to 1$ при $m \to \infty$. Рассмотрим теперь s'.

$$s' = (\mathbf{w}_m - \mathbf{w}_0)^\intercal(\mathbf{\Sigma}_m + \Delta\mathbf{\Sigma}_m)^{-1}(\mathbf{w}_m - \mathbf{w}_0) = (\mathbf{w}_m - \mathbf{w}_0)\mathbf{\Sigma}_m^{-1}(\mathbf{w}_m - \mathbf{w}_0) +$$
$$+ (\mathbf{w}_m - \mathbf{w}_0)^\intercal\mathbf{\Sigma}_m^{-1}\left(\sum_{k=1}^{\infty}(\Delta\mathbf{\Sigma}_m\mathbf{\Sigma}_m^{-1})^k\right)(\mathbf{w}_m - \mathbf{w}_0) = s_0 + \Delta s.$$

Рассмотрим исходы Ω''', для которых

$$\|\Delta\mathbf{\Sigma}_m\|\|\mathbf{\Sigma}_m^{-1}\| \leq \frac{1}{2q_1(m)}.$$

Из условия имеем $\mathbb{P}(\Omega''') \to 1$ при $m \to \infty$.

$$\Delta s \leq 2\|\mathbf{w}_m - \mathbf{w}_0\|^2\|\mathbf{\Sigma}_m^{-1}\|\|\mathbf{\Sigma}_m^{-1}\|\|\Delta\mathbf{\Sigma}_m\|.$$

Выберем те исходы $\tilde{\Omega}$ из Ω''', для которых неравенство $\|\mathbf{w}_m^j - \mathbf{w}_0^j\| \leq g(m)\sqrt{\|\mathbf{\Sigma}_m\|}$ выполнено для всех $j = 1, \ldots, n$. Для исходов из $\tilde{\Omega}$ имеем

$$\Delta s \leq n g^2(m)(\|\mathbf{\Sigma}_m\|\|\mathbf{\Sigma}_m^{-1}\|)/q_1(m).$$

В силу $\|\mathbf{\Sigma}_m^{-1}\|\|\mathbf{\Sigma}_m\| = o(q_1(m))$ выберем

$$g(m) = \left(\frac{q_1(m)}{\|\mathbf{\Sigma}_m^{-1}\|\|\mathbf{\Sigma}_m\|}\right)^{1/4},$$

тогда

$$\Delta s \leq n\sqrt{\frac{\|\mathbf{\Sigma}_m\|\|\mathbf{\Sigma}_m^{-1}\|}{q_1(m)}} = o(1).$$

Отсюда получаем, что $\Delta s \xrightarrow{p} 0$ при $m \to \infty$. Получаем, что

$$s_m = s + \Delta s + s'' + s''' \xrightarrow{d} \chi^2(n) \text{ при } m \to \infty,$$

поскольку $s \sim \chi^2(n)$, а $\Delta s + s'' + s''' \xrightarrow{p} 0$ при $m \to \infty$, что и требовалось. □

Подход байесовской статистики.

В байесовском подходе параметры линейной регрессии $\mathbf{w}_1$ и $\mathbf{w}_2$ считаются не фиксированными, а получающимися из некоторых априорных распределений $p_1(\mathbf{w}_1)$ и $p_2(\mathbf{w}_2)$. Подход классической статистики при этом можно смоделировать, выбрав в качестве априорных распределений два равномерных псевдораспределения на всем пространстве $\mathbb{R}^n$.

Как и ранее пусть $(\mathbf{X}_1, \mathbf{y}_1)$ и $(\mathbf{X}_2, \mathbf{y}_2)$ есть выборки, полученные из f_1 и f_2 соответственно. Считаем при этом, что объекты в первой и второй выборках разные. Тогда имеем следующие модели

$$\mathbf{y}_1 = \mathbf{X}_1\mathbf{w}_1 + \boldsymbol{\varepsilon}_1,\ \boldsymbol{\varepsilon}_1 \sim N(0,\ \sigma_1^2\mathbf{I}),\ \mathbf{w}_1 \sim p_1(\mathbf{w}_1)$$

$$\mathbf{y}_2 = \mathbf{X}_2\mathbf{w}_2 + \boldsymbol{\varepsilon}_2,\ \boldsymbol{\varepsilon}_2 \sim N(0,\ \sigma_2^2\mathbf{I}),\ \mathbf{w}_2 \sim p_2(\mathbf{w}_2).$$

Считаем далее, что $p_1(\mathbf{w}_1)$ и $p_2(\mathbf{w}_2)$ есть нормальные распределения, то есть

$$p_1(\mathbf{w}_1) = N(\mathbf{w}_1|\mathbf{v}_1,\ \mathbf{\Sigma}_1^{-1}),\ p_2(\mathbf{w}_2) = N(\mathbf{w}_2|\mathbf{v}_2,\ \mathbf{\Sigma}_2^{-1}).$$

Получаем, что функции совместного правдоподобия имеют вид

$$p(\mathbf{y}_k, \mathbf{w}_k|\mathbf{X}_k) = p(\mathbf{y}_k|\mathbf{X}_k, \mathbf{w}_k)p(\mathbf{w}_k) = N(\mathbf{y}_k - \mathbf{X}_k\mathbf{w}_k|\mathbf{0}, \sigma_k^2\mathbf{I})N(\mathbf{w}_k|\mathbf{v}_k, \mathbf{\Sigma}_k^{-1}), k = 1, 2.$$

Пользуясь формулой Байеса, получаем для апостериорного распределения параметров $\mathbf{w}_1$ и $\mathbf{w}_2$

$$p(\mathbf{w}_k|\mathbf{X}_k,\ \mathbf{y}_k) = \frac{p(\mathbf{y}_k,\ \mathbf{w}_k|\mathbf{X}_k)}{p(\mathbf{y}_k|\mathbf{X}_k)} \propto p(\mathbf{y}_k,\ \mathbf{w}_k|\mathbf{X}_k),\ k = 1,\ 2,$$

откуда $p(\mathbf{w}_k|\mathbf{X}_k,\ \mathbf{y}_k) = N(\mathbf{w}_k|\hat{\mathbf{w}}_k,\ \tilde{\mathbf{\Sigma}}_k^{-1})$, где

$$\hat{\mathbf{w}}_k = \left(\mathbf{\Sigma}_k + \frac{1}{\sigma_k^2}\mathbf{X}_k^\intercal\mathbf{X}_k\right)^{-1}\left(\mathbf{\Sigma}_k\mathbf{v}_k + \frac{1}{\sigma_k^2}\mathbf{X}_k^\intercal\mathbf{y}_k\right)$$

есть оценка максимума апостериорной вероятности, а

$$\tilde{\mathbf{\Sigma}}_k = \mathbf{\Sigma}_k + \frac{1}{\sigma_k^2}\mathbf{X}_k^\intercal\mathbf{X}_k,\ k = 1,\ 2.$$

Тогда для s-score для пары апостериорных распределений имеем

$$-2\log s = (\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1)^\intercal\left(\left(\mathbf{\Sigma}_1 + \frac{1}{\sigma_1^2}\mathbf{X}_1^\intercal\mathbf{X}_1\right)^{-1} + \left(\mathbf{\Sigma}_2 + \frac{1}{\sigma_2^2}\mathbf{X}_2^\intercal\mathbf{X}_2\right)^{-1}\right)^{-1}(\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1).$$

Найдем далее распределение $-2\log s$ в условиях истинности гипотезы H_0 о совпадении моделей, то есть в условиях того, что $\mathbf{w}_1 = \mathbf{w}_2 = \mathbf{w}$. В условиях истинности гипотезы H_0 имеем

$$\hat{\mathbf{w}}_k|\mathbf{w} \sim N(\mathbf{m}_k,\ \mathbf{\Omega}_k),\ \text{где}$$

$$\mathbf{m}_k = \mathbf{w} + (\mathbf{\Sigma}_k + \mathbf{A}_k)^{-1}\mathbf{\Sigma}_k(\mathbf{v}_k - \mathbf{w}),$$

$$\mathbf{\Omega}_k = (\mathbf{\Sigma}_k + \mathbf{A}_k)^{-1}(\mathbf{I} - \mathbf{\Sigma}_k(\mathbf{\Sigma}_k + \mathbf{A}_k)^{-1}) = \tilde{\mathbf{\Sigma}}_k^{-1} - \tilde{\mathbf{\Sigma}}_k^{-1}\mathbf{\Sigma}_k\tilde{\mathbf{\Sigma}}_k^{-1},$$

где введено обозначение $\mathbf{A}_k = \frac{1}{\sigma_k^2}\mathbf{X}_k^\intercal\mathbf{X}_k$.

Рассмотрим теперь подробнее выражение для $-2\log s$. Введем обозначения

$$\mathbf{b}_k = \tilde{\mathbf{\Sigma}}_k^{-1}\mathbf{\Sigma}_k(\mathbf{v}_k - \mathbf{w}),\ k = 1,\ 2.$$

Заметим, что при условии $\lambda_{\min}(\mathbf{X}_k^\intercal\mathbf{X}_k) \to \infty$ при $m_k \to \infty$, $k = 1,\ 2$ получим, что $\|\mathbf{b}_k\| \to 0$ при $m_k \to \infty$, так как $\|\tilde{\mathbf{\Sigma}}_k\| \to \infty$ при $m_k \to \infty$. Сформулируем утверждение, аналогичное предыдущим, относительно асимптотического распределения $-2\log s$.

Утверждение 7. Пусть имеются две модели линейной регрессии, описанные выше. Пусть также для некоторой возрастающей функции $q_1(m)$ такой, что $q_1(m) \to \infty$ при $m \to \infty$ выполнено

$$\frac{\lambda_{\max}(\mathbf{X}_k^\intercal\mathbf{X}_k)}{\lambda_{\min}(\mathbf{X}_k^\intercal\mathbf{X}_k)} = o(q_1(m_k))\ \text{при}\ m_k \to \infty,\ k = 1,\ 2,$$

$$\lambda_{\min}(\mathbf{X}_k^\intercal\mathbf{X}_k) = \Omega(q_1^2(m_k))\ \text{при}\ m_k \to \infty,\ k = 1,\ 2.$$

Тогда $-2\log s \xrightarrow{d} \chi^2(n)$ при $m_1,\ m_2 \to \infty$, если $q_1(m_2)/q_1(m_1) = \Theta(1)$.

Доказательство. Покажем, что требуемое выполнено в силу выполнения условий утверждения 2. Введем обозначение $m = \max(m_1, m_2)$. Для этого требуется показать, что выполнены следующие условия при $m_1,\ m_2 \to \infty$

$$\|\mathbf{b}_2 - \mathbf{b}_1\|^2 \|(\mathbf{\Omega}_1 + \mathbf{\Omega}_2)^{-1}\| q_1(m) \to 0, \tag{4.3}$$

$$\frac{\lambda_{max}(\mathbf{\Omega}_1 + \mathbf{\Omega}_2)}{\lambda_{min}(\mathbf{\Omega}_1 + \mathbf{\Omega}_2)} = o(q_1(m)), \tag{4.4}$$

$$\|\Delta\mathbf{\Omega}\| \|(\mathbf{\Omega}_1 + \mathbf{\Omega}_2)^{-1}\| q_1(m) \to 0, \tag{4.5}$$

где

$$\Delta\mathbf{\Omega} = (\mathbf{\Sigma}_1 + \mathbf{A}_1)^{-1}\mathbf{\Sigma}_1(\mathbf{\Sigma}_1 + \mathbf{A}_1)^{-1} + (\mathbf{\Sigma}_2 + \mathbf{A}_2)^{-1}\mathbf{\Sigma}_2(\mathbf{\Sigma}_2 + \mathbf{A}_2)^{-1}.$$

Рассмотрим достаточно большие m_1, m_2 такие, что $\|\mathbf{A}_k\| \geq \|\mathbf{\Sigma}_k\|$ и $\|\mathbf{\Sigma}_k\| \|(\mathbf{\Sigma}_k + \mathbf{A}_k)^{-1}\| \leq 1/2,\ k = 1,\ 2.$

$$\lambda_{\max}(\Delta\mathbf{\Omega}) \leq \|(\mathbf{\Sigma}_1 + \mathbf{A}_1)^{-1}\|^2 \|\mathbf{\Sigma}_1\| + \|(\mathbf{\Sigma}_2 + \mathbf{A}_2)^{-1}\|^2 \|\mathbf{\Sigma}_2\| =$$
$$= \frac{\|\mathbf{\Sigma}_1\|}{\lambda^2_{\min}(\mathbf{\Sigma}_1 + \mathbf{A}_1)} + \frac{\|\mathbf{\Sigma}_2\|}{\lambda^2_{\min}(\mathbf{\Sigma}_2 + \mathbf{A}_2)} \leq \frac{\|\mathbf{\Sigma}_1\|}{\lambda^2_{\min}(\mathbf{A}_1)} + \frac{\|\mathbf{\Sigma}_2\|}{\lambda^2_{\min}(\mathbf{A}_2)}. \tag{4.6}$$

С учетом (4.6) получаем

$$\lambda_{\min}(\mathbf{\Omega}_1 + \mathbf{\Omega}_2) \geq \lambda_{\min}((\mathbf{\Sigma}_1 + \mathbf{A}_1)^{-1} + (\mathbf{\Sigma}_2 + \mathbf{A}_2)^{-1}) - \lambda_{\max}(\Delta\mathbf{\Omega}) \geq \lambda_{\min}((\mathbf{\Sigma}_1 + \mathbf{A}_1)^{-1}) +$$
$$+ \lambda_{\min}((\mathbf{\Sigma}_2 + \mathbf{A}_2)^{-1}) - \lambda_{\max}(\Delta\,\mathbf{\Omega}) = \frac{1}{\lambda_{\max}(\mathbf{\Sigma}_1 + \mathbf{A}_1)} + \frac{1}{\lambda_{\max}(\mathbf{\Sigma}_2 + \mathbf{A}_2)} - \lambda_{\max}(\Delta\mathbf{\Omega}) \geq$$
$$\geq \frac{1}{2\lambda_{\max}(\mathbf{A}_1)} + \frac{1}{2\lambda_{\max}(\mathbf{A}_2)} - \frac{\|\mathbf{\Sigma}_1\|}{\lambda^2_{\min}(\mathbf{A}_1)} - \frac{\|\mathbf{\Sigma}_2\|}{\lambda^2_{\min}(\mathbf{A}_2)}. \tag{4.7}$$

С учетом рассмотрения достаточно больших m_1 и m_2 имеем

$$\lambda_{\max}(\mathbf{\Omega}_1 + \mathbf{\Omega}_2) \leq \lambda_{\max}(\mathbf{\Omega}_1) + \lambda_{\max}(\mathbf{\Omega}_2) \leq 2\lambda_{\max}((\mathbf{\Sigma}_1 + \mathbf{A}_1)^{-1}) + 2\lambda_{\max}((\mathbf{\Sigma}_2 + \mathbf{A}_2)^{-1}) =$$
$$= \frac{2}{\lambda_{\min}(\mathbf{\Sigma}_1 + \mathbf{A}_1)} + \frac{2}{\lambda_{\min}(\mathbf{\Sigma}_2 + \mathbf{A}_2)} \leq \frac{2}{\lambda_{\min}(\mathbf{A}_1)} + \frac{2}{\lambda_{\min}(\mathbf{A}_2)}. \tag{4.8}$$

Выберем теперь m_1 и m_2 так, что

$$\lambda^2_{\min}(\mathbf{A}_k) \geq 4\|\mathbf{\Sigma}_k\| \lambda_{\max}(\mathbf{A}_k),\ k = 1,\ 2.$$

Это возможно сделать, исходя из условия. Тогда для таких m_1 и m_2 с учетом (4.7) и (4.8) получаем

$$\frac{\lambda_{\max}(\mathbf{\Omega}_1 + \mathbf{\Omega}_2)}{\lambda_{\min}(\mathbf{\Omega}_1 + \mathbf{\Omega}_2)} \leq 8\frac{\dfrac{1}{\lambda_{\min}(\mathbf{A}_1)} + \dfrac{1}{\lambda_{\min}(\mathbf{A}_2)}}{\dfrac{1}{\lambda_{\max}(\mathbf{A}_1)} + \dfrac{1}{\lambda_{\max}(\mathbf{A}_2)}} \leq 8\frac{\dfrac{q_1(m_1)c(m_1)}{\lambda_{\max}(\mathbf{A}_1)} + \dfrac{q_1(m_2)c(m_2)}{\lambda_{\max}(\mathbf{A}_2)}}{\dfrac{1}{\lambda_{\max}(\mathbf{A}_1)} + \dfrac{1}{\lambda_{\max}(\mathbf{A}_2)}} \leq$$
$$\leq 8q_1(\max(m_1,\ m_2))\max(c(m_1),\ c(m_2)),\ \text{где}\ c(m) \to 0\ \text{при}\ m \to \infty.$$

Отсюда получаем, что

$$\frac{\lambda_{\max}(\mathbf{\Omega}_1+\mathbf{\Omega}_2)}{\lambda_{\min}(\mathbf{\Omega}_1+\mathbf{\Omega}_2)} = o(q_1(m)) \text{ при } m_1,\ m_2 \to \infty.$$

Рассмотрим теперь член, содержащий смещение и покажем, что

$$\|\mathbf{b}_2-\mathbf{b}_1\|\|(\mathbf{\Omega}_1+\mathbf{\Omega}_2)^{-1}\|q_1(m) \to 0 \text{ при } m \to \infty.$$

$$\|\mathbf{b}_2-\mathbf{b}_1\|^2 \le \|\mathbf{b}_1\|^2+\|\mathbf{b}_2\|^2 \le \|\mathbf{v}_1-\mathbf{w}\|^2\|\mathbf{\Sigma}_1\|^2\|(\mathbf{\Sigma}_1+\mathbf{A}_1)^{-1}\|^2+\|\mathbf{v}_2-\mathbf{w}\|^2\|\mathbf{\Sigma}_2\|^2\|(\mathbf{\Sigma}_2+\mathbf{A}$$
$$\le C\left[\frac{1}{\lambda^2_{\min}(\mathbf{A}_1)}+\frac{1}{\lambda^2_{\min}(\mathbf{A}_2)}\right], \text{где } C = O(1) = \max(\|\mathbf{v}_2-\mathbf{w}\|^2\|\mathbf{\Sigma}_2\|^2, \|\mathbf{v}_1-\mathbf{w}\|^2\|\mathbf{\Sigma}_1\|^2).$$

Выберем m_1 и m_2 как и ранее так, что

$$\lambda^2_{\min}(\mathbf{A}_k) \ge 4\|\mathbf{\Sigma}_k\|\lambda_{\max}(\mathbf{A}_k),\ k=1,\ 2.$$

$$\frac{1}{\lambda^2_{\min}(\mathbf{A}_k)} = \frac{1}{\lambda_{\min}(\mathbf{A}_k)}\frac{\lambda_{\max}(\mathbf{A}_k)}{\lambda_{\min}(\mathbf{A}_k)}\frac{1}{\lambda_{\max}(\mathbf{A}_k)} \le \frac{d(m_k)}{q_1(m_k)}\frac{1}{\lambda_{\max}(\mathbf{A}_k)}, \text{ где } d(m)\to 0 \text{ при } m \to$$

$$\|\mathbf{b}_2-\mathbf{b}_1\|^2\|(\mathbf{\Omega}_1+\mathbf{\Omega}_2)^{-1}\| \le C\left[\frac{1}{\lambda^2_{\min}(\mathbf{A}_1)}+\frac{1}{\lambda^2_{\min}(\mathbf{A}_2)}\right]\frac{1}{\lambda_{\min}(\mathbf{\Omega}_1+\mathbf{\Omega}_2)} \le$$
$$\le C\left[\frac{1}{\lambda^2_{\min}(\mathbf{A}_1)}+\frac{1}{\lambda^2_{\min}(\mathbf{A}_2)}\right]\frac{1}{\frac{1}{4\lambda_{\max}(\mathbf{A}_1)}+\frac{1}{4\lambda_{\max}(\mathbf{A}_2)}} \le$$
$$\le \frac{4C\max(d(m_1),\ d(m_2))}{q_1(m)}\frac{\frac{q_1(m)}{q_1(m_1)}\frac{1}{\lambda_{\max}(\mathbf{A}_1)}+\frac{q_1(m)}{q_1(m_2)}\frac{1}{\lambda_{\max}(\mathbf{A}_2)}}{\frac{1}{\lambda_{\max}(\mathbf{A}_1)}+\frac{1}{\lambda_{\max}(\mathbf{A}_2)}} = o\left(\frac{1}{q_1(m)}\right) \text{ при } m_1,$$

что получено с учетом

$$\frac{q_1(m)}{q_1(m_1)} = \Theta(1),\ \frac{q_1(m)}{q_1(m_2)} = \Theta(1).$$

Осталось проверить третье условие, то есть

$$\|\Delta\mathbf{\Omega}\|\|(\mathbf{\Omega}_1+\mathbf{\Omega}_2)^{-1}\|.$$

Имеем

$$\|\Delta\mathbf{\Omega}\|\|(\mathbf{\Omega}_1+\mathbf{\Omega}_2)^{-1} = \|(\mathbf{\Sigma}_1+\mathbf{A}_1)^{-1}\mathbf{\Sigma}_1(\mathbf{\Sigma}_1+\mathbf{A}_1)^{-1}+(\mathbf{\Sigma}_2+\mathbf{A}_2)^{-1}\mathbf{\Sigma}_2(\mathbf{\Sigma}_2+\mathbf{A}_2)^{-1}\|\|(\mathbf{\Omega}_1+$$
$$\le \left(\|\mathbf{\Sigma}_1\|\|(\mathbf{\Sigma}_1+\mathbf{A}_1)^{-1}\|^2+\|\mathbf{\Sigma}_2\|\|(\mathbf{\Sigma}_2+\mathbf{A}_2)^{-1}\|^2\right)\|(\mathbf{\Omega}_1+\mathbf{\Omega}_2)^{-1}\| \le$$
$$\le \max(\|\mathbf{\Sigma}_1\|,\|\mathbf{\Sigma}_2\|)\left[\frac{1}{\lambda^2_{\min}(\mathbf{A}_1)}+\frac{1}{\lambda^2_{\min}(\mathbf{A}_2)}\right]\|(\mathbf{\Omega}_1+\mathbf{\Omega}_2)^{-1}\| = o\left(\frac{1}{q_1(m)}\right) \text{ при } m_1, m_2$$

так как последнее выражение совпадает с точностью до константы с выражением, оцененным в (4.9). Таким образом, получаем требуемое, то есть

$$-2\log s \xrightarrow{d} \chi^2(n).$$

□

Заметим, что в предыдущем утверждении считалось, что дисперсии шума известны, то есть σ_1^2 и σ_2^2 известны. Рассмотрим теперь случай, когда σ_1^2 и σ_2^2 неизвестны. В этом случае вместо σ_1^2 и σ_2^2, как и в случае с подходом классической статистики, используем их несмещенные оценки

$$\hat{\sigma}_k^2 = \frac{(\mathbf{y}_k - \mathbf{X}_k\hat{\mathbf{w}}_k^{ML})^\intercal(\mathbf{y}_k - \mathbf{X}_k\hat{\mathbf{w}}_k^{ML})}{m_k - n},\ k = 1,\ 2,$$

где $\hat{\mathbf{w}}_k^{ML}$ есть оценка максимума правдоподобия (4.1) для вектора параметров модели k. Введем обозначения

$$\tilde{\mathbf{A}_k} = \frac{1}{\hat{\sigma}_k^2}\mathbf{X}_k^\intercal\mathbf{X}_k,\ k = 1,\ 2,$$

$$\Delta\Omega_k = (\mathbf{\Sigma}_k + \tilde{\mathbf{A}}_k)^{-1} - (\mathbf{\Sigma}_k + \mathbf{A}_k)^{-1} + (\mathbf{\Sigma}_k + \mathbf{A}_k)^{-1}\mathbf{\Sigma}_k(\mathbf{\Sigma}_k + \mathbf{A}_k)^{-1},\ k = 1,\ 2.$$

Для s-score для пары апостериорных распределений имеем

$$-2\log s = (\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1)^\intercal\left(\left(\mathbf{\Sigma}_1 + \frac{1}{\hat{\sigma}_1^2}\mathbf{X}_1^\intercal\mathbf{X}_1\right)^{-1} + \left(\mathbf{\Sigma}_2 + \frac{1}{\hat{\sigma}_2^2}\mathbf{X}_2^\intercal\mathbf{X}_2\right)^{-1}\right)^{-1}(\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1), \text{где}$$

$$\hat{\mathbf{w}}_k = \left(\mathbf{\Sigma}_k + \frac{1}{\hat{\sigma}_k^2}\mathbf{X}_k^\intercal\mathbf{X}_k\right)^{-1}\left(\mathbf{\Sigma}_k\mathbf{v}_k + \frac{1}{\hat{\sigma}_k^2}\mathbf{X}_k^\intercal\mathbf{y}_k\right).$$

Отметим, что $\hat{\mathbf{w}}_k$ уже не имеет нормальное распределение, поскольку $\hat{\sigma}_k^2$ есть случайная величина. Введем обозначение

$$\tilde{\mathbf{w}}_k = \left(\mathbf{\Sigma}_k + \frac{1}{\sigma_k^2}\mathbf{X}_k^\intercal\mathbf{X}_k\right)^{-1}\left(\mathbf{\Sigma}_k\mathbf{v}_k + \frac{1}{\sigma_k^2}\mathbf{X}_k^\intercal\mathbf{y}_k\right),$$

что совпадает с выражением для оценки максимума апостериорной вероятности в случае, когда $\sigma_k^2,\ k = 1,\ 2$ известны. Обозначим также

$$-2\log\tilde{s} = (\tilde{\mathbf{w}}_2 - \tilde{\mathbf{w}}_1)^\intercal\left(\left(\mathbf{\Sigma}_1 + \frac{1}{\hat{\sigma}_1^2}\mathbf{X}_1^\intercal\mathbf{X}_1\right)^{-1} + \left(\mathbf{\Sigma}_2 + \frac{1}{\hat{\sigma}_2^2}\mathbf{X}_2^\intercal\mathbf{X}_2\right)^{-1}\right)^{-1}(\tilde{\mathbf{w}}_2 - \tilde{\mathbf{w}}_1).$$

Утверждение 8. Пусть имеются две модели линейной регрессии, описанные выше. Пусть также для некоторой возрастающей функции $q_1(m)$ такой, что $q_1(m) \to \infty$ и $q_1(m) = o(m^{1/4})$ при $m \to \infty$ выполнено

$$\frac{\lambda_{\max}(\mathbf{X}_k^\intercal \mathbf{X}_k)}{\lambda_{\min}(\mathbf{X}_k^\intercal \mathbf{X}_k)} = o(q_1(m_k)) \text{ при } m_k \to \infty,\ k = 1,\ 2,$$

$$\lambda_{\min}(\mathbf{X}_k^\intercal \mathbf{X}_k) = \Omega(q_1^2(m_k)) \text{ при } m_k \to \infty,\ k = 1,\ 2.$$

Тогда $-2 \log \tilde{s} \xrightarrow{d} \chi^2(n)$ при $m_1,\ m_2 \to \infty$, если $q_1(m_2)/q_1(m_1) = \Theta(1)$.

Доказательство. Покажем, что требуемое выполнено в силу выполнения условий утверждения 2. Введем обозначение $m = \max(m_1, m_2)$. Для этого требуется показать, что выполнены следующие условия при $m_1,\ m_2 \to \infty$

$$\begin{gathered} \|\mathbf{b}_2 - \mathbf{b}_1\|^2 \|(\mathbf{\Omega}_1 + \mathbf{\Omega}_2)^{-1}\| q_1(m) \to 0, \\ \frac{\lambda_{max}(\mathbf{\Omega}_1 + \mathbf{\Omega}_2)}{\lambda_{min}(\mathbf{\Omega}_1 + \mathbf{\Omega}_2)} = o(q_1(m)), \\ \|\Delta\mathbf{\Omega}\| \|(\mathbf{\Omega}_1 + \mathbf{\Omega}_2)^{-1}\| q_1(m) \to 0. \end{gathered}$$

Здесь $\Delta\mathbf{\Omega} = \Delta\mathbf{\Omega}_1 + \Delta\mathbf{\Omega}_2$, где

$$\Delta\mathbf{\Omega}_k = (\mathbf{\Sigma}_k + \tilde{\mathbf{A}}_k)^{-1} - (\mathbf{\Sigma}_k + \mathbf{A}_k)^{-1} + (\mathbf{\Sigma}_k + \mathbf{A}_k)^{-1}\mathbf{\Sigma}_k(\mathbf{\Sigma}_k + \mathbf{A}_k)^{-1}.$$

Первые два свойства выполнены, поскольку они полностью совпадают со свойствами, выполнение которых доказано в предыдущем утверждении. Покажем выполнение последнего свойства.

Отметим, как и ранее, что

$$\mathbb{P}(|\hat{\sigma}_k^2 - \sigma_k^2| \geq \varepsilon\sigma_k^2) \leq \frac{2}{\varepsilon^2(m_k - n)}$$

с учетом ${\hat{\sigma_k}}^2(m_k - n)/\sigma_k^2 \sim \chi^2(m_k - n)$. Отсюда, взяв $\varepsilon(m) = \omega(1/\sqrt{m}) = g(m)/\sqrt{m}$, где $g(m) \to \infty$ при $m \to \infty$, получим, что

$$\mathbb{P}\left(\frac{\hat{\sigma}_k^2}{\sigma_k^2} \in [1 - \varepsilon(m_k),\ 1 + \varepsilon(m_k)]\right) \to 1 \text{ при } m_k \to \infty,\ k = 1,\ 2. \tag{4.10}$$

Выберем m_1 и m_2 достаточно большими так, что $\varepsilon(m) \leq 1/4$ при $m \geq \min(m_1,\ m_2)$. Рассмотрим те исходы Ω', для которых

$$\frac{\hat{\sigma}_k^2}{\sigma_k^2} \in [1 - \varepsilon(m_k),\ 1 + \varepsilon(m_k)],\ k = 1,\ 2.$$

Из (4.10) имеем $\mathbb{P}(\Omega') \to 1$ при $m_1,\, m_2 \to \infty$. Обозначим $\delta_k(m_k) = \sigma_k^2/\hat{\sigma}_k^2 - 1$.

$$\|\Delta\mathbf{\Omega}_k\| \le \|(\mathbf{\Sigma}_k+\mathbf{A}_k)^{-1}\mathbf{\Sigma}_k(\mathbf{\Sigma}_k+\mathbf{A}_k)^{-1}\|+\|(\mathbf{\Sigma}_k+\mathbf{A}_k)^{-1}\|\|(\mathbf{I}+\delta_k\mathbf{A}_k(\mathbf{\Sigma}_k+\mathbf{A}_k)^{-1})^{-1}-\mathbf{I}\|$$
$$\overset{(4.8)}{\le} \frac{\|\mathbf{\Sigma}_k\|}{\lambda^2_{\min}(\mathbf{A}_k)}+\frac{1}{\lambda^2_{\min}(\mathbf{A}_k)}\left\|\sum_{l=1}^{\infty}(-1)^l\delta_k^l(\mathbf{A}_k(\mathbf{\Sigma}_k+\mathbf{A}_k)^{-1})^l\right\| \le \frac{\|\mathbf{\Sigma}_k\|}{\lambda^2_{\min}(\mathbf{A}_k)}+\frac{2\delta_k}{\lambda_{\min}(\mathbf{A}_k)}.$$

Выберем m_1 и m_2 как и ранее так, что

$$\lambda^2_{\min}(\mathbf{A}_k) \ge 4\|\mathbf{\Sigma}_k\|\lambda_{\max}(\mathbf{A}_k),\ k = 1,\, 2.$$

Тогда получим

$$\|\Delta\mathbf{\Omega}\|\|(\mathbf{\Omega}_1+\mathbf{\Omega}_2)^{-1}\| \le \left(\frac{\|\mathbf{\Sigma}_1\|}{\lambda^2_{\min}(\mathbf{A}_1)}+\frac{2\delta_1}{\lambda_{\min}(\mathbf{A}_1)}+\frac{\|\mathbf{\Sigma}_2\|}{\lambda^2_{\min}(\mathbf{A}_2)}+\frac{2\delta_2}{\lambda_{\min}(\mathbf{A}_2)}\right)\frac{4}{\frac{1}{\lambda_{\max}(\mathbf{A}_1)}+}$$
$$= o\left(\tfrac{1}{q_1(m)}\right)+8\frac{\frac{\delta_1}{\lambda_{\min}(\mathbf{A}_1)}+\frac{\delta_2}{\lambda_{\min}(\mathbf{A}_2)}}{\frac{1}{\lambda_{\max}(\mathbf{A}_1)}+\frac{1}{\lambda_{\max}(\mathbf{A}_2)}} \le o\left(\tfrac{1}{q_1(m)}\right)+8q_1(m)\max(c(m_1),c(m_2))\max(\delta_1($$

где $c(m) \to 0$ при $m \to \infty$. Выбирая

$$g(m) = \left(\frac{m^{1/4}}{q_1(m)}\right)^2,$$

получаем

$$\|\Delta\mathbf{\Omega}\|\|(\mathbf{\Omega}_1+\mathbf{\Omega}_2)^{-1}\| \le o\left(\tfrac{1}{q_1(m)}\right)+8\max(c(m_1),\, c(m_2))/q_1(m) = o\left(\tfrac{1}{q_1(m)}\right),$$

что и требовалось. □

Замечание 1. Отметим, однако, что вообще говоря в рассматриваемом случае неверно, что $-2\log s \overset{d}{\to} \chi^2(n)$. Это связано с тем, что $\|\hat{\mathbf{w}}_k - \tilde{\mathbf{w}}_k\|m_k^{1/2} = \Omega(1)$ по вероятности. При этом, если считать, что матрицы признаков $\mathbf{X}_k$, $k = 1,\, 2$ $\lambda_{\max}$ генерируются случайнои независимо по элементам из нормального или равномерного распределения по столбцам, то $\lambda_{\max}(\mathbf{A}_k) = \Omega(m_k)$, так как $\mathbb{E}\mathrm{tr}(\mathbf{A}_k) \propto m_k$, а $\mathrm{tr}(\mathbf{A}_k) \le n\lambda_{\max}$.

Обобщенно-линейные модели. Случай независимых моделей

В данной секции будет рассмотрена задача сравнения моделей для обобщенно-линейных моделей, частным случаем которых является, например, логистическа регрессия. Рассматривается при этом случай, когда множества

объектов, по которым построены модели не пересекаются. Это может иметь место, например, при обучении по огромным выборкам, когда предпочтительнее оказывается обучение по подвыборкам, а затем комбинация результатов, например, с помощью простого голосования.

Определение обобщенно-линейной модели.

Определение 1. *Экспоненциальным семейством распределений* называется семейство распределений $P_{\mathbf{y}|\boldsymbol{\theta}}$ для $\mathbf{y} \in \mathbb{R}^{\tilde{q}}$, где $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ и $\mathbf{u}(\mathbf{y}) \in \mathbb{R}^q$ с плотностями

$$f(\mathbf{y}|\boldsymbol{\theta}) = h(\mathbf{y}) \exp(\boldsymbol{\eta}^\intercal(\boldsymbol{\theta})\mathbf{u}(\mathbf{y}) - b(\boldsymbol{\theta})),$$

где $h(\mathbf{y}) \geq 0$. Здесь Θ есть выпуклое множество вида

$$\Theta = \{\boldsymbol{\theta} : \ 0 < \int h(\mathbf{y}) \exp(\boldsymbol{\eta}^\intercal(\boldsymbol{\theta})\mathbf{u}(\mathbf{y}))d\mathbf{y} < \infty\}.$$

Замечание 1. Далее предполагаем, что $b(\boldsymbol{\theta})$ бесконечное число раз дифференцируемо на $\Theta_0 = \operatorname{int}\Theta$.

Определение 2. Параметр некоторого семейства распределений $\boldsymbol{\theta}$ из экспоненциального семейства называется *натуральным*, если плотность распределения представима в виде

$$f(\mathbf{y}|\boldsymbol{\theta}) = h(\mathbf{y}) \exp(\boldsymbol{\theta}^\intercal\mathbf{u}(\mathbf{y}) - b(\boldsymbol{\theta})),$$

то есть $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$.

Определение 3. *Обобщенно-линейной моделью* называется модель порождения данных, то есть пар $(\mathbf{x}_i,\ y_i)$, определяемая следующими свойствами

- Случайные вектора $\{y_i\}_{i=1}^m$, $y_i \in \mathbb{R}^q$ независимы в совокупности и имеют плотности из экспоненциального семейства с натуральными параметрами $\boldsymbol{\theta}_i$
$$f(y_i|\boldsymbol{\theta}_i) = h(y_i) \exp(\boldsymbol{\theta}_i^\intercal\mathbf{u}(y_i) - b(\boldsymbol{\theta}_i)),\ i = 1,\ \ldots,\ m.$$
- $\mathbb{E}\mathbf{u}(y_i) = g^{-1}(\mathbf{w}^\intercal\mathbf{x}_i)$, где $g : \mathbb{R}^q \to \mathbb{R}^q$ есть функция связи.

Определение 4. *Натуральной функцией связи* называется функция связи вида $g = \mu^{-1}$, где $\mu = \partial b(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$.

Замечание 2. Отметим, что $\mathbb{E}\mathbf{u}(y_i) = \partial b(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, откуда для натуральной функции связи имеем

$$\boldsymbol{\theta}_i = \mathbf{w}^\intercal\mathbf{x}_i,\ \mathbb{E}\mathbf{u}(y_i) = \partial b(\boldsymbol{\theta}_i)/\partial\boldsymbol{\theta}_i|_{\boldsymbol{\theta}_i = \mathbf{w}^\intercal\mathbf{x}_i}.$$

Подход классической статистики.

Рассмотрим две модели логистической регрессии f_1 и f_2. Пусть $\mathbf{w}_1$ и $\mathbf{w}_2$ есть неизвестные параметры, соответствующие моделям f_1 и f_2, а $(\mathbf{X}_1,\ \mathbf{y}_1)$ и $(\mathbf{X}_2,\ \mathbf{y}_2)$ есть выборки, полученные из f_1 и f_2 соответственно. Считаем при этом, что объекты в первой и второй выборках разные. Тогда имеем следующие модели

$$y_{ki} \sim \mathrm{Be}(p_{ki}),\ \text{где}\ p_{ki} = g(\mathbf{w}_k^\intercal\mathbf{x}_{ki}) = \frac{1}{1 + \exp(-\mathbf{w}_k^\intercal\mathbf{x}_{ki})},\ k = 1,\ 2,\ i = 1,\ \ldots,\ m_k.$$

Функция правдоподобия, соответствующая таким моделям имет вид

$$p(\mathbf{y}_k|\mathbf{X}_k,\ \mathbf{w}_k) = \prod_{i=1}^{m_k} p_{ki}^{y_{ki}}(1-p_{ki})^{1-y_{ki}},\ k=1,\ 2.$$

Оценки максимума правдоподобия даются как

$$\hat{\mathbf{w}}_k = \arg\max_{\mathbf{w}_k} p(\mathbf{y}_k|\mathbf{X}_k,\ \mathbf{w}_k). \tag{4.11}$$

Отметим, что в отличие от линейной регрессии, где оценки максимума правдоподобия можно было получить в явном виде, получение $\hat{\mathbf{w}}_k$ из (4.11) в явном виде невозможно, поэтому далее покажем, что функция правдоподобия является вогнутой, а потому имеет единственный максимум.

Функция правдоподобия и ее свойства.

Для анализа функции правдоподобия введем обозначение

$$l_k(\mathbf{w}_k) = l(\mathbf{y}_k|\mathbf{X}_k, \mathbf{w}_k) = -\log p(\mathbf{y}_k|\mathbf{X}_k,\ \mathbf{y}_k) = -\sum_{i=1}^{m_k} (y_{ki}\log p_{ki} + (1-y_{ki})\log(1-p_{ki})) \tag{4.12}$$

Продифференцируем $l_k(\mathbf{w}_k)$ по $\mathbf{w}_k$ и с учетом $g'(x) = g(x)\cdot g(-x)$ получим

$$\frac{\partial l_k(\mathbf{w}_k)}{\partial \mathbf{w}_k} = -\sum_{i=1}^{m_k} \mathbf{x}_{ki}(y_{ki} - g(\mathbf{x}_{ki}^{\intercal}\mathbf{w}_k)) = \mathbf{X}_k^{\intercal}(\mathbf{g}_k - \mathbf{y}_k). \tag{4.13}$$

Найдем гессиан $l_k(\mathbf{w}_k)$ и покажем, что он положительно определен, откуда и получим выпуклость $l_k(\mathbf{w}_k)$, а, значит, вогнутость функции правдоподобия $p(\mathbf{w}_k|\mathbf{X}_k,\ \mathbf{y}_k)$ по $\mathbf{w}_k$ и единственность ее максимума.

$$\mathbf{H}_k = \frac{\partial^2 l_k(\mathbf{w}_k)}{\partial \mathbf{w}_k^2} = \sum_{i=1}^{m_k} \mathbf{x}_{ki} g(\mathbf{x}_{ki}^{\intercal}\mathbf{w}_k) g(-\mathbf{x}_{ki}^{\intercal}\mathbf{w}_k)\mathbf{x}_{ki}^{\intercal} = \mathbf{X}_k^{\intercal}\mathbf{R}_k\mathbf{X}_k, \tag{4.14}$$

где

$$\mathbf{R}_k = \operatorname{diag}\left(g(\mathbf{w}_k^{\intercal}\mathbf{x}_{ki})(1-g(\mathbf{w}_k^{\intercal}\mathbf{x}_{ki})),\ i=1,\ \ldots,\ m_k\right).$$

Рассмотрим произвольный вектор $\mathbf{u} \neq \mathbf{0}$.

$$\mathbf{u}^{\intercal}\mathbf{H}_k\mathbf{u} = \mathbf{u}^{\intercal}\mathbf{X}_k^{\intercal}\mathbf{R}_k\mathbf{X}_k\mathbf{u} = (\mathbf{X}_k\mathbf{u})^{\intercal}\,\mathbf{R}_k\,(\mathbf{X}_k\mathbf{u}) > 0,$$

откуда $\mathbf{H}_k$ положительно определенная при условии, что $\mathbf{X}_k\mathbf{u} \neq \mathbf{0}$, то есть столбцы матрицы признаков не линейно зависимы. Таким образом, имеем вогнутость функции правдоподобия, а, значит, единственность ее максимума – оценки максимума правдоподобия $\hat{\mathbf{w}}_k$.

Оценка максимума правдоподобия $\hat{\mathbf{w}}_k$ в модели логистической регрессии уже не имеет нормального распределения, однако имеет место ее асимптотическая нормальность, которая дается следующей теоремой [2].

Теорема 3 ([2]). Пусть имеется модель логистической регрессии, описанная выше. Пусть значения признаков ограничены, то есть $\exists C > 0 : |x_{kij}| \leq C \, \forall \, i = 1, \ldots, m_k, j = 1, \ldots, n$. Пусть также гессиан $\mathbf{H}_k$ является строго положительно определенным, причем в некоторой окрестности $N(\mathbf{w}_k)$ выполнено

$$\frac{\lambda_{\max}(\mathbf{H}_k)}{\lambda_{\min}(\mathbf{H}_k)} = O(1),$$

$$\lambda_{\min}(\mathbf{H}_k) \to \infty \text{ при } m_k \to \infty.$$

Тогда $\mathbf{H}_k^{1/2}(\hat{\mathbf{w}}_k - \mathbf{w}_k) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_n)$.

Замечание 3. Более слабые требования, при которых имеется асимптотическая нормальность, приведены в [3]. Так условие

$$\frac{\lambda_{\max}(\mathbf{H}_k)}{\lambda_{\min}(\mathbf{H}_k)} = O(1),$$

можно заменить на выполнение в некоторой окрестности $N(\mathbf{w}_k)$

$$\exists \, \delta > 0 : \; \frac{\lambda_{\min}(\mathbf{H}_k)}{\lambda_{\max}^{1/2+\delta}(\mathbf{H}_k)} = \Omega(1).$$

Подход байесовской статистики.

Линейная регрессия. Случай зависимых моделей

До сих пор считалось, что сравниваемые модели независимы, то есть объекты выборок, их представляющие являются независимыми, то есть множества объектов не пересекаются. Это, однако, не означает, что в разных выборках не может быть объектов с идентичными значениями признакового описания и целевой переменной. Однако требуется, чтобы все объекты, в том числе и такие, были независимыми, что в случае с кредитным скорингом подразумевает, что выборки людей для каждой из моделей не пересекаются.

Однако не всегда такое предположение оказывается выполненным. Так, например, идея бэггинга состоит в сэмплировании выборок той же длины из исходной путем выбора с возвразением. Затем для каждой из полученных новой выборки строится и оценивается своя модель и осуществляется простое голосование моделей. Выборки объектов в разных моделях бэггинга пересекаются, а потому не являются независимыми.

В данном разделе будет рассмотрен случай зависимых выборок между моделями.

4.2. Обоснование вида функции сходства

Несмотря на прореживание мультимодели, она может не являться (s, α) – адекватной, то есть может содержать похожие модели. По этой причине поставим задачу сравнения моделей, решение которой можно использовать при построении адекватной мультимодели.

Задачу сравнения пары моделей можно сформулировать следующим образом.

- Даны две модели f_1 и f_2, векторы параметров моделей $\mathbf{w}_1$, $\mathbf{w}_2$.
- Имеем выборки $(\mathbf{X}_1,\ \mathbf{y}_1)$ и $(\mathbf{X}_2,\ \mathbf{y}_2)$,
  $y_{1,i} = f_1(\mathbf{x}_{1,i},\ \mathbf{w}_1), \quad y_{2,i} = f_2(\mathbf{x}_{2,i},\ \mathbf{w}_2).$
- Априорные распределения параметров моделей $\mathbf{w}_1 \sim p_1(\mathbf{w})$, $\mathbf{w}_2 \sim p_2(\mathbf{w})$.
- Апостериорные распределения $p(\mathbf{w}_1|\mathbf{X}_1,\ \mathbf{y}_1)$ и $p(\mathbf{w}_2|\mathbf{X}_2,\ \mathbf{y}_2)$, обозначаемые далее $g_1(\mathbf{w})$ и $g_2(\mathbf{w})$.

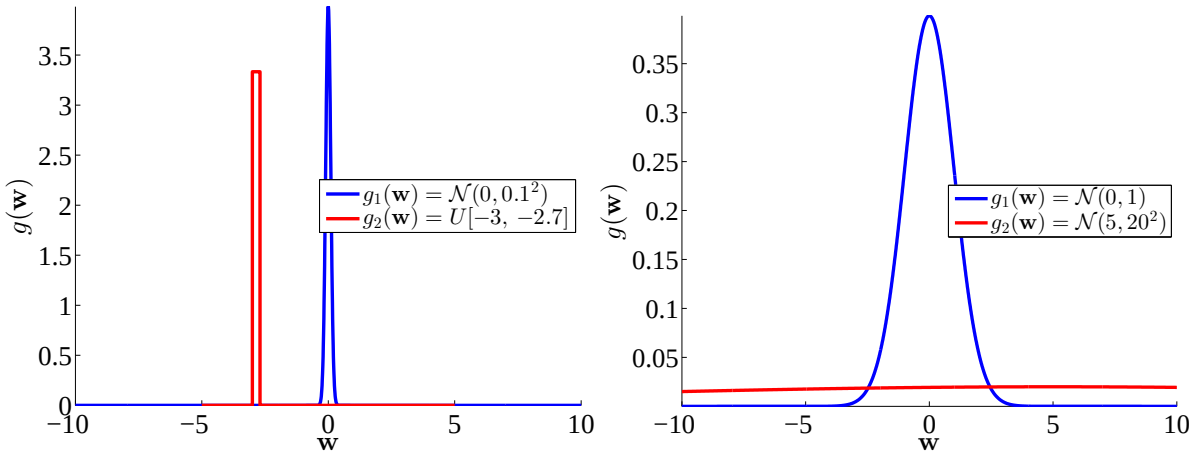
Требуется построить функцию сходства, определенную на паре распределений $g_1(\mathbf{w})$ и $g_2(\mathbf{w})$. Она должна удовлетворять следующим требованиям.

1. определена в случае несовпадения носителей,
2. $s(g_1,\ g_2) \leq s(g_1,\ g_1)$,
3. $s \in [0,\ 1]$,
4. $s(g_1,\ g_1) = 1$,
5. близка к 1, если $g_2(\mathbf{w})$ — малоинформативное распределение,
6. симметрична, $s(g_1,\ g_2) = s(g_2,\ g_1)$.

Отметим, что свойства 2–4 и свойство 6 в списке требований к функции сходства являются техническими. Так свойство 3 задает диапазон принимаемых значений функции сходства s, а свойства 2 и 4 указывают, что для пары совпадающих распределений значение функции съодства должно быть максимально возможным, причем сходство распределения с самим собой не ниже, чем с любым другим распределение. Свойство 1 определяет возможность сравнения моделей, определенных на разных признаковых пространствах. Свойство 5 является основным в решаемой задаче, поскольку обеспечивает неотличимость модели, про параметры которой ничего неизвестно от любой другой модели. Поясним это свойство и соответствующее ему отличие задачи сравнения моделей от задачи различения пары распределений.

На рис. 4.1а приведена пара распределений g_1, g_2, $g_1(w) = N(0,\ 0.1^2)$, $g_2(w) = U[-3,\ -2.7]$. Эти два распределения существенно отличаются, что можно выразить, например, через большое значение расстояния Дженсона-Шеннона между ними.

$$\rho_{JS}(g_1,\ g_2) \gg 0.$$



(а) Случай разных различимых распределений. (б) Случай разных неразличимых распределений.

Рис. 4.1: Иллюстрация различий между отличием апостериорных распределений параметров пары моделей и различимостью моделей.

Рассматривая это апостериорные распределения вероятностей на параметры моделей, заметим, что модели, им соответствующие также являются существенно разными, поскольку про параметр w первой модели известно, что его значение близко к нулю, а для второй модели $w \in [-3,\ -2.7]$.

На рис. 4.1б приведена пара распределений $g_1,\ g_2,\ g_1(w) = N(0,\ 1),\ g_2(w) = N(5,\ 20^2)$. Эти два распределения существенно отличаются, что можно выразить, например, через большое значение расстояния Дженсона-Шеннона между ними.

$$\rho_{JS}(g_1,\ g_2) \gg 0.$$

Рассматривая это апостериорные распределения вероятностей на параметры моделей, заметим, что из-за неинформативности второго распределения про значения параметра w для второй модели почти ничего неизвестно, а потому отличить вторую модель от первой нельзя.

Дадим формальное определение информативности распределения.

Определение информативности

Определение 22. Назовем распределение $g_2(\cdot) :\ \Omega \to \mathbb{R}^+$ *неинформативным* относительно распределения $g_1(\cdot) :\ \Omega \to \mathbb{R}^+$ с конечным носителем $\mathrm{supp}(g_1) = A$, если $\exists\, B :\ A \subseteq B$, что

$$\forall\, \mathbf{v} \in B :\ g_2(\mathbf{w}) = \frac{1}{|B|},$$

то есть $g_2(\cdot)$ есть равномерное распределение на множестве B.

Обобщим теперь понятие неинформативности на случай двух распределений $g_1(\cdot)$, $g_2(\cdot)$, которые определены на несовпадающих пространствах, то есть $g_1 : \Omega \times \Omega_1 \to \mathbb{R}^+$, $g_2 : \Omega \times \Omega_2 \to \mathbb{R}^+$.

Определение 23. Назовем распределение $g_2(\cdot) : \Omega \times \Omega_2 \to \mathbb{R}^+$ *неинформативным* относительно распределения $g_1(\cdot) : \Omega \times \Omega_1 \to \mathbb{R}^+$ с конечным носителем $\mathrm{supp}(g_1) = A$, если $\Omega_1 = \varnothing$, то есть g_1 определено на подпространстве области определения g_2 и $\exists\, \tau > 0,\ B : A \times [-\tau, \tau]^{\dim(\Omega_2)} \subseteq B$, что

$$\forall\, \mathbf{v} \in B : g_2(\mathbf{w}) = \frac{1}{|B|},$$

то есть $g_2(\cdot)$ есть равномерное распределение на множестве B.

Замечание 1. Отметим, что по приведенному определению неинформативность распределения g_2 относительно g_1 имеет место, когда g_1 не имеет дополнительных по сравнению с g_2 признаков, то есть $\Omega_1 = \varnothing$ и g_2 неинформативно относительно некоторого расширения $\overline{g}_1 : \Omega \times \Omega_2 \to \mathbb{R}^+$ распределения g_1, где $\overline{g}_1(\mathbf{w}, \mathbf{w}_2) = g_1(\mathbf{w}) g_\tau(\mathbf{w}_2)$, где $\mathrm{supp}(g_\tau) \subseteq [-\tau, \tau]^{\dim(\Omega_2)}$.

Определение 24. Назовем последовательность распределений $g_2^1(\cdot), \ldots, g_2^k(\cdot), \ldots\ \Omega \to \mathbb{R}^+$ *малоинформативной* на Ω, если выполнены следующие условия

$$\forall\, a > 0,\ g_2^k(\cdot)|_A \to U(A), \text{ где } A = \{\mathbf{w} : \|\mathbf{w}\| \le a\}, \tag{4.15}$$

$$\exists\, 0 \le B < \infty,\ \exists\, k_0 : \forall\, k \ge k_0 \sup_{\{\mathbf{w}:\, \|\mathbf{w}\| \ge B\}} g_2^k(\mathbf{w}) \le \sup_{\{\mathbf{w}:\, \|\mathbf{w}\| \le B\}} g_2^k(\mathbf{w}), \tag{4.16}$$

где в условии (4.15) $g_2^k(\cdot)|_A$ есть сужение распределения $g_2^k(\cdot)$ на множество A, то есть

$$g_2^k|_A(\mathbf{w}) = \begin{cases} 0, \text{ если } \|\mathbf{w}\| > a, \\ \dfrac{g_2^k(\mathbf{w})}{\int_A g_2^k(\mathbf{v}) d\mathbf{v}},\ \mathbf{w} \in A. \end{cases}$$

Сходимость в свойстве (4.15) понимается равномерная, то есть

$$g_2^k(\cdot)|_A \to U(A) \Longleftrightarrow \sup_{\mathbf{w} \in A} |g_2^k(\mathbf{w}) - 1/|A|| \to 0 \text{ при } k \to \infty.$$

Отметим, что для последовательности непрерывных плотностей распределения $g_2^k(\cdot)$ равномерную сходимость можно заменить на поточечную в силу эквивалентности этих понятий для непрерывных функций на ограниченном множестве.

Замечание 1. Отметим, что определение *неинформативности* дано относительно некоторого конкретного распределения с конечным носителем, в то время как определение *малоинформативности* последовательности распределений абсолютно, а не относительно. Отметим, однако, что *малоинформативность* последовательности распределений имеет вид лишь для соответствующего пространства Ω. Свойство исчезает при рассмотрении более широкого пространства вида $\tilde{\Omega} = \Omega \times \Omega_1$.

Замечание 2. Свойство (4.15) из определения *малоинформативности* означает сходимость к равномерному распределению рассматриваемой последовательности распределений для любого конечного множества и является характеристическим для вводимого понятия. Свойство же (4.16) является техническим, но также существенным, что будет в дальнейшем показано путем сравнения введенного понятия малоинформативности со стандартным, порожденным KL – дивергенцией.

Таким образом, пользуясь введенным определением неинформативности распределения относительно другого распределения и малоинформативной последовательности распределений, уточним требования к функции $s(g_1, g_2)$ сходства распределений.

Определение 25. Функция сходства $s(g_1, g_2)$, определенная на паре распределений $g_1(\mathbf{w}) : \Omega \times \Omega_1 \to \mathbb{R}^+$ и $g_2(\mathbf{w}) : \Omega \times \Omega_2 \to \mathbb{R}^+$, где $\Omega = \mathbb{R}^k$, $\Omega_1 = \mathbb{R}^{k_1}$, $\Omega_2 = \mathbb{R}^{k_2}$, $k,\ k_1,\ k_2 \geq 0$ называется *корректной*, если она удовлетворяет следующим требованиям.

1. $s(g_1,\ g_2)$ определена $\forall\, k,\ k_1,\ k_2 \geq 0$, если $g_1(\mathbf{w},\ \mathbf{w}_1),\ g_2(\mathbf{w},\ \mathbf{w}_2) < \infty\ \forall\, \mathbf{w} \in \Omega,\ \mathbf{w}_1 \in \Omega_1,\ \mathbf{w}_2 \in \Omega_2$,
2. $s(g_1,\ g_2) \leq s(g_1,\ g_1)$,
3. $s(g_1,\ g_2) \in [0,\ 1]$,
4. $s(g_1,\ g_1) = 1$,
5. - Если g_2 является неинформативным относительно g_1, то $s(g_1, g_2) = 1$;
   - $\forall\, g_1 :\ k_1 = 0$, то есть $g_1(\mathbf{w},\ \mathbf{w}_1) = g_1(\mathbf{w})$, для малоинформативной последовательности распределений $g_2^1,\ \ldots,\ g_2^k,\ \ldots$ выполнено
   $$s(g_1,\ g_2^k) \to 1 \text{ при } k \to \infty,$$
6. $s(g_1,\ g_2) = s(g_2,\ g_1)$.

Определение 26. Назовем *тривиальной функцией сходства* функцию сходства, которая не различает никакую пару распределений, то есть $s_{tr}(g_1, g_2) \equiv 1$.

Замечание 1. Непосредственной проверкой убеждаемся, что тривиальная функция сходства является корректной.

Приведем теперь пример малоинформативных последовательностей распределений $g_2^1,\ \ldots,\ g_2^k,\ \ldots,\ g_2^k :\ \mathbb{R}^k \to \mathbb{R}^+$.

- $g_2^k = U(A_k)$, где $A_k = \{\mathbf{w} :\ \|\mathbf{w}\| \leq k\}$
  Данная последовательность является малоинформативной, что непосредственно следует из определения малоинформативной последовательности при $B = 0$.
- $g_2^k = N(\mathbf{m}_k,\ \mathbf{\Sigma}_k)$, где $\sup_k \|\mathbf{m}_k\| < \infty,\ \|\mathbf{\Sigma}_k^{-1}\| \to 0$ при $k \to \infty$
  Данная последовательность также является малоинформативной. Покажем сначала выполнение свойства (4.16). Для доказательства рассмотрим $B = \sup_k \|\mathbf{m}_k\|$, тогда
  $$\frac{\sup_{\{\mathbf{w}:\ \|\mathbf{w}\| \geq B\}} g_2^k(\mathbf{w})}{\sup_{\{\mathbf{w}:\ \|\mathbf{w}\| \leq B\}} g_2^k(\mathbf{w})} = \frac{\sup_{\{\mathbf{w}:\ \|\mathbf{w}\| \geq B\}} g_2^k(\mathbf{w})}{g_2^k(\mathbf{m}_k)} \leq 1.$$

Свойство (4.15) следует из

$$\frac{\max_A g_2^k(\mathbf{w})}{\min_A g_2^k(\mathbf{w})} \leq \max_A \exp(1/2(\mathbf{w}-\mathbf{m}_k)^\intercal \|\mathbf{\Sigma}_k^{-1}\|(\mathbf{w}-\mathbf{m}_k)) \leq \exp(\|\mathbf{\Sigma}_k^{-1}\|(a^2+\|\mathbf{m}_k\|^2))$$

при $k \to \infty$, где $A = \{\mathbf{w} : \|\mathbf{w}\| \leq a\}$.

Покажем теперь, что известные расстояния между распределениями не удовлетворяют перечисленными требованиям к функции сходства, в частности требованию 5, которое определяет специфику решаемой задачи, а потому не являются корректными. При этом для перехода от расстояния / дивергенции к функции близости будет использовать следующее преобразование

$$s_\rho(g_1,\ g_2) = \exp(-\rho(g_1,\ g_2)), \tag{4.17}$$

которое позволяет автоматически выполнить требования 2, 3, 4 при условии неотрицательности дивергенции / расстояния и равенства нулю для пары совпадающих распределений.

Теорема 4 (Адуенко, 2014)**.** Перечисленным требованиям к функции сходства для пары распределений не удовлетворяют а) дивергенция Кульбака-Лейблера; б) расстояние Дженсона-Шеннона; в) расстояние Хеллингера; г) расстояние Бхаттачарайа.

Доказательство. а)

$$D_{KL}(g_1,\ g_2) = \int g_1(\mathbf{w}) \log \frac{g_1(\mathbf{w})}{g_2(\mathbf{w})} d\mathbf{w}$$

есть дивергенция Кульбака-Лейблера. Так как дивергенция Кульбака-Лейблера принимает неотрицательные значения и $D_{KL}(g, g) = 0$, для $s_{KL}(\cdot, \cdot)$ выполнены требования 2, 3, 4. Покажем, что при этом не выполнены требования 1, 5, 6.

Требование 6 не выполнено в силу несимметричности дивергенции Кульбака-Лейблера. $D_{KL}(g_1, g_2) \neq D_{KL}(g_2, g_1)$, откуда $s_{KL}(g_1, g_2) \neq s_{KL}(g_2, g_1)$. Требование 1 не выполененно, так как $D_{KL}(g_1,\ g_2)$ не определено, если $g_1(x) \neq 0$, $g_2(x) = 0$ на множестве ненулевой меры относительно g_1. Требование 5 не выполнено, так как

$$D_{KL}(\mathcal{N}(0,\ 1),\ \mathcal{N}(0,\ \sigma^2) = \frac{1}{2}\left(-\frac{1}{\sigma^2} + 1 + \log\sigma^2\right) \to \infty \text{ при } \sigma^2 \to \infty,$$

откуда $s_{KL}(\mathcal{N}(0,\ 1),\ \mathcal{N}(0,\ \sigma^2)) \to 0$ при $\sigma^2 \to \infty$, хотя последовательность распределений $N(0,\ \sigma^2)$ является малоинформативной.

б)

$$D_{JS}(g_1,\ g_2) = \sqrt{\tfrac{1}{2}D_{KL}(g_1,\ \tfrac{1}{2}(g_1+g_2)) + \tfrac{1}{2}D_{KL}(g_2,\ \tfrac{1}{2}(g_1+g_2))}$$

есть расстояние Дженсона-Шеннона. Так как расстояние Дженсона-Шеннона является расстоянием, для $s_{JS}(\cdot,\ \cdot)$ выполнены требования 2, 3, 4. По этой же причине выполнено свойство 6. Свойство 1 также выполнено, так как носитель g_1 содержится в носителе $\frac{1}{2}(g_1+g_2)$, а также носитель g_2 содержится в носителе $\frac{1}{2}(g_1+g_2)$, то есть $\mathrm{supp}(g_1),\ \mathrm{supp}(g_2) \subseteq \mathrm{supp}\left(\frac{1}{2}(g_1+g_2)\right)$, а потому дивергенция Кульбака-Лейблера определена.

Покажем, что свойство 5 не выполнено. Для этого рассмотрим пару распределений $g_1(w) = N(0,\ 1),\ g_2(w) = N(0,\ \sigma^2)$. Рассмотрим $\sigma^2 \geq 4e$. Тогда при $x \in [-1,\ 1]$

$$g_1(w) = \frac{1}{\sqrt{2\pi}}\exp(-w^2/2) \geq \frac{1}{\sqrt{2\pi e}} \geq \frac{2}{\sqrt{2\pi}\sigma} \geq 2g_2(w).$$

Отсюда имеем

$$D^2_{JS}(g_1,\ g_2) \geq \int g_1(w)\log\frac{2g_1(w)}{g_1(w)+g_2(w)}dw \geq \int_{-1}^{1} g_1(w)\log\frac{2g_1(w)}{g_1(w)+g_2(w)}dw \geq \int_{-1}^{1} g_1(w)\log(4/3)dw = \log(4/3)(2\Phi(1)-1). \quad (4.18)$$

Из (4.18) получаем

$$D_{JS}(\mathcal{N}(0,\ 1),\ \mathcal{N}(0,\ \sigma^2)) \geq \sqrt{\log(4/3)(2\Phi(1)-1)} > 0 \text{ при } \sigma^2 \geq 4e.$$

Отсюда

$$s_{JS}(\mathcal{N}(0,\ 1),\ \mathcal{N}(0,\ \sigma^2)) \leq \exp(-1-\sqrt{\log(4/3)(2\Phi(1)-1)}) \not\to 1 \text{ при } \sigma^2 \to \infty.$$

в)

$$D_H(g_1,\ g_2) = \sqrt{1-\int\sqrt{g_1(\mathbf{w})g_2(\mathbf{w})}d\mathbf{w}}$$

есть расстояние Хеллингера. Так как расстояние Хеллингера является расстоянием, для $s_H(\cdot,\ \cdot)$ выполнены требования 2, 3, 4. По этой же причине выполнено свойство 6. Свойство 1 формально выполнено, так как для пары распределений с несовпадающими носителями $D_H(g_1,\ g_2) \equiv 1$, что ведет к невыполненности свойства 5 и тривиальности результата сравнения распределений с несовпадающими носителями. Покажем, что и для распределений с совпадающими носителями свойство 5 не выполнено. Для этого рассмотрим пару распределений $g_1(w) = N(0,\ 1),\ g_2(w) = N(0,\ \sigma^2)$.

$$\int\sqrt{g_1(w)g_2(w)}dw = \frac{1}{\sqrt{2\pi\sigma}}\int e^{-\frac{w^2}{4}\left(1+\frac{1}{\sigma^2}\right)}dw = \frac{\sqrt{2}}{\sqrt{\sigma+1/\sigma}} \to 0 \text{ при } \sigma^2 \to \infty.$$

Отсюда

$$s_H(\mathcal{N}(0,\ 1),\ \mathcal{N}(0,\ \sigma^2)) \not\to 1 \text{ при } \sigma^2 \to \infty.$$

г)

$$D_B(g_1,\, g_2) = -\log \int \sqrt{g_1(w)g_2(w)}dw$$

есть расстояние Бхаттачарайа. Так как расстояние Бхаттачарайа является расстоянием, для $s_B(\cdot,\, \cdot)$ выполнены требования 2, 3, 4. По этой же причине выполнено свойство 6. Свойство 1 не выполнено, так как для пары распределений с несовпадающими носителями $\int \sqrt{g_1(w)g_2(w)}dw = 0$. Покажем, что свойство 5 не выполнено. Как и для расстояния Хеллингера рассмотрим пару распределений $g_1(w) = N(0,\, 1)$, $g_2(w) = N(0,\, \sigma^2)$. Имеем

$$\int \sqrt{g_1(w)g_2(w)}dw = \frac{1}{\sqrt{2\pi\sigma}} \int e^{-\frac{w^2}{4}\left(1+\frac{1}{\sigma^2}\right)}dw = \frac{\sqrt{2}}{\sqrt{\sigma + 1/\sigma}} \to 0 \text{ при } \sigma^2 \to \infty.$$

Отсюда

$$s_B(\mathcal{N}(0,\, 1),\, \mathcal{N}(0,\, \sigma^2)) \to 0 \neq 1 \text{ при } \sigma^2 \to \infty.$$

□

Отметим, что приведенный выше результат не является частным и выполнен для гораздо более широкого класса дивергенций – дивергенций Брегмана и f-дивергенций. Опишем далее этот класс и покажем, что дивергенции Брегмана и f-дивергенции не удовлетворяют свойству 5 требований к функции сходства.

Определение 27. *Дивергенцией Брегмана* $D_F(p,\, q) : \Omega \times \Omega \to \mathbb{R}$, где Ω – выпуклое множество, называется заданная строго выпуклой непрерывно дифференцируемой функцией $F :\, \Omega \to \mathbb{R}$ функция двух аргументов вида

$$D_F(p,\, q) = F(p) - F(q) - \langle \nabla F(q),\, p - q \rangle.$$

Замечание 1. Частными случаями дивергенции Брегмана являются, например,

- Квадрат L_2 расстояния между распределениями: $\int (p(x) - q(x))^2 dx = D_F(p,\, q)$, где $F(p) = \int p^2(x)dx$;
- Дивергенция Кульбака-Лейблера: $\int p(x) \log(p(x)/q(x))dx = D_F(p,\, q)$, где $F(p) = \int p(x) \log p(x)dx - \int p(x)dx$, где учтено условие нормировки: $\int p(x)dx = \int q(x)dx = 1$.

В примерах выше скалярным произведением является стандартное скалярное произведение в пространстве $L_2(\mathbb{R})$ $\langle p(x),\, q(x) \rangle = \int p(x)q(x)dx$.

Определение 28. *Симметризованной дивергенцией Брегмана* $\tilde{D}_F(p,\, q) : \Omega \times \Omega \to \mathbb{R}$ назовем

$$\tilde{D}_F(p,\, q) = \sqrt{\tfrac{1}{2} D_F\left(p, \tfrac{1}{2}(p+q)\right) + \tfrac{1}{2} D_F\left(q, \tfrac{1}{2}(p+q)\right)}.$$

Теорема 5 (Адуенко, 2016)**.** Функции сходства, порожденные дивергенциями Брегмана и симметризованными дивергенциями Брегмана в соответствии с (4.17), не являются корректными.

Доказательство. Докажем сначала, что данное утверждение выполнено для дивергенций Брегмана. Пусть F – некотора строго выпуклая функция, порождающая дивергенцию Брегмана. Рассмотрим случай пары несовпадающих распределений $(g_1,\ g_2)$, $g_1 \neq g_2$, что g_2 является неинформативным относительно g_1, то есть g_1 имеет конечный носитель, а

$$g_2 = U(A),\ \mathrm{supp}(g_1) \subseteq A.$$

Требование 5 к функции сходства s_F, порожденной дивергенцией Брегмана D_F, для этой пары распределений принимает вид

$$s_F(g_1,\ g_2) = 1 \Longleftrightarrow D_F(g_1,\ g_2) = 0.$$

Тогда получаем

$$F(g_1) - F(g_2) - \langle \nabla F(g_2),\ g_1 - g_2 \rangle = 0. \tag{4.19}$$

Так как $g_1 \neq g_2$ условие (4.19) противоречит строгой выпуклости функции F, откуда имеем требуемое.

Покажем теперь, что функции сходства, порождаемые симметризованными дивергенциями Брегмана, также не являются корректными. Как и в предыдуем случае, рассмотрим случай пары несовпадающих распределений $(g_1,\ g_2)$, $g_1 \neq g_2$, что g_2 является неинформативным относительно g_1, то есть g_1 имеет конечный носитель, а

$$g_2 = U(A),\ \mathrm{supp}(g_1) \subseteq A.$$

Требование 5 к функции сходства $\tilde{s}_F$, порожденной симметризованной дивергенцией Брегмана $\tilde{D}_F$, для этой пары распределений принимает вид

$$\tilde{s}_F(g_1,\ g_2) = 1 \Longleftrightarrow \tilde{D}_F(g_1,\ g_2) = 0.$$

Это условие эквивалетно следующему

$$D_F\left(g_1, \tfrac{1}{2}(g_1+g_2)\right) + D_F\left(g_2, \tfrac{1}{2}(g_1+g_2)\right) = 0.$$

Запишем левую часть равенства в явном виде, получим эквивалетное условие корректности функции сходства

$$D_F\left(g_1, \tfrac{1}{2}(g_1+g_2)\right)+D_F\left(g_2, \tfrac{1}{2}(g_1+g_2)\right) = F(g_1)-F\left(\tfrac{1}{2}(g_1+g_2)\right)-\langle \nabla F\left(\tfrac{1}{2}(g_1+g_2)\right),$$
$$+F(g_2)-F\left(\tfrac{1}{2}(g_1+g_2)\right)+\langle \nabla F\left(\tfrac{1}{2}(g_1+g_2)\right), \tfrac{1}{2}(g_1-g_2)\rangle = F(g_1)+F(g_2)-2F\left(\tfrac{1}{2}(g_1+g_2)\right.$$

Так как $g_1 \neq g_2$ условие (4.20) противоречит строгой выпуклости функции F, откуда имеем требуемое. □

Рассмотрим теперь понятие f - дивергенций и покажем, что функции сходства, порожденные этим классом дивергенций, не являются корректными.

Определение 29. *f-дивергенцией* $d_f(p,\ q) : \Omega \times \Omega \to \mathbb{R}$, где Ω – множество распределений с конечными плотностями над $\mathbb{R}^k$, называется заданная выпуклой функцией $f :\ \mathbb{R} \to \mathbb{R},\ f(1) = 0$ функция двух аргументов вида

$$d_f(p,\ q) = \int_{\mathbb{R}^k} f\left(\frac{p(\mathbf{w})}{q(\mathbf{w})}\right) q(\mathbf{w}) d\mathbf{w}.$$

Замечание 1. Частными случаями f-дивергенций являются, например,

- Дивергенция Кульбака-Лейблера для $f(t) = t \log t$;
- Квадрат расстояния Хеллингера для $f(t) = 1 - \sqrt{t},\ f(t) = (\sqrt{t} - 1)^2$;
- χ^2 – дивергенция для $f(t) = (t-1)^2$

$$d_{\chi^2}(p,\ q) = \int_{\mathbb{R}^k} \left(\frac{p(\mathbf{w})}{q(\mathbf{w})} - 1\right)^2 q(\mathbf{w}) d\mathbf{w};$$

- Расстояние полной вариации для $f(t) = \frac{1}{2}|t-1|$

$$d_{TV}(p,\ q) = \tfrac{1}{2} \int_{\mathbb{R}^k} |p(\mathbf{w}) - q(\mathbf{w})| d\mathbf{w}.$$

Замечание 2. Данное определение f – дивергенций корректно, если $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$ в силу наличия особенности в аргументе функции f при $q(\mathbf{w}) \to 0$. Отметим, однако, что для некоторых функций f возможно корректное определение f – дивергенции и для случая, когда $\mathrm{supp}(p) \not\subseteq \mathrm{supp}(q)$. Примером такой функции f является $f(t) = 1 - \sqrt{t}$, порождающая квадрат расстояния Хеллингера.

Отметим теперь следующее важное свойство f – дивергенций для случая их корректного определения для произвольной функции f, то есть в случае $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$.

Утверждение 9.

$$\forall\, f,\ p,\ q :\ \mathrm{supp}(p) \subseteq \mathrm{supp}(q) \text{ выполнено } d_f(p,\ q) \geq 0.$$

При этом для любой выпуклой функции f, отличной от линейной, равенство $d_f(p,\ q) = 0$ имеет место тогда и только тогда, когда $p(\mathbf{w}) \equiv q(\mathbf{w})$.

Доказательство. Для доказательства этого утверждения воспользуемся неравенством Йенсена. Пусть $\boldsymbol{\xi}$ – случайный вектор, имеющий плотность распределения $q(\mathbf{w})$. Введем случайную величину $\eta = p(\boldsymbol{\xi})/q(\boldsymbol{\xi})$. Тогда из неравенства Йенсена имеем

$$\mathbb{E}_q f(\eta) \geq f(\mathbb{E}_q \eta).$$

Отсюда получаем

$$\int q(\mathbf{w}) f\left(\frac{p(\mathbf{w})}{q(\mathbf{w})}\right) d\mathbf{w} \geq f\left(\int q(\mathbf{w}) \frac{p(\mathbf{w})}{q(\mathbf{w})} d\mathbf{w}\right) = f(1) = 0.$$

При этом равенство достигается тогда и только тогда, когда f – линейна или η – вырожденная случайная величина. Вырожденность случайной величины η означает, что

$$\exists\, C:\ \forall\, \mathbf{w} \in \mathrm{supp}(q):\ p(\mathbf{w}) = Cq(\mathbf{w}).$$

В силу условия нормировки и $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$ имеем

$$1 = \int_{\mathbf{w}\in\mathrm{supp}(q)} p(\mathbf{w})d\mathbf{w} = C\int_{\mathbf{w}\in\mathrm{supp}(q)} q(\mathbf{w})d\mathbf{w} = C.$$

□

Замечание 1. Отметим, что условие $\mathrm{supp}(p) \subseteq \mathrm{supp}(q)$ в частности выполнено, если носитель $q(\mathbf{w})$ совпадает со всем пространством, то есть $\mathrm{supp}(q) = \mathbb{R}^k$.
Замечание 2. Отметим, что при $\mathrm{supp}(p) \not\subseteq \mathrm{supp}(q)$ условие неотрицательности f – дивергенции может быть не выполнено.

В качестве примера рассмотрим $f(t) = (1 - \sqrt[4]{t})(0.5 - \sqrt[4]{t})$. Пусть $\mathrm{supp}(q) \neq \mathbb{R}^k$, а $p(\mathbf{w}) = 0.5q(\mathbf{w})\ \forall\ \mathbf{w} \in \mathrm{supp}(q(\mathbf{w}))$, а для $\mathbf{w} \notin \mathrm{supp}(q(\mathbf{w}))$ доопределим $p(\mathbf{w})$ произвольным образом с учетом условия нормировки. Тогда

$$d_f(p,\, q) = \int_{\mathbb{R}^k} \left(0.5q(\mathbf{w}) - 1.5\sqrt[4]{p(\mathbf{w})q^3(\mathbf{w})} + \sqrt{p(\mathbf{w})q(\mathbf{w})}\right) d\mathbf{w} =$$
$$0.5 - 1.5\sqrt[4]{0.5} + \sqrt{0.5} = (1 - \sqrt{0.5})(0.5 - \sqrt{0.5}) < 0.$$

Покажем теперь, что функции сходства, порожденные f – дивергенциями в соответствии с (4.17), не являются корректными.
Теорема 6 (Адуенко, 2016)**.** Функции сходства, порожденные f – дивергенциями в соответствии с (4.17), не являются корректными.

Доказательство. Рассмотрим случай пары несовпадающих распределений $(g_1,\, g_2)$, $g_1 \neq g_2$, что g_2 является неинформативным относительно g_1, то есть g_1 имеет конечный носитель, а

$$g_2 = U(A),\ \mathrm{supp}(g_1) \subseteq A.$$

Тогда $\mathrm{supp}(g_1) \subseteq \mathrm{supp}(g_2)$ и f – дивергенция корректна определена для этой пары распределений для любой выпуклой функции f, для которой $f(1) = 0$. Требование 5 к функции сходства s_f, порожденной f – дивергенцией d_f, для этой пары распределений принимает вид

$$s_f(g_1,\, g_2) = 1 \Longleftrightarrow d_f(g_1,\, g_2) = 0.$$

Однако в силу утверждения (9)

$$d_f(g_1,\, g_2) = 0 \Longleftrightarrow g_1(\mathbf{w}) \equiv g_2(\mathbf{w}),$$

что не выполнено в силу выбора пары распределений $g_1,\, g_2$. Полученное противоречие завершает доказательство. □

4.3. Предлагаемая функция сходства

Как было показано ранее, функции сходства, порождаемые дивергенциями Брегмана, симметризованными дивергенциями Брегмана и f – дивергенциями, не являются корректными и в частности не удовлетворяет свойству 5, являющемуся определяющем в решаемой задаче. Предложим далее функцию сходства s и покажем, что она является корректной.

Определение 30. Назовем функцией сходства *s-score* пары распределений $g_1 : \mathbb{R}^k \to \mathbb{R}^+$, $g_2 : \mathbb{R}^k \to \mathbb{R}^+$, определенных на одном пространстве, функцию вида

$$s_0(g_1,\ g_2) = \frac{\int g_1(\mathbf{w})g_2(\mathbf{w})d\mathbf{w}}{\max_{\mathbf{b}\in\mathbb{R}^k}\int g_1(\mathbf{v})g_2(\mathbf{v}-\mathbf{b})d\mathbf{v}}.$$

Определение 31. Назовем пару распределений $(g_1^\tau(\mathbf{w},\ \mathbf{w}_1,\ \mathbf{w}_2) : \mathbb{R}^k \times \mathbb{R}^{k_1} \times \mathbb{R}^{k_2} \to \mathbb{R}^+,\ g_2^\tau(\mathbf{w},\ \mathbf{w}_1,\ \mathbf{w}_2) : \mathbb{R}^k \times \mathbb{R}^{k_1} \times \mathbb{R}^{k_2} \to \mathbb{R}^+)$ τ – расширением пары распределений $(g_1(\mathbf{w},\ \mathbf{w}_1) : \mathbb{R}^k \times \mathbb{R}^{k_1} \to \mathbb{R}^+,\ g_2(\mathbf{w},\ \mathbf{w}_2) : \mathbb{R}^k \times \mathbb{R}^{k_2} \to \mathbb{R}^+)$, если

$$g_1^\tau(\mathbf{w},\ \mathbf{w}_1,\ \mathbf{w}_2) = g_1(\mathbf{w},\ \mathbf{w}_1)U(Q_2^\tau),\ g_2^\tau(\mathbf{w},\ \mathbf{w}_1,\ \mathbf{w}_2) = g_2(\mathbf{w},\ \mathbf{w}_2)U(Q_1^\tau), \text{ где}$$

$$Q_1^\tau = \{\mathbf{w}_1 :\ \|\mathbf{w}_1\|_\infty \le \tau\},\ Q_2^\tau = \{\mathbf{w}_2 :\ \|\mathbf{w}_2\|_\infty \le \tau\}.$$

Определение 32. Назовем функцией сходства *s-score* пары распределений $g_1(\mathbf{w},\ \mathbf{w}_1)$, $g_2(\mathbf{w},\ \mathbf{w}_2)$, $g_1 : \mathbb{R}^k \times \mathbb{R}^{k_1} \to \mathbb{R}^+$, $g_2 : \mathbb{R}^k \times \mathbb{R}^{k_2} \to \mathbb{R}^+$ функцию вида

$$s(g_1,\ g_2) = \lim_{\tau\to 0} s_0(g_1^\tau,\ g_2^\tau).$$

Утверждение 10. Функция сходства $s(g_1,\ g_2)$ определена корректно.

Доказательство. Покажем, что предел в определении функции сходства существует и конечен, а потому функция сходства $s(g_1,\ g_2)$ определена корректно. Для упрощения обозначений далее опустим верхний индекс τ и вместо Q_1^τ, Q_2^τ используем Q_1, Q_2 соответственно. Имеем по определению

$$s(g_1,\ g_2) = \lim_{\tau\to 0} \frac{\int_{Q_1} d\mathbf{w}_1 \int_{Q_2} d\mathbf{w}_2 \int d\mathbf{w} g_1(\mathbf{w},\ \mathbf{w}_1)g_2(\mathbf{w},\ \mathbf{w}_2)}{\max_{\mathbf{b},\,\mathbf{b}_1,\,\mathbf{b}_2} \int_{Q_1} d\mathbf{v}_1 \int_{Q_2} d\mathbf{v}_2 \int d\mathbf{v} g_1(\mathbf{v}-\mathbf{b},\ \mathbf{v}_1-\mathbf{b}_1)g_2(\mathbf{v},\ \mathbf{v}_2-\mathbf{b}_2)}. \tag{4.21}$$

Рассмотрим числитель и знаменатель в (4.21) раздельно.

$$\int_{Q_1} d\mathbf{w}_1 \int_{Q_2} d\mathbf{w}_2 \int d\mathbf{w} g_1(\mathbf{w},\ \mathbf{w}_1)g_2(\mathbf{w},\ \mathbf{w}_2) = |Q_1||Q_2| \int g_1(\mathbf{w},\ \hat{\mathbf{w}}_1)g_2(\mathbf{w},\ \hat{\mathbf{w}}_2)d\mathbf{w},$$

где $\hat{\mathbf{w}}_1 \in Q_1$, $\hat{\mathbf{w}}_2 \in Q_2$.

Для знаменателя аналогично имеем

$$\max_{\mathbf{b},\,\mathbf{b}_1,\,\mathbf{b}_2} \int_{Q_1} d\mathbf{v}_1 \int_{Q_2} d\mathbf{v}_2 \int g_1(\mathbf{v}-\mathbf{b},\ \mathbf{v}_1-\mathbf{b}_1)g_2(\mathbf{v},\ \mathbf{v}_2-\mathbf{b}_2)d\mathbf{v} =$$

$$|Q_1||Q_2| \max_{\mathbf{b},\,\mathbf{b}_1,\,\mathbf{b}_2} \int g_1(\mathbf{v}-\mathbf{b},\tilde{\mathbf{v}}_1-\mathbf{b}_1)g_2(\mathbf{v},\tilde{\mathbf{v}}_2-\mathbf{b}_2)d\mathbf{v} = |Q_1||Q_2| \max_{\mathbf{b},\,\mathbf{v}_1,\,\mathbf{v}_2} \int g_1(\mathbf{v}-\mathbf{b},\mathbf{v}_1)g_2(\mathbf{v}$$

где в силу $\mathbf{b} \in \mathbb{R}^k$, $\mathbf{b}_1 \in \mathbb{R}^{k_1}$, $\mathbf{b}_2 \in \mathbb{R}^{k_2}$ максимизация в последнем равенстве уже производится по произвольным $\mathbf{b} \in \mathbb{R}^k$, $\mathbf{v}_1 \in \mathbb{R}^{k_1}$, $\mathbf{v}_2 \in \mathbb{R}^{k_2}$.
Тогда равенство (4.21) приобретает вид

$$s(g_1,\, g_2) = \lim_{\tau \to 0} \frac{|Q_1||Q_2| \int g_1(\mathbf{w},\, \hat{\mathbf{w}}_1) g_2(\mathbf{w},\, \hat{\mathbf{w}}_2) d\mathbf{w}}{|Q_1||Q_2| \max_{\mathbf{b},\, \mathbf{v}_1,\, \mathbf{v}_2} \int g_1(\mathbf{v} - \mathbf{b},\, \mathbf{v}_1) g_2(\mathbf{v},\, \mathbf{v}_2) d\mathbf{v}} =$$
$$\frac{\int g_1(\mathbf{w},\, \mathbf{0}) g_2(\mathbf{w},\, \mathbf{0}) d\mathbf{w}}{\max_{\mathbf{b},\, \mathbf{v}_1,\, \mathbf{v}_2} \int g_1(\mathbf{v} - \mathbf{b},\, \mathbf{v}_1) g_2(\mathbf{v},\, \mathbf{v}_2) d\mathbf{v}} \text{ в силу } \hat{\mathbf{w}}_1 \in Q_1,\ \hat{\mathbf{w}}_2 \in Q_2 \text{ и } \tau \to 0.$$

Таким образом, получаем, что предел существует, конечен, так как при $\mathbf{b} = \mathbf{0}$, $\mathbf{v}_1 = \mathbf{0}$, $\mathbf{v}_2 = \mathbf{0}$ значением максимизируемого значения в знаменателе совпадает со значением числителя, а $s(g_1,\, g_2) \geq 0$, что и требовалось. □

Теорема 7 (Адуенко, 2014)**.** Предалагаемая функция сходства s-score является корректной.

Доказательство. Покажем, что для s-score выполнены требования к функции сходства, которые задают корректность функции сходства.

Выполнимость свойства 1 показана в утверждении 1. Свойство 6 выполнено в силу симметричности определения s-score

$$s(g_1,\, g_2) = \frac{\int g_1(\mathbf{w},\, \mathbf{0}) g_2(\mathbf{w},\, \mathbf{0}) d\mathbf{w}}{\max_{\mathbf{b},\, \mathbf{v}_1,\, \mathbf{v}_2} \int g_1(\mathbf{v} - \mathbf{b},\, \mathbf{v}_1) g_2(\mathbf{v},\, \mathbf{v}_2) d\mathbf{v}} =$$
$$\frac{\int g_2(\mathbf{w},\, \mathbf{0}) g_1(\mathbf{w},\, \mathbf{0}) d\mathbf{w}}{\max_{\mathbf{b},\, \mathbf{v}_1,\, \mathbf{v}_2} \int g_1(\mathbf{v},\, \mathbf{v}_1) g_2(\mathbf{v} - \mathbf{b},\, \mathbf{v}_2) d\mathbf{v}} = s(g_2,\, g_1).$$

Так как $s(g_1,\, g_2) \geq 0$ в силу неотрицательности плотностей распределений и

$$\max_{\mathbf{b},\, \mathbf{v}_1,\, \mathbf{v}_2} \int g_1(\mathbf{v} - \mathbf{b},\, \mathbf{v}_1) g_2(\mathbf{v},\, \mathbf{v}_2) d\mathbf{v} \overset{\mathbf{b}=\mathbf{0},\, \mathbf{v}_1=\mathbf{0},\, \mathbf{v}_2=\mathbf{0}}{\geq} \int g_1(\mathbf{v},\, \mathbf{v}_1) g_2(\mathbf{v},\, \mathbf{v}_2) d\mathbf{v},$$

получаем выполнимость условия 3, то есть $s(g_1,\, g_2) \in [0,\, 1]$. Покажем теперь выполнимость условия 3. Оно имеет вид

$$\frac{\int g_1^2(\mathbf{w}) d\mathbf{w}}{\max_{\mathbf{b} \in \mathbb{R}^k} \int g_1(\mathbf{v}) g_1(\mathbf{v} - \mathbf{b}) d\mathbf{v}} = 1.$$

Рассматривая $g_2(\mathbf{v}) = g_1(\mathbf{v} - \mathbf{b})$ и пользуясь неравенство Коши-Буняковского, получим

$$\int g_1(\mathbf{v}) g_1(\mathbf{v} - \mathbf{b}) d\mathbf{v} \leq \sqrt{\int g_1^2(\mathbf{v}) d\mathbf{v}} \sqrt{\int g_1^2(\mathbf{w} - \mathbf{b}) d\mathbf{w}} = \int g_1^2(\mathbf{w}) d\mathbf{w},$$

причем при $\mathbf{b} = \mathbf{0}$ неравенство обращается в равенство. Отсюда имеем выполнимость условия 4. Выполнимость условий 3 и 4 совместно влечет выполнимость

также и условия 2.

Покажем теперь выполнимость условия 5. Рассмотрим сначала случай, когда распределение $g_2 : \Omega \times \Omega_2 \to \mathbb{R}^+$ является неинформативным относительно распределения $g_1 : \Omega \to \mathbb{R}^+$. В этом случае, обозначив $A = \mathrm{supp}(g_1)$, имеем

$$\exists\, \tau > 0,\ B :\ A \times [-\tau,\, \tau]^{\dim(\Omega_2)} \subseteq B,$$

что $g_2 = U(B)$. Тогда для функции сходства имеем

$$s(g_1, g_2) = \frac{\int g_1(\mathbf{w}) g_2(\mathbf{w},\, \mathbf{0}) d\mathbf{w}}{\max_{\mathbf{b},\, \mathbf{v}_2} \int g_1(\mathbf{v}) g_2(\mathbf{v} - \mathbf{b},\, \mathbf{v}_2) d\mathbf{v}} = \frac{\frac{1}{|B|}}{\frac{1}{|B|} \max_{\mathbf{b},\, \mathbf{v}_2} \int_A g_1(\mathbf{v}) I[(\mathbf{v} - \mathbf{b},\, \mathbf{v}_2) \in B]}$$

$$\text{так как } \int_A g_1(\mathbf{v}) I[(\mathbf{v} - \mathbf{b},\, \mathbf{v}_2) \in B] d\mathbf{v} \le \int_A g_1(\mathbf{v}) d\mathbf{v} = 1$$

и при $\mathbf{b} = \mathbf{0},\ \mathbf{v}_2 = \mathbf{0}$ равенство достигается.

Рассмотрим теперь малоинформативную последовательность распределений $g_2^1,\ \ldots,\ g_2^k,\ \ldots$, где $g_2^k :\ \mathbb{R}^k \times \mathbb{R}^{k_2} :\ \mathbb{R}^+$. Покажем, что для произвольного распределения $g_1 : \mathbb{R}^k \to \mathbb{R}^+$ выполнено

$$s(g_1,\, g_2^k) \to 1 \text{ при } k \to \infty.$$

Таким образом, требуется показать, что

$$\frac{\int g_1(\mathbf{w}) g_2^k(\mathbf{w},\, \mathbf{0}) d\mathbf{w}}{\max_{\mathbf{b},\, \mathbf{v}_2} \int g_1(\mathbf{v}) g_2^k(\mathbf{v} - \mathbf{b},\, \mathbf{v}_2) d\mathbf{v}} \to 1 \text{ при } k \to \infty.$$

Обозначим $Q_a = \{\mathbf{w},\, \mathbf{w}_2 :\ \|(\mathbf{w},\, \mathbf{w}_2)| \ge a\}$, $R_a = \{\mathbf{w},\, \mathbf{w}_2 :\ \|(\mathbf{w},\, \mathbf{w}_2)| \le a\}$. Из определения малоинформативной последовательности имеем

$$\exists\, 0 \le B < \infty,\ \exists\, k_0 :\ \forall\, k \ge k_0 \sup_{Q_B} g_2^k(\mathbf{w},\, \mathbf{w}_2) \le \sup_{R_B} g_2^k(\mathbf{w},\, \mathbf{w}_2).$$

Зафиксируем $\varepsilon > 0$. Определим B_ε так, что

$$\int_{\{\mathbf{w}:\, \|\mathbf{w}\| \ge B_\varepsilon\}} g_1(\mathbf{v}) d\mathbf{v} < \varepsilon.$$

Определим $\tilde{B} = \max(B,\, B_\varepsilon)$. Зафиксируем также $\delta > 0$. Из определения малоинформативной последовательности имеем

$$\exists\, k_\delta :\ \forall\, k \ge k_\delta \frac{\sup_{R_{\tilde{B}}} g_2^k(\mathbf{w},\, \mathbf{w}_2)}{\inf_{R_{\tilde{B}}} g_2^k(\mathbf{w},\, \mathbf{w}_2)} \le 1 + \delta.$$

Определим $\tilde{k} = \max(k_\delta,\, k_0)$. Тогда для $k \ge \tilde{k}$ имеем

$$\int g_1(\mathbf{w}) g_2^k(\mathbf{w},\, \mathbf{0}) d\mathbf{w} \ge \int_{\{\mathbf{w}:\, \|\mathbf{w}\| \le B\}} g_1(\mathbf{w}) g_2^k(\mathbf{w},\, \mathbf{0}) d\mathbf{w} \ge$$

$$\inf_{\{\mathbf{w}:\, \|\mathbf{w}\| \le B\}} g_2^k(\mathbf{w}, \mathbf{0}) \int_{\{\mathbf{v}:\, \|\mathbf{v}\| \le B\}} g_1(\mathbf{v}) d\mathbf{v} \ge (1-\varepsilon) \inf_{R_B} g_2^k(\mathbf{w}, \mathbf{w}_2) \ge (1-\varepsilon)(1-\delta) \sup_{R_B} g_2^k(\mathbf{w}, \mathbf{w}$$

(

Аналогично для знаменателя выражения для s-score с учетом свойства (4.16) имеем

$$\forall\, \mathbf{b},\ \mathbf{v}_2 \int g_1(\mathbf{v}) g_2^k(\mathbf{v}-\mathbf{b},\ \mathbf{v}_2) d\mathbf{v} \le \sup_{\mathbf{v},\,\mathbf{v}_2} g_2^k(\mathbf{v},\ \mathbf{v}_2) = \sup_{R_B} g_2^k(\mathbf{v},\ \mathbf{v}_2). \qquad (4.23)$$

Тогда из (4.22) и (4.23) получаем

$$\forall\, k \ge \tilde{k}:\ s(g_1,\ g_2^k) \ge (1-\varepsilon)(1-\delta).$$

С учетом произвольности выбора ε, δ получаем требуемое. □

4.4. KL-информативность

Рассмотрим теперь еще альтернативное определение информативности [**?**] и порожденную им функцию сходства, основанное на подсчете дивергенции Кульбака-Лейблера между распределением и его уточнением, полученным путем использования другого распределения как априорного. Покажем, что хотя такая функция сходства удовлетворяет части требований к функции сходства, в частности требованию 5 относительно одного из пары распределений, она не является корректной и потому не подходит для решения поставленной задачи сравнения моделей.

Определение 33. Назовем распределение $\tilde{g}_1(\mathbf{w})$ *уточнением* $g_1(\mathbf{w}) : \mathbb{R}^k \to \mathbb{R}^+$ относительно $g_2(\mathbf{w}) : \mathbb{R}^k \to \mathbb{R}^+$, если

$$\tilde{g}_1(\mathbf{w}) = \frac{g_1(\mathbf{w}) g_2(\mathbf{w})}{\int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v}} \text{ и } \int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v} > 0.$$

Определение 34. Назовем *KL-информативностью* распределения $g_2(\mathbf{w}) : \mathbb{R}^k \to \mathbb{R}^+$ относительно распределения $g_1(\mathbf{w}) : \mathbb{R}^k \to \mathbb{R}^+$

$$I_{KL}(g_2|g_1) = \begin{cases} \infty, \text{ если } \int g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w} = 0, \\ D_{KL}(\tilde{g}_1 || g_1), \text{ если } \int g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w} > 0. \end{cases}$$

Замечание 1. Отметим, что для распределений, определенных на одном пространстве, KL-информативность определена корректно, поскольку $\mathrm{supp}(\tilde{g}_1) \subseteq \mathrm{supp}(g_1)$.

Замечание 2. Отметим, что введенное определение *уточнения* $\tilde{g}_1(\mathbf{w})$ распределения $g_1(\mathbf{w})$ относительно $g_2(\mathbf{w})$ фактически есть апостериорное распределение на параметр $\mathbf{w}$ для модели с правдоподобием, задаваемым $g_1(\mathbf{w})$ и априорным распределением $g_2(\mathbf{w})$. *KL-информативность* распределения $g_2(\mathbf{w})$ относительно $g_1(\mathbf{w})$ тогда означает в терминах дивергенции Кульбака-Лейблера объем информации, которую несет априорное распределение по сравнению с тем, что уже получено из данных.

Определение 35. Назовем *несимметричным KL-сходством* (или *функцией сходства, порожденной KL-информативностью*) распределения $g_2(\mathbf{w}) : \mathbb{R}^k \to \mathbb{R}^+$ относительно распределения $g_1(\mathbf{w}) : \mathbb{R}^k \to \mathbb{R}^+$

$$s_{KL}(g_2|g_1) = \exp(-I_{KL}(g_2|g_1)).$$

Рассмотрим несколько частных случаев для иллюстрации свойств введенного несимметричного KL-сходства.

Лемма 1. Пусть для $g_1(\mathbf{w}) : \mathbb{R}^k \to \mathbb{R}^+$ $\mathrm{supp}(g_1) = A$, а $g_2(\mathbf{w}) : \mathbb{R}^k \to \mathbb{R}^+$ принимает постоянное положительное значение на A, то есть $g_2(\mathbf{w}) \equiv c > 0 \; \forall \; \mathbf{w} \in A$. Тогда g_2 будет *KL-неинформативным* относительно g_1, то есть $I_{KL}(g_2|g_1) = 0$, а $s_{KL}(g_2|g_1) = 1$.

Доказательство. Лемма непосредственно следует из определения уточнения $\tilde{g}_1(\mathbf{w})$ для g_1 относительно g_2, поскольку

$$g_1\tilde{(\mathbf{w})} = \begin{cases} 0, \; \mathbf{w} \notin A, \\ \dfrac{cg_1(\mathbf{w})}{\int_A cg_1(\mathbf{w})d\mathbf{w}}, \; \mathbf{w} \in A, \end{cases} \quad = g_1(\mathbf{w}), \text{ а } D_{KL}(g_1||g_1) = 0.$$

□

Следствие 3. Несимметричное KL-сходство удовлетворяет первой части требования 5 к корректной функции сходства относительно g_1 для пары распределений $g_1 \; \mathbb{R}^k \to \mathbb{R}^+$, $g_2 : \mathbb{R}^k \to \mathbb{R}^+$, определенных на одном пространстве.

Доказательство. Требуемое непосредственно следует из того, что если $A = \mathrm{supp}(g_1)$ и $g_2 = U(B)$, $A \subseteq B$, то $g_2(\mathbf{v}) \equiv \frac{1}{|B|}$ для $\mathbf{v} \in A$, откуда из леммы получаем $s_{KL}(g_2|g_1) = 1$. □

Замечание 1. Выполнение первой части требования 5 относительно g_1 означает выполнение его для распределения $g_2(\mathbf{w}) = U(B)$, что $A = \mathrm{supp}(g_1) \subseteq B$. Однако при $g_1(\mathbf{w}) = U(B)$, что $A = \mathrm{supp}(g_2) \subset B$ свойство 5 относительно g_2 не выполнено, так как $\tilde{g}_1 \neq g_1$ в силу $\mathrm{sup}(\tilde{g}_1) = A \neq B = \mathrm{supp}(g_1)$. Такая несимметричность имеет место в силу несимметричности KL-информативности, а потому несимметричности и функции сходства, ею порожденной.

Лемма 2. Пусть $g_1(\mathbf{w}) = N(\mathbf{w}|\mathbf{m}_1, \; \mathbf{\Sigma}_1)$, $g_2(\mathbf{w}) = N(\mathbf{w}|\mathbf{m}_2, \; \mathbf{\Sigma}_2)$, $\mathbf{w} \in \mathbb{R}^k$. Тогда выражение для KL-информативности $I_{KL}(g_2|g_1)$ распределения g_2 относительно распределения g_1 имеет вид

$$I_{KL}(g_2|g_1) = \tfrac{1}{2}\left(tr(\mathbf{\Sigma}_1^{-1}(\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})^{-1}) - k + \log(\det\mathbf{\Sigma}_1 \det(\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})) + \right.$$
$$\left. (\mathbf{m}_1 - \mathbf{m}_2)^{\intercal}\mathbf{\Sigma}_2^{-1}(\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})^{-1}\mathbf{\Sigma}_1^{-1}(\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})^{-1}\mathbf{\Sigma}_2^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\right). \quad (4.24)$$

Доказательство. Получим сначала выражение для $\tilde{g}_1(\mathbf{w})$. С учетом того, что $g_1(\mathbf{w})$ и $g_2(\mathbf{w})$ есть нормальные распределения, то $\tilde{g}_1(\mathbf{w})$ тоже нормальное распределение, так как сопряженное к нормальному распределению, также нормальное. Тогда $\tilde{g}_1(\mathbf{w}) = N(\mathbf{w}|\mathbf{m}, \mathbf{\Sigma})$, где для параметров распределения имеем

$$\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1},\ \mathbf{\Sigma}^{-1}\mathbf{m} = \mathbf{\Sigma}_1^{-1}\mathbf{m}_1 + \mathbf{\Sigma}_2^{-1}\mathbf{m}_2, \text{ откуда}$$

$$\mathbf{\Sigma} = (\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})^{-1},\ \mathbf{m} = (\mathbf{\Sigma}_1^{-1} + \mathbf{\Sigma}_2^{-1})^{-1}(\mathbf{\Sigma}_1^{-1}\mathbf{m}_1 + \mathbf{\Sigma}_2^{-1}\mathbf{m}_2).$$

Подставляя полученные выражения для $\mathbf{m}$, $\mathbf{\Sigma}$ в выражение для дивергенции Кульбака-Лейблера для пары нормальных распределений, получаем

$$\begin{aligned} D_{KL}(\tilde{g}_1||g_1) &= \tfrac{1}{2}\left(tr(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}) + (\mathbf{m}-\mathbf{m}_1)^{\intercal}\mathbf{\Sigma}_1^{-1}(\mathbf{m}_2-\mathbf{m}_1) - k + \log(\det\mathbf{\Sigma}_1/\det\mathbf{\Sigma})\right) = \\ &= \tfrac{1}{2}\left(tr(\mathbf{\Sigma}_1^{-1}(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-1}) - k + \log(\det\mathbf{\Sigma}_1\det(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})) + \right. \\ &\quad \left. (\mathbf{m}_1-\mathbf{m}_2)^{\intercal}\mathbf{\Sigma}_2^{-1}(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-1}\mathbf{\Sigma}_1^{-1}(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-1}\mathbf{\Sigma}_2^{-1}(\mathbf{m}_1-\mathbf{m}_2)\right). \end{aligned}$$

□

Следствие 4. При $\|\mathbf{\Sigma}_2^{-1}\| \to 0$ $I_{KL}(g_2|g_1) \to 0$.

Доказательство.

$$\|(\mathbf{m}_1-\mathbf{m}_2)^{\intercal}\mathbf{\Sigma}_2^{-1}(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-1}\mathbf{\Sigma}_1^{-1}(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-1}\mathbf{\Sigma}_2^{-1}(\mathbf{m}_1-\mathbf{m}_2)\| \leq$$
$$\|\mathbf{m}_1-\mathbf{m}_2\|^2\|\mathbf{\Sigma}_2^{-1}\| \to 0 \text{ при } \|\mathbf{\Sigma}_2^{-1}\| \to 0,$$

$$tr(\mathbf{\Sigma}_1^{-1}(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})^{-1}) \to k \text{ при } \|\mathbf{\Sigma}_2^{-1}\| \to 0,$$
$$\log(\det\mathbf{\Sigma}_1\det(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})) \to 0 \text{ при } \|\mathbf{\Sigma}_2^{-1}\| \to 0,$$

откуда и имеем требуемое. □

Получим теперь упрощение выражения для KL-информативности.

Лемма 3. Пусть $g_1(\mathbf{w})$, $g_2(\mathbf{w}) : \mathbb{R}^n \to \mathbb{R}^+$ есть пара распределений. Тогда

$$I_{KL}(g_2|g_1) = \frac{\int g_1(\mathbf{w})g_2(\mathbf{w})\log g_2(\mathbf{w})d\mathbf{w}}{\int g_1(\mathbf{v})g_2(\mathbf{v})d\mathbf{v}} - \log\int g_1(\mathbf{v})g_2(\mathbf{v})d\mathbf{v}. \tag{4.25}$$

Доказательство. По определению уточнения $\tilde{g}_1$ имеем для KL-информативности

$$I_{KL}(g_2|g_1) = D_{KL}(\tilde{g}_1||g_1) = \int \frac{g_1(\mathbf{w})g_2(\mathbf{w})}{\int g_1(\mathbf{v})g_2(\mathbf{v})d\mathbf{v}}\log\left(\frac{\frac{g_1(\mathbf{w})g_2(\mathbf{w})}{\int g_1(\mathbf{v})g_2(\mathbf{v})d\mathbf{v}}}{g_1(\mathbf{w})}\right) =$$
$$= \frac{\int g_1(\mathbf{w})g_2(\mathbf{w})\log g_2(\mathbf{w})d\mathbf{w}}{\int g_1(\mathbf{v})g_2(\mathbf{v})d\mathbf{v}} - \log\int g_1(\mathbf{v})g_2(\mathbf{v})d\mathbf{v}.$$

□

Покажем теперь, что в случае, когда пара распределений g_1, g_2 определены на одном пространстве, то есть $g_1,\ g_2 : \mathbb{R}^k \to \mathbb{R}^+$, несимметичное KL-сходство удовлетворяет второй части требования 5 к корректной функции сходства относительно g_1.

Лемма 4. Для несимметричного KL-сходства выполнена вторая часть требования 5 к корректной функции сходства относительно g_1, то есть для малоинформативной последовательности $g_2^1,\ \ldots,\ g_2^l,\ \ldots,\ g_2^l : \mathbb{R}^k \to \mathbb{R}^+$ для любого распределения $g_1 : \mathbb{R}^k \to \mathbb{R}^+$ $s_{KL}(g_1,\ g_2^l) \to 1$ при $l \to \infty$.

Доказательство. Обозначим $Q_a = \{\mathbf{w} :\ \|\mathbf{w}| \geq a\}$, $R_a = \{\mathbf{w} :\ \|\mathbf{w}| \leq a\}$. Из определения малоинформативной последовательности имеем

$$\exists\, 0 \leq B < \infty,\ \exists\, k_0 :\ \forall\, k \geq k_0 :\ \sup_{Q_B} g_2^k(\mathbf{w}) \leq \sup_{R_B} g_2^k(\mathbf{w}).$$

Зафиксируем $\varepsilon > 0$. Определим B_ε так, что

$$\int_{\mathbf{w} \in Q_{B_\varepsilon}} g_1(\mathbf{v}) d\mathbf{v} < \varepsilon.$$

Определим $\tilde{B} = \max(B,\ B_\varepsilon)$. Зафиксируем также $\delta > 0$. Из определения малоинформативной последовательности имеем

$$\exists\, k_\delta :\ \forall\, k \geq k_\delta\ \frac{\sup_{R_{\tilde{B}}} g_2^k(\mathbf{w})}{\inf_{R_{\tilde{B}}} g_2^k(\mathbf{w})} \leq 1 + \delta.$$

Также из свойства малоинформативности последовательности имеем, что

$$\exists k_1 :\ \forall\, k \geq k_1 :\ \sup\ g_2^k(\mathbf{w}) < 1/e.$$

Это следует из того, что при $k \geq k_0$ $\sup g_2^k = \sup_{R_B}(g_2^k) \to 0$ в силу того, что $\forall\, a > 0\ g_2^k(\cdot)|_{Q_a} \to U(Q_a)$, откуда, взяв $a > B$, получаем $\sup_{R_B} g_2^k(\mathbf{w}) \to 1/|Q_a|$, что можно сделать сколь угодно малым.

Определим $\tilde{k} = \max(k_\delta,\ k_0,\ k_1)$. Тогда для $k \geq \tilde{k}$ с учетом $\log g_2^k(\mathbf{w}) < 0$ имеем

$$I_{KL}(g_2^k|g_1) = \frac{\int g_1(\mathbf{w}) g_2(\mathbf{w}) \log g_2(\mathbf{w}) d\mathbf{w}}{\int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v}} - \log \int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v} =$$

$$\frac{\int_{R_{\tilde{B}}} g_1(\mathbf{w}) g_2(\mathbf{w}) \log g_2(\mathbf{w}) d\mathbf{w} + \int_{Q_{\tilde{B}}} g_1(\mathbf{w}) g_2(\mathbf{w}) \log g_2(\mathbf{w}) d\mathbf{w}}{\int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v}} - \log \int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v} =$$

$$\frac{\int_{R_{\tilde{B}}} g_1(\mathbf{w}) g_2(\mathbf{w}) \log g_2(\mathbf{w}) d\mathbf{w}}{\int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v}} - \log \int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v} + \frac{\int_{Q_{\tilde{B}}} g_1(\mathbf{w}) g_2(\mathbf{w}) \log g_2(\mathbf{w}) d\mathbf{w}}{\int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v}} =$$

$$\frac{\log g_2(\hat{\mathbf{w}}) \int_{R_{\tilde{B}}} g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w}}{\int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v}} - \log \int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v} + \frac{\int_{Q_{\tilde{B}}} g_1(\mathbf{w}) g_2(\mathbf{w}) \log g_2(\mathbf{w}) d\mathbf{w}}{\int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v}} =$$

$$= \log g_2(\hat{\mathbf{w}}) - \log \int g_1(\mathbf{v}) g_2(\mathbf{v}) d\mathbf{v} + \frac{\int_{Q_{\tilde{B}}} g_1(\mathbf{w}) \frac{g_2(\mathbf{w})}{g_2(\hat{\mathbf{w}})} \log \left(\frac{g_2(\mathbf{w})}{g_2(\hat{\mathbf{w}})} \right) d\mathbf{w}}{\int g_1(\mathbf{v}) \frac{g_2(\mathbf{v})}{g_2(\hat{\mathbf{w}})} d\mathbf{v}}. \quad (4.26)$$

Здесь $\hat{\mathbf{w}} \in R_{\tilde{B}}$. Для второго слагаемого в (4.26) получим

$$\log \int g_1(\mathbf{v}) g_2^k(\mathbf{v}) d\mathbf{v} = \log \left(\int_{R_{\tilde{B}}} g_1(\mathbf{v}) g_2^k(\mathbf{v}) d\mathbf{v} + \int_{Q_{\tilde{B}}} g_1(\mathbf{v}) g_2^k(\mathbf{v}) d\mathbf{v} \right) \geq$$

$$\log \int_{R_{\tilde{B}}} g_1(\mathbf{v}) g_2^k(\mathbf{v}) d\mathbf{v} = \log g_2(\tilde{\mathbf{w}}) + \log \int_{R_{\tilde{B}}} g_1(\mathbf{v}) d\mathbf{v} \geq \log g_2(\tilde{\mathbf{w}}) + \log(1-\varepsilon).$$

Здесь $\tilde{\mathbf{w}} \in R_{\tilde{B}}$. Рассмотрим теперь числитель и знаменатель третьего слагаемого в (4.26). Для знаменателя имеем

$$\int g_1(\mathbf{v}) \tfrac{g_2(\mathbf{v})}{g_2(\hat{\mathbf{w}})} d\mathbf{v} \geq \int_{R_{\tilde{B}}} g_1(\mathbf{v}) \tfrac{g_2(\mathbf{v})}{g_2(\hat{\mathbf{w}})} d\mathbf{v} \geq \frac{1}{1+\delta} \int_{R_{\tilde{B}}} g_1(\mathbf{v}) d\mathbf{v} \geq (1-\varepsilon)(1-\delta).$$

Для числителя третьего слагаемого в (4.26) имеем следующие оценки.

$$\frac{g_2^k(\mathbf{w})}{g_2^k(\hat{\mathbf{w}})} \leq \frac{\sup_{R_{\tilde{B}}} g_2^k(\mathbf{w})}{\inf_{R_{\tilde{B}}} g_2^k(\mathbf{w})} \leq 1+\delta.$$

Тогда

$$\int_{Q_{\tilde{B}}} g_1(\mathbf{w}) \tfrac{g_2(\mathbf{w})}{g_2(\hat{\mathbf{w}})} \log \left(\tfrac{g_2(\mathbf{w})}{g_2(\hat{\mathbf{w}})} \right) d\mathbf{w} \leq (1+\delta) \log(1+\delta) \int_{Q_{\tilde{B}}} g_1(\mathbf{w}) d\mathbf{w} \leq \varepsilon\delta(1+\delta).$$

Отсюда с учетом (4.26) имеем

$$I_{KL}(g_2^k | g_1) \leq \log(\hat{\mathbf{w}}) - \log(\tilde{\mathbf{w}}) + \log(1-\varepsilon) + \frac{\varepsilon\delta(1+\delta)}{(1-\varepsilon)(1-\delta)} \leq \log(1+\delta) + \log(1-\varepsilon) + \frac{\varepsilon\delta(1}{(1-\varepsilon}$$

С учетом произвольности выбора $\varepsilon > 0$ и $\delta > 0$ имеем требуемое. □

Замечание 1. Выполнение второй части требования 5 относительно g_1 означает выполнение его для малоинформативной последовательности $g_2^1, \ldots, g_2^l, \ldots$ относительно g_1, но не для малоинформативной последовательности $g_1^1, \ldots, g_1^l, \ldots$ относительно g_2. Так, рассмотрев $g_2 = U(\{\mathbf{w} : \|\mathbf{w}\| \leq 1\})$ и взяв $g_1^l = U(\{\mathbf{w} : \|\mathbf{w}\| \leq l\})$, получим $\tilde{g}_1^l = g_2 \neq g_1^l$. Выражение для $I_{KL}(g_2 | g_1^l)$ с учетом (4.25) имеет вид

$$\frac{\frac{k!}{l^k} k! \log k! \frac{1}{k!}}{\frac{1}{k!} \frac{k!}{l^k} k!} - \log \left(\tfrac{1}{k!} \tfrac{k!}{l^k} k! \right) = k \log l \not\to 0 \text{ при } l \to \infty.$$

Такая несимметричность имеет место в силу несимметричности KL-информативности, а потому несимметричности и функции сходства, ею порожденной.

Отметим, что до сих пор рассматривалась пара распределений g_1, g_2, определенных на одном пространстве, то есть $g_1 : \mathbb{R}^k \to \mathbb{R}^+$, $g_2 : \mathbb{R}^k \to \mathbb{R}^+$. Определим теперь KL-информативность и функцию сходства, ею порожденную, для пары распределений $g_1(\mathbf{w},\ \mathbf{w}_1) :\ \mathbb{R}^k \times \mathbb{R}^{k_1} \to \mathbb{R}^+$, $g_2(\mathbf{w},\ \mathbf{w}_2) :\ \mathbb{R}^k \times \mathbb{R}^{k_2} \to \mathbb{R}^+$. Для этого используем тот же способ, что и для определения предлагаемой функции сходства s-score, через τ – расширение распределений равномерными.

Определение 36. Назовем *KL-информативностью* распределения $g_2(\mathbf{w}, \mathbf{w}_2) : \mathbb{R}^k \times \mathbb{R}^{k_2} \to \mathbb{R}^+$ относительно распределения $g_1(\mathbf{w},\ \mathbf{w}_1) : \mathbb{R}^k \times \mathbb{R}^{k_1} \to \mathbb{R}^+$ следующий предел

$$I_{KL}(g_2|g_1) = \lim_{\tau \to 0} I(g_2^\tau | g_1^\tau).$$

Определение 37. Назовем *несимметричным KL-сходством* (или *функцией сходства, порожденной KL-информативностью*) распределения $g_2(\mathbf{w},\ \mathbf{w}_2) : \mathbb{R}^k \times \mathbb{R}^{k_2} \to \mathbb{R}^+$ относительно распределения $g_1(\mathbf{w},\ \mathbf{w}_1) : \mathbb{R}^k \times \mathbb{R}^{k_1} \to \mathbb{R}^+$

$$s_{KL}(g_2|g_1) = \exp(-I_{KL}(g_2|g_1)).$$

Получим теперь выражения для KL-информативности в общем случае, пользуясь (4.25). Рассмотрим сначала случай, когда $k_1 = 0$.

$$\int g_1^\tau(\mathbf{w},\ \mathbf{w}_2) g_2(\mathbf{w},\ \mathbf{w}_2) d\mathbf{w} d\mathbf{w}_2 = \frac{1}{|Q_2^\tau|} \int_{Q_2^\tau} d\mathbf{w}_2 \int d\mathbf{w} g_1(\mathbf{w}) g_2(\mathbf{w},\ \mathbf{w}_2) = \int g_1(\mathbf{w}) g_2(\mathbf{w},\ \tilde{\mathbf{w}}_2) d\mathbf{w},\ \tilde{\mathbf{w}}_2 \in Q_2^\tau.$$

$$\int g_1^\tau(\mathbf{w}, \mathbf{w}_2) g_2(\mathbf{w}, \mathbf{w}_2) \log g_2(\mathbf{w}, \mathbf{w}_2) d\mathbf{w} d\mathbf{w}_2 = \frac{1}{|Q_2^\tau|} \int_{Q_2^\tau} d\mathbf{w}_2 \int g_1(\mathbf{w}) g_2(\mathbf{w}, \mathbf{w}_2) \log g_2(\mathbf{w}$$
$$\int g_1(\mathbf{w}) g_2(\mathbf{w},\ \hat{\mathbf{w}}_2) \log g_2(\mathbf{w},\ \hat{\mathbf{w}}_2) d\mathbf{w},\ \hat{\mathbf{w}}_2 \in Q_2^\tau.$$

Тогда при $\tau \to 0$ получаем

$$I_{KL}(g_2|g_1) = \frac{\int g_1(\mathbf{w}) g_2(\mathbf{w},\ \mathbf{0}) \log g_2(\mathbf{w},\ \mathbf{0}) d\mathbf{w}}{\int g_1(\mathbf{w}) g_2(\mathbf{w},\ \mathbf{0}) d\mathbf{w}} - \int g_1(\mathbf{w}) g_2(\mathbf{w},\ \mathbf{0}) d\mathbf{w}.$$

Отсюда в силу совпадения выражения для KL-информативности для случая $k_1 = 0$, $k_2 \neq 0$ со случаем $k_1 = k_2 = 0$ с точностью до замены $g_2(\mathbf{w})$ на $g_2(\mathbf{w},\ \mathbf{0})$ получаем, что при $k_1 = 0$ свойство 5 из требований к корректной функции сходства выполнено для несимметричного KL-сходства относительно g_2, но не относительно g_1, как указывалось ранее.

Рассмотрим теперь случай, когда $k_1 \neq 0$. В этом случае аналогично имеем

$$\int g_1^\tau(\mathbf{w}, \mathbf{w}_1, \mathbf{w}_2) g_2^\tau(\mathbf{w}, \mathbf{w}_1, \mathbf{w}_2) d\mathbf{w} d\mathbf{w}_1 d\mathbf{w}_2 = \frac{1}{|Q_1^\tau||Q_2^\tau|} \int_{Q_1^\tau} d\mathbf{w}_1 \int_{Q_2^\tau} d\mathbf{w}_2 \int d\mathbf{w} g_1(\mathbf{w}, \mathbf{w}_1$$

$$\int g_1(\mathbf{w},\ \tilde{\mathbf{w}}_1) g_2(\mathbf{w},\ \tilde{\mathbf{w}}_2) d\mathbf{w},\ \tilde{\mathbf{w}}_1 \in Q_1^\tau,\ \tilde{\mathbf{w}}_2 \in Q_2^\tau.$$

$$\int g_1^\tau(\mathbf{w},\ \mathbf{w}_1,\ \mathbf{w}_2) g_2^\tau(\mathbf{w},\ \mathbf{w}_1,\ \mathbf{w}_2) \log g_2^\tau(\mathbf{w},\ \mathbf{w}_1,\ \mathbf{w}_2) d\mathbf{w} d\mathbf{w}_1 \mathbf{w}_2 =$$

$$\frac{1}{|Q_1^\tau||Q_2^\tau|} \int_{Q_1^\tau} d\mathbf{w}_1 \int_{Q_2^\tau} d\mathbf{w}_2 \int g_1(\mathbf{w},\ \mathbf{w}_1) g_2(\mathbf{w},\ \mathbf{w}_2)(\log g_2(\mathbf{w},\ \mathbf{w}_2) - \log |Q_1^\tau|) d\mathbf{w} =$$

$$\int g_1(\mathbf{w}, \hat{\mathbf{w}}_1) g_2(\mathbf{w}, \hat{\mathbf{w}}_2) \log g_2(\mathbf{w}, \hat{\mathbf{w}}_2) d\mathbf{w} - \log |Q_1^\tau| \int g_1(\mathbf{w}, \hat{\mathbf{w}}_1) g_2(\mathbf{w}, \hat{\mathbf{w}}_2) d\mathbf{w}, \hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2 \in Q$$

Тогда получим при $\tau \to 0$

$$I_{KL}(g_2|g_1) + \log |Q_1^\tau| \to \frac{\int g_1(\mathbf{w},\ \mathbf{0}) g_2(\mathbf{w},\ \mathbf{0}) \log g_2(\mathbf{w},\ \mathbf{0}) d\mathbf{w}}{\int g_1(\mathbf{w},\ \mathbf{0}) g_2(\mathbf{w},\ \mathbf{0}) d\mathbf{w}} - \int g_1(\mathbf{w}, \mathbf{0}) g_2(\mathbf{w}, \mathbf{0}) d\mathbf{w},$$

откуда при $k_1 > 0$ $I_{KL}(g_2|g_1) = \infty$ и $s_{KL}(g_1,\ g_2) = 0$. Таким образом, распределение g_2 будет информативным для g_1 всегда, если $k_1 > 0$, то есть если есть набор параметров, про которые во второй модели известно, что они равны $\mathbf{0}$, а для первой модели есть некоторая неопределенность относительно их значений. Отметим, что при $k_1 = 0$ свойство 5 выполнено относительно g_2, но в силу несимметричности функции сходства, порожденной KL-информативностью, это свойство не выполнено относительно g_1 и в частности при $k_1 \neq 0$, $k_2 = 0$. Опишем далее нарушения требований корректности к функции сходства для несимметричного KL-сходства.

Утверждение 11. Несимметричное KL-сходство не удовлетворяет свойствам 2 и 4 из требований к корректной функции сходства, то есть модель отличима от самой себя.

Доказательство. Свойство 2 требует выполнения $s_{KL}(g_1,\ g_1) = 1$, откуда $I_{KL}(g_1|g_1) = 0$, что не выполнено. Так как $\tilde{g}_1$ есть уточнение распределения g_1 с помощью g_1, то в случае, если g_1 не является равномерным, то $\tilde{g}_1 \neq g_1$, откуда $I_{KL}(g_1|g_1) > 0$.

Рассмотрим в качестве примера $g_1 : \mathbb{R}^k \to \mathbb{R}^+$, $g_1(\mathbf{w}) = N(\mathbf{w}|\mathbf{m},\ \mathbf{\Sigma})$. Тогда, пользуясь (4.24), получим

$$I_{KL}(g_1|g_1) = \tfrac{1}{2}\left(tr\left(\mathbf{\Sigma}^{-1}\tfrac{1}{2}\mathbf{\Sigma}\right) - k + \log(\det \mathbf{\Sigma}\ \det(2\mathbf{\Sigma}^{-1}))\right) = \tfrac{k}{2}\left(\log 2 - \tfrac{1}{2}\right) > 0.$$

Таким образом, свойство 4 не выполнено. Отсюда же следует при $s(g_1,\ g_1) < 1$ в силу выполнения свойства 5 относительно g_1, что для малоинформативной последовательности $g_2^1,\ \ldots,\ g_2^l,\ \ldots\ s_{KL}(g_1,\ g_2^l) \to 1$ при $l \to \infty$, откуда $\exists l_0 :$ $\forall\, l \geq l_0 :\ s(g_1,\ g_2^l) \geq \frac{1+s(g_1,\, g_1)}{2} > s(g_1,\ g_1)$, что означает невыполнение свойства 2. $\square$

Ранее также было показано, что несимметричное KL-сходство удовлетворяет свойству 5 относительно g_1, но не относительно g_2 в силу несимметричности, что можно сформулировать следующим образом.

Утверждение 12. Несимметричное KL-сходство не удовлетворяет свойству 6 к корректной функции сходства, то есть не является симметричным, а также не удовлетворяет свойству 5 относительно g_2.

Замечание 1. Для $g_1(\mathbf{w},\, \mathbf{w}_1) : \mathbb{R}^k \times \mathbb{R}^{k_1} \to \mathbb{R}^+$, $g_2(\mathbf{w},\, \mathbf{w}_2) : \mathbb{R}^k \times \mathbb{R}^{k_2} \to \mathbb{R}^+$ при $k_1 \neq 0$ $s_{KL}(g_1,\, g_2) \equiv 0$. Таким образом, независимо от точности знания $\mathbf{w}_1$ первая модель считается отличимой от второй, где $\mathbf{w}_1 = \mathbf{0}$.

Замечание 2. Устранить несимметричность KL-сходства и достичь выполнения свойства 5 и относительно g_1, и относительно g_2, можно, определив симметричный вариант KL-информативности как

$$\tilde{s}_{KL}(g_1,\, g_2) = \max(s_{KL}(g_1,\, g_2),\, s_{KL}(g_2,\, g_1)),$$

где $s_{KL}(g_1,\, g_2)$ есть несимметричное KL-сходство относительно g_1, а $s_{KL}(g_2,\, g_1)$ есть несимметричное KL-сходство относительно g_2. В силу выполнения для первого из них требования 5 к корректной функции сходства относительно g_1, а для второго – выполнения этого свойства относительно g_2, получим, что $\tilde{s}_{KL}(g_1,\, g_2)$ является симметричной (то есть удовлетворяет свойству 6) и удовлетворяет требованию 5 к корректной функции сходства.

Отметим, однако, что свойства 2 и 4 после симметризации по-прежнему остаются невыполненными и возникают новые проблемы, связанные с использованием оператора взятия максимума, что приводит к тому, что слишком много пар распределений начинают считаться неразличимыми, что приближает симметризованное KL-сходство к тривиальному, для которого $s_{tr}(g_1,\, g_2) \equiv 1$. Для иллюстрации рассмотрим следующий пример в одномерном случае, то есть для $g_1,\, g_2 : \mathbb{R} \to \mathbb{R}^+$ (см. рис. 4.2), где

$$g_1(w) = \tfrac{1}{2} I(w \in [-1,\, 1]),\; g_2 = \begin{cases} \varepsilon,\; w \in [-1,\, 1], \\ 0,\; |w| \in (1, 2), \\ (\tfrac{1}{2} - \varepsilon) e^{2-|w|}, |w| \geq 2. \end{cases}$$

Так как $g_2(w) \equiv \varepsilon = \text{const}$ для $w \in [-1,\, 1]$ для уточнения $\tilde{g}_1$ для g_1 относительно g_2 получим $\tilde{g}_1 = g_1 = U[-1,\, 1]$, откуда $\tilde{s}_{KL}(g_1,\, g_2) = 1$, хотя информация о параметре w в рассматриваемом примере для двух моделей существенно различается.

Таким образом, были проанализированы функции сходства, порожденные дивергенциями Брегмана, f – дивергенциями и KL-информативностью, широко использующиеся в задачах сравнения распределений [50–54] и определений информативности одного распределения относительно другого [55–57] и было показано, что они не являются корректными, а потому не применимы для решаемой задачи сравнения моделей. Была также предложена функция сходства

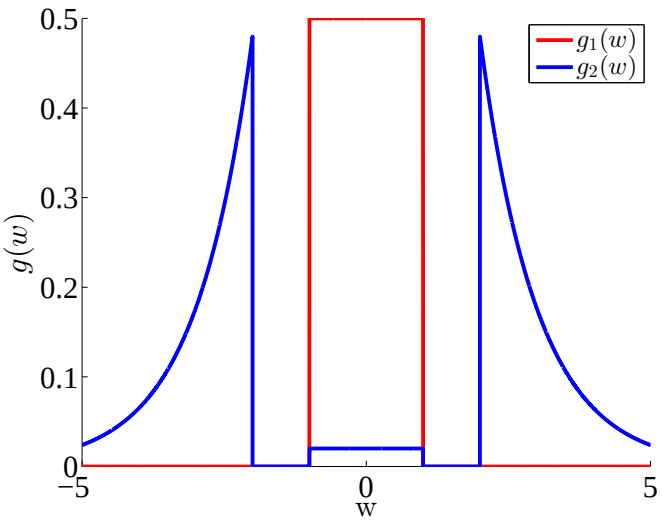


Рис. 4.2: Пример разных моделей, неразличимых с помощью симметризованного KL-сходства, $\varepsilon = 0.02$.

s-score и было показано, что предлагаемая функция сходства является корректной. Рассмотрим далее статистические свойства распределения предлагаемой функции сходства s-score в условиях истинности гипотезы о совпадении моделей.

4.5. Свойства предлагаемой функции сходства

Рассмотрим далее несколько теорем, показывающих наличие свойст монотонности для предлагаемой функции сходства при выполнении некоторых ограничений.

Теорема 8 (Адуенко, 2014)**.** Пусть модели, задаваемые математическим ожиданием и ковариционной матрицей апостериорного распределения параметров $(\mathbf{m}_1,\ \mathbf{\Sigma}_1)$ и $(\mathbf{m}_2,\ \mathbf{\Sigma}_2)$ считаются различимыми, если

$$\mathrm{s-score}\left(N(\mathbf{m}_1,\ \mathbf{\Sigma}_1),\ N(\mathbf{m}_2,\ \mathbf{\Sigma}_2)\right) \leq C \in (0,\ 1).$$

Тогда, если указанные модели различимы по приведенному критерию, то и модели, задаваемые $(\mathbf{m}_1,\ \mathbf{\Sigma}_1)$ и $(\mathbf{m}_2,\ \boldsymbol{O})$ будут различимыми согласно приведенному критерию.

Доказательство.

$$\mathrm{s-score}\left(N(\mathbf{m}_1,\ \mathbf{\Sigma}_1),\ N(\mathbf{m}_2,\ \boldsymbol{O})\right) = \exp\left(-\tfrac{1}{2}(\mathbf{m}_1-\mathbf{m}_2)^{\intercal}(\mathbf{\Sigma}_1)^{-1}(\mathbf{m}_1-\mathbf{m}_2)\right) \leq$$
$$\leq \exp\left(-\tfrac{1}{2}(\mathbf{m}_1-\mathbf{m}_2)^{\intercal}(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2)^{-1}(\mathbf{m}_1-\mathbf{m}_2)\right) \leq C.$$

Приведенное выше справедливо, так как матрица $\mathbf{\Sigma}_1^{-1} - (\mathbf{\Sigma}_1+\mathbf{\Sigma}_2)^{-1}$ неотрицательно определена

$$\mathbf{\Sigma}_1^{-1} - (\mathbf{\Sigma}_1+\mathbf{\Sigma}_2)^{-1} = \mathbf{\Sigma}_1^{-1}(\mathbf{I} - (\mathbf{I}+\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1})^{-1}),$$

а в силу положительной определенности $\mathbf{\Sigma}_1$ и неотрицательной определенности $\mathbf{\Sigma}_2$ собственные значения матрицы $(\mathbf{I}+\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1})$ не меньше единицы, откуда все собственные значения матрицы $\mathbf{I}-(\mathbf{I}+\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{-1})^{-1}$ неотрицательны. □

Теорема 9 (Адуенко, 2014)**.** Пусть модели, задаваемые математическим ожиданием и ковариционной матрицей апостериорного распределения параметров $(\mathbf{m}_1,\ \mathbf{\Sigma}_1)$ и $(\mathbf{m}_2,\ \mathbf{\Sigma}_2)$ считаются различимыми, если

$$\mathrm{s}-\mathrm{score}\,(N(\mathbf{m}_1,\ \mathbf{\Sigma}_1),\ N(\mathbf{m}_2,\ \mathbf{\Sigma}_2)) \le C \in (0,\ 1).$$

Тогда, если указанные модели различимы по приведенному критерию, то и модели, задаваемые $(\mathbf{m}_1,\ \mathbf{\Sigma}_1)$ и $(\mathbf{m}_2,\ \lambda\mathbf{\Sigma}_2)$, $\lambda \in [0,\ 1]$ будут различимы согласно приведенному критерию.

Доказательство.

$$\begin{aligned}\mathrm{s}-\mathrm{score}\,(N(\mathbf{m}_1,\ \mathbf{\Sigma}_1),\ N(\mathbf{m}_2,\ \lambda\mathbf{\Sigma}_2)) &= \exp\left(-\tfrac{1}{2}(\mathbf{m}_1-\mathbf{m}_2)^{\intercal}(\mathbf{\Sigma}_1+\lambda\mathbf{\Sigma}_2)^{-1}(\mathbf{m}_1-\mathbf{m}_2)\right.\\ &\le \exp\left(-\tfrac{1}{2}(\mathbf{m}_1-\mathbf{m}_2)^{\intercal}(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2)^{-1}(\mathbf{m}_1-\mathbf{m}_2)\right) \le C.\end{aligned}$$

Приведенное выше справедливо, так как матрица $(\mathbf{\Sigma}_1+\lambda\mathbf{\Sigma}_2)^{-1}-(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2)^{-1}$ неотрицательно определена

$$(\mathbf{\Sigma}_1+\lambda\mathbf{\Sigma}_2)^{-1}-(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2)^{-1} = (\mathbf{\Sigma}_1+\lambda\mathbf{\Sigma}_2)^{-1}(\mathbf{I}-(\mathbf{I}+(1-\lambda)\mathbf{\Sigma}_2(\mathbf{\Sigma}_1+\lambda\mathbf{\Sigma}_2)^{-1})^{-1}),$$

а в силу положительной определенности $\mathbf{\Sigma}_1$ и неотрицательной определенности $\mathbf{\Sigma}_2$ собственные значения матрицы $(\mathbf{I}+(1-\lambda)\mathbf{\Sigma}_2(\mathbf{\Sigma}_1+\lambda\mathbf{\Sigma}_2)^{-1})$ не меньше единицы, откуда все собственные значения матрицы $\mathbf{I}-(\mathbf{I}+(1-\lambda)\mathbf{\Sigma}_2(\mathbf{\Sigma}_1+\lambda\mathbf{\Sigma}_2)^{-1})^{-1}$ неотрицательны. □

Замечание 1. Приведенные теоремы можно интерпретировать следующим образом. Если модель отличима от модели с некоторым средним и большей неопределенностью, то она отличима и от модели с тем же средним, но меньшей неопределенностью. Обратно: если модель неотличима от некоторой модели с меньшей неопределенностью, то она неотличима и от модели с большей неопределенностью с тем же средним.

Рассмотрим теперь теорему, которая дает верхнюю и ниэжнюю оценку на число попарно различимых моделей при некоторых условиях.

Теорема 10 (Адуенко, 2014)**.** Пусть рассматриваются K моделей с $\|\mathbf{m}_1\| = \ldots = \|\mathbf{m}_K\| = \lambda_1 > 0$ и $\mathbf{\Sigma}_1 = \ldots = \mathbf{\Sigma}_K = \lambda_2\mathbf{I}$. В качестве критерия отличимости моделей рассматривается следующий: модели с номерами $i \ne j$ разные, если

$$\mathrm{s}-\mathrm{score}\,(N(\mathbf{m}_i,\ \mathbf{\Sigma}_i),\ N(\mathbf{m}_j,\ \mathbf{\Sigma}_j)) \le C \in (0,\ 1).$$

Тогда максимальное число попарно различимых моделей, которое может быть в наборе, есть

$$K_{max} = \left\lfloor \sqrt{\pi}\frac{n\Gamma(\frac{n+1}{2})}{(n-1)\Gamma(\frac{n}{2}+1)}\frac{1}{\int_0^{\theta/2}\sin^{n-2}\varphi d\varphi} \right\rfloor,$$

при C, близком к 1, имеем

$$K_{max} \approx \left\lfloor \sqrt{\pi} \frac{n\Gamma(\frac{n+1}{2})}{(n-1)\Gamma(\frac{n}{2}+1)} \frac{2^{\frac{n-1}{2}}}{(1-\rho)^{\frac{n-1}{2}}} \right\rfloor .$$

Здесь $\theta \in [0, \pi]$, $\cos\theta = \rho = \max(-1, 1 + 2\lambda_2/\lambda_1^2 \ln C)$, n – размерность признакового пространства. При этом можно построить K_{min} попарно различимых моделей, где

$$K_{min} = \left\lfloor \sqrt{\pi} \frac{n\Gamma(\frac{n+1}{2})}{(n-1)\Gamma(\frac{n}{2}+1)} \frac{1}{\int_0^\theta \sin^{n-2}\varphi d\varphi} \right\rfloor ,$$

что при C, близком к 1, имеем

$$K_{min} \approx \left\lfloor \sqrt{\pi} \frac{n\Gamma(\frac{n+1}{2})}{(n-1)\Gamma(\frac{n}{2}+1)} \frac{1}{2^{\frac{n-1}{2}}(1-\rho)^{\frac{n-1}{2}}} \right\rfloor .$$

Доказательство. Заметим, что вектора $\mathbf{m}_1, \ldots, \mathbf{m}_K$ однозначно задаются точками на сфере Ω_n радиуса λ_1 в пространстве $\mathbb{R}^n$. Условие различимости двух моделей с номерами $i \neq j$ с учетом сделанных предположений записывается в виде

$$\frac{1}{2\lambda_2}(\mathbf{m}_i - \mathbf{m}_j)^\intercal(\mathbf{m}_i - \mathbf{m}_j) \leq -2\ln C,$$

что с учетом $\|\mathbf{m}_i\| = \|\mathbf{m}_j\| = \lambda_1$ переписывается в виде

$$\mathbf{m}_i^\intercal \mathbf{m}_j \geq \lambda_1^2 + 2\lambda_2 \ln C, \text{ или}$$

$$\cos\theta = \cos\widehat{\mathbf{m}_i\mathbf{m}_j} \geq \max\left(1 + 2\frac{\lambda_2}{\lambda_1^2}\ln C,\ -1\right) = \rho. \tag{4.27}$$

В силу произвольности $i \neq j$ две модели различаются, если вектор средних одной модели $\mathbf{m}_i$ не лежит в конусе $C_j(\theta)$ с осью в $\mathbf{m}_j$ и углом раствора θ, определяемым условием $\theta \in [0, \pi]$ и (4.27).

Лемма 5. Приведенное условие можно заменить равносильным: две модели различаются, если конус $C_i(\theta/2)$ с осью $\mathbf{m}_i$ и углом раствора $\theta/2$ не пересекается с конусом $C_j(\theta/2)$ с осью $\mathbf{m}_j$ и углом раствора $\theta/2$.

Доказательство. Пусть $\mathbf{m}_i \in C_j(\theta)$. Рассмотрим вектор $\mathbf{m} = \frac{1}{2}(\mathbf{m}_i + \mathbf{m}_j)$. В силу $\|\mathbf{m}_i\| = \|\mathbf{m}_j\|$ $\widehat{\mathbf{m}_i\mathbf{m}} = \widehat{\mathbf{m}\mathbf{m}_j} = 1/2\widehat{\mathbf{m}_i\mathbf{m}_j} \leq \theta/2$. Отсюда $\mathbf{m} \in C_i(\theta/2)$, $\mathbf{m} \in C_j(\theta/2)$, откуда $C_i(\theta/2) \cap C_j(\theta/2) \neq \varnothing$.

С другой стороны, если $\exists \mathbf{m} \in C_i(\theta/2) \cap C_j(\theta/2)$, то, рассматривая трехгранный угол, образованный векторами $\mathbf{m}$, $\mathbf{m}_i$ и $\mathbf{m}_j$, получим (по свойству плоских углов трехгранного угла)

$$\widehat{\mathbf{m}_i\mathbf{m}_j} \leq \widehat{\mathbf{m}_i\mathbf{m}} + \widehat{\mathbf{m}\mathbf{m}_j} \leq \theta/2 + \theta/2 = \theta,$$

откуда $\mathbf{m}_j \in C_i(\theta)$. Лемма доказана. □

Вернемся к доказательству теоремы. Опишем вокруг каждого вектора $\mathbf{m}_1, \ldots, \mathbf{m}_K$ конус с осью в этом векторе и углом раствора $\theta/2$. Каждый конус высечен на сфере Ω_n радиуса λ_1 некоторый участок. Условие попарной различимости всех моделей тогда в силу леммы 1 можно сформулировать так:

Следствие 5. Рассматриваемые модели попарно различимы тогда и только тогда, когда высекаемые конусами $C_1(\theta/2), \ldots, C_K(\theta/2)$ на сфере Ω_n радиуса λ_1 участки $P_1, \ldots, P_K$ не пересекаются.

Кроме того, из условия различимости в виде $\mathbf{m}_j \notin C_i(\theta)$ получаем следующее утверждение.

Лемма 6. Модель с номером i отличима от всех остальных моделей тогда и только тогда, когда

$$\mathbf{m}_i \notin \cup_{j=1,\, j\neq i}^{K} C_j(\theta) \Longleftrightarrow \mathbf{m}_i \notin \cup_{j=1,\, j\neq i}^{K} (C_j(\theta) \cap \Omega_n).$$

Пользуясь следствием 5, получим верхнюю оценку на число моделей. Пользусь леммой 6 выполним построение, доказывающее нижнюю оценку на число моделей. Из следствия 5 заключаем, что

$$S(P_1 \cup \ldots \cup P_K) = \sum_{i=1}^{K} S(P_i) = KS(P_i) \leq S(\Omega_n), \text{ где} \tag{4.28}$$

$S(\Omega_n)$ – площадь сферы радиуса λ_1 в $\mathbb{R}^n$, а $S(P_i)$ – площадь участка, высекаемого на Ω_n конусом $C_i(\theta/2)$.

Лемма 7.

$$S(\Omega_n \cap C_i(\alpha)) = \lambda_1^{n-1} S_{n-1} \int_0^{\alpha} \sin^{n-2}\varphi d\varphi,$$

где S_{n-1} – площадь единиченой сферы в пространстве $\mathbb{R}^{n-1}$.

Доказательство. Рассматривая в качестве выделенной оси ось конуса $C_i(\alpha)$ и обозначив x – координату вдоль этой оси для искомой площади имеем

$$S(\Omega_n \cap C_i(\alpha)) = \lambda_1^{n-1} S_{n-1} \int_{\cos\alpha}^{1} (1-x^2)^{\frac{n-2}{2}} \frac{dx}{\sqrt{1-x^2}} \overset{x=\cos\varphi}{=} \lambda_1^{n-1} S_{n-1} \int_0^{\alpha} \sin^{n-2}\varphi d\varphi.$$

Из (4.28) получаем

$$K \leq K_{max} = \frac{S(\Omega_n)}{S(P_i)} = \frac{S_n \lambda_1^{n-1}}{S_{n-1}\lambda_1^{n-1} \int_0^{\theta/2} \sin^{n-2}\varphi d\varphi} = \frac{n\dfrac{\pi^{n/2}}{\Gamma(n/2+1)}}{(n-1)\dfrac{\pi^{(n-1)/2}}{\Gamma((n-1)/2+1)}} \cdot \frac{1}{\int_0^{\theta/2} \sin$$

$$= \frac{\sqrt{\pi} n \Gamma(\frac{n-1}{2}+1)}{(n-1)\Gamma(\frac{n}{2}+1)} \cdot \frac{1}{\int_0^{\theta/2} \sin^{n-2}\varphi d\varphi}. \tag{4.29}$$

□

При малом θ (что соответствует C, близкому к 1), производя приближение $\sin\varphi \approx \varphi$ и $\theta \approx \sin\theta = \sqrt{1-\rho^2} \approx \sqrt{2(1-\rho)}$ в (4.29) получаем

$$K_{max} \approx \sqrt{\pi}\frac{n\Gamma(\frac{n+1}{2})}{(n-1)\Gamma(\frac{n}{2}+1)}\frac{2^{\frac{n-1}{2}}}{(1-\rho)^{\frac{n-1}{2}}}.$$

Перейдем к построению примера для получения нижней оценки $K \geq K_{min}$. В силу леммы 6, если

$$\cup_{i=1}^{K} C_i(\theta) \cap \Omega_n \neq \Omega_n,$$

то есть $\exists\, \mathbf{m} \in \Omega_n : \ \mathbf{m} \notin \cup_{i=1}^{K} C_i(\theta)$, то можно добавить еще одну модель с средним вектором, равным $\mathbf{m}$, и она по построению будет отлична от всех уже имеющихся моделей. Тогда получаем следующую процедуру:

- Шаг 1. Выбираем произвольный вектор $\mathbf{m}_1 \in \Omega_n$
- Шаг k. Выбираем произвольный вектор $\mathbf{m}_k \in \Omega_n \setminus \cup_{i=1}^{k-1} C_i(\theta)$, если таковой существует.

По построению получаем, что

$$K \geq \frac{S(\Omega_n)}{S(C_i(\theta) \cap \Omega_n)}.$$

Пользуясь леммой 7 аналогично получению верхней оценки получаем

$$K \geq K_{min} = \frac{S(\Omega_n)}{S(C_i(\theta) \cap \Omega_n)} = \frac{S_n \lambda_1^{n-1}}{S_{n-1}\lambda_1^{n-1}\int_0^\theta \sin^{n-2}\varphi d\varphi} = \frac{n\dfrac{\pi^{n/2}}{\Gamma(n/2+1)}}{(n-1)\dfrac{\pi^{(n-1)/2}}{\Gamma((n-1)/2+1)}} \cdot \frac{}{\int_0^\theta}$$

$$= \frac{\sqrt{\pi}n\Gamma(\frac{n-1}{2}+1)}{(n-1)\Gamma(\frac{n}{2}+1)} \cdot \frac{1}{\int_0^\theta \sin^{n-2}\varphi d\varphi}. \quad (4.30)$$

При малом θ (что соответствует C, близкому к 1), производя приближение $\sin\varphi \approx \varphi$ и $\theta \approx \sin\theta = \sqrt{1-\rho^2} \approx \sqrt{2(1-\rho)}$ в (4.30) получаем

$$K_{min} \approx \sqrt{\pi}\frac{n\Gamma(\frac{n+1}{2})}{(n-1)\Gamma(\frac{n}{2}+1)}\frac{1}{2^{\frac{n-1}{2}}(1-\rho)^{\frac{n-1}{2}}}.$$

□

Пример 1. При $n = 2$ полученные результаты есть $K_{min} = \pi/\theta$, $K_{max} = 2\pi/\theta$. При $n = 3$, $\theta = \pi/6$ $K_{min} = 14$, $K_{max} = 58$. При $n = 3$, $\theta = \pi/3$ $K_{min} = 4$, $K_{max} = 14$.

4.6. Алгоритмы построения $(s,\ \alpha)$ – адекватных мультимоделей

Определение 38. Мультимодель (смесь моделей или многоуровневая модель) называется $(s,\ \alpha)$ – адекватной, если все модели $f_1,\ \ldots,\ f_l$, входящие в нее, являются попарно статистически различимыми с помощью функции сходства s на уровне значимости α. Под статистической различимостью моделей предполагается статистическая различимость апостериорных распределений параметров моделей $p(\mathbf{w}_k|\mathbf{y},\ \mathbf{X},\ \alpha,\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K)$, $k = \overline{1,\ K}$ для мультимодели и $p(\mathbf{w}_k|\mathbf{y},\ \mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K)$, $k = \overline{1,\ K}$ для многоуровневой модели.

Замечание 1. Отметим, что определение $(s,\ \alpha)$ – адекватной мультимодели существенно зависит от используемой функции сходства s. Так, если $s = s_{tr}$, которая считает любые две модели неразличимыми, то только одиночная модель может являться $(s,\ \alpha)$ – адекватной. Напротив, если $s = s_\delta$, то есть любые два сколь угодно близких, но несовпадающих распределений считаются различимыми, то даже мультимодель, состоящая из миллиона похожих моделей будет считаться адекватной.

Алгоритмы построения $(s,\ \alpha)$ – адекватных многоуровневых моделей

Пусть имеется оптимальная обученная многоуровневая модель, заданная совместным правдоподобием

$$p(\mathbf{y}, \mathbf{w}_1, \ldots, \mathbf{w}_K|\mathbf{X}, \mathbf{A}_1^*, \ldots, \mathbf{A}_K^*) = \prod_{k=1}^{K}\left[\frac{\sqrt{\det \mathbf{A}_k^*}}{(2\pi)^{n/2}}\exp(-\tfrac{1}{2}\mathbf{w}_k^\intercal \mathbf{A}_k^* \mathbf{w}_k)\prod_{i\in\mathcal{I}_k}\sigma(y_i\mathbf{w}_k^\intercal \mathbf{x}_i)\right],$$

где $\{1,\ \ldots,\ m\} = \mathcal{I} = \mathcal{I}_1 \sqcup \ldots \sqcup \mathcal{I}_K$ есть разбиение множества индексов объектов по их принадлежности области действия каждой из моделей. Пусть также $\mathbf{w}_1^*,\ \ldots,\ \mathbf{w}_K^*$ есть оценки максимума апостериорной вероятности на векторы параметров моделей, входящих в многоуровневую модель, полученные в результате обучения (1.6).

В силу того, что многоуровневая модель представляет собой совокупность K моделей, каждая из которых действует в своей части признакового пространства Ω_k, где $\mathbb{R}^n = \Omega_1 \sqcup \ldots \Omega_K$, для апостериорных распределений на векторы параметров моделей имеем

$$p(\mathbf{w}_k|\mathbf{y},\ \mathbf{X},\ \mathbf{A}_1^*,\ \ldots\ \mathbf{A}_K^*) = p(\mathbf{w}_k|\mathbf{y}_{\mathcal{I}_k},\ \mathbf{X}_{\mathcal{I}_k},\ \mathbf{A}_k^*),\ k = \overline{1,\ K},$$

где $p(\mathbf{w}_k|\mathbf{y}_{\mathcal{I}_k},\ \mathbf{X}_{\mathcal{I}_k},\ \mathbf{A}_k^*)$ есть апостериорное распределение для вектора параметров одиночной логистической модели на выборке $(\mathbf{X}_{\mathcal{I}_k},\ \mathbf{y}_{\mathcal{I}_k})$. Пользуясь нормальной аппроксимацией, получаем

$$p(\mathbf{w}_k|\mathbf{y}_{\mathcal{I}_k},\ \mathbf{X}_{\mathcal{I}_k},\ \mathbf{A}_k^*) \approx g_k(\mathbf{w}_k) = N(\mathbf{w}_k|\mathbf{w}_k^*,\ \mathbf{\Sigma}_k),\ \text{где}$$

$$\mathbf{\Sigma}_k = \left(\mathbf{X}_{\mathcal{I}_k}^\intercal \mathbf{R}_k \mathbf{X}_{\mathcal{I}_k} + \mathbf{A}_k^*\right)^{-1},\ \text{где}\ \mathbf{R}_k = \operatorname{diag}(\sigma(\mathbf{w}_k^{*\intercal}\mathbf{x}_i)\sigma(-\mathbf{w}_k^{*\intercal}\mathbf{x}_i),\ i \in \mathcal{I}_k).$$

Так как модели, входящие в многоуровневую модель, оптимизируются независимо и никаких ограничений на их похожесть не накладывается, а разбиение признакового пространства на области действия моделей может не отражать реальной неоднородности в данных, построенная многоуровневая модель может быть не $(s,\ \alpha)$ – адекватной. Опишем далее способы построения адекватной многоуровневой модели по данной оптимальной и обученной.

Пусть задана некоторая функция сходства s. Обозначим $\mathbf{S} = \|s_{kl}(g_k(\mathbf{w}_k),\ g_l(\mathbf{w}_l))\|,\ k,\ l\ =\ \overline{1,\ K}$ матрицу значений попарных сходств моделей, входящих в многоуровневую модель, а $\mathbf{T} = \|t_{kl}(g_k(\mathbf{w}_k), g_l(\mathbf{w}_l))\|, k, l = \overline{1,\ K}$ матрицу соответствующих достигаемых уровней значимости в условиях истинности гипотезы о совпадении моделей, то есть $t_{kl} = \mathbb{P}(s(g_k(\mathbf{w}_k),\ g_l(\mathbf{w}_l)) < s_{kl}|\mathbf{w}_k = \mathbf{w}_l)$.

Отметим, что случайность здесь происходит из того, что g_k и g_l есть апостериорные распределения на $\mathbf{w}_k$, $\mathbf{w}_l$ соответственно, полученные по выборке. Так как $\mathbf{y}$ есть случайные вектор, то и g_k, g_l случайны. Так $g_k(\mathbf{w}_k) = N(\mathbf{w}_k|\mathbf{w}_k^*,\ \mathbf{\Sigma}_k)$, $g_l(\mathbf{w}_l) = N(\mathbf{w}_l|\mathbf{w}_l^*,\ \mathbf{\Sigma}_l)$, при этом $\mathbf{w}_k^*$ и $\mathbf{w}_l^*$ случайные векторы, имеющие некоторое распределение, а $\mathbf{\Sigma}_k$, $\mathbf{\Sigma}_l$ есть случайный матрицы, также имеющие некоторое распределение. Конкретный вид распределения значений функции сходства в условиях истинности гипотезы о совпадении моделей, то есть $F_s(x) = \mathbb{P}(s(g_k,\ g_l) < x|\mathbf{w}_k = \mathbf{w}_l)$, по которому рассчитываются уровни значимости t_{kl}, зависит от функции сходства s и считается вычисленным отдельно. Для предлагаемой функции сходства s-score это распределение дается теоремой **??**.

Далее рассматриваем матрицу достигаемых уровней значимости $\mathbf{T}$ и предложим несколько методов построения $(s,\ \alpha)$ – адекватной многоуровневой модели. Отметим, что если все достигаемые уровни значимости в матрице $\mathbf{T} = \|t_{kl}\|,\ k,\ l\ =\ \overline{1,\ K}$ не превосходят α, то есть $\forall\ k,\ l,\ k \neq l\ t_{kl} \leq \alpha$, то исходная обученная оптимальная многоуровневая модель уже является $(s,\ \alpha)$ – адекватной. Пусть далее это не так, то есть $\exists\ k,\ l,\ k \neq l\ t_{kl} > \alpha$. Рассмотрим несколько методов объединения моделей для построения $(s,\ \alpha)$ – адекватной мультимодели.

Метод последовательного парного объединения по наибольшему сходству. Этот метод основан на поиске двух наиболее близких друг к другу моделей, объединении их в одну оптимальную и вновь обученную. Затем производится пересчет элементов матрицы $\mathbf{T}$, соответствующих сходству объединенной модели с остальными. Итерации продолжаются до тех пор, пока $\exists\ k,\ l,\ k \neq l\ t_{kl} \geq \alpha$. Такая идея приводит к следующему алгоритму.

1. Находим $[k^*,\ l^*] = \arg\max_{k<l} t_{kl}$
2. Если $t_{k^*l^*} < \alpha$, останавливаемся. Построенная на данном шаге модель является $(s,\ \alpha)$ – адекватной. Иначе переходим на шаг 3.
3. Объединяем модели с номерами k^*, l^* и производим оптимизацию и обучение полученной модели, а также пересчет апостериорного распределения

на вектор параметров объединенной модели.

$$\mathcal{I}_{k^*} \sqcup \mathcal{I}_{l^*} \to \mathcal{I}_{k^*},$$

$$\mathbf{A}^*_{k^*} = \arg\max_{\mathbf{A}_{k^*}} p(\mathbf{y}_{\mathcal{I}_{k^*}} | \mathbf{X}_{\mathcal{I}_{k^*}},\ \mathbf{A}_{k^*}),$$

$$\mathbf{w}^*_{k^*} = \arg\max_{\mathbf{w}_{k^*}} p(\mathbf{y}_{\mathcal{I}_{k^*}},\ \mathbf{w}_{k^*} | \mathbf{X}_{\mathcal{I}_{k^*}},\ \mathbf{A}^*_{k^*}),$$

$$\mathbf{\Sigma}^*_{k^*} = \left(\mathbf{X}^\intercal_{\mathcal{I}_{k^*}} \mathbf{R}_{k^*} \mathbf{X}_{\mathcal{I}_{k^*}} + \mathbf{A}^*_{k^*}\right)^{-1}, \text{где} \mathbf{R}_{k^*} = \operatorname{diag}(\sigma(\mathbf{w}^{*\intercal}_{k^*}\mathbf{x}_i)\sigma(-\mathbf{w}^{*\intercal}_{k^*}\mathbf{x}_i), i \in \mathcal{I}_{k^*}),$$

$$g_{k^*}(\mathbf{w}_{k^*}) = N(\mathbf{w}_{k^*} | \mathbf{w}^*_{k^*},\ \mathbf{\Sigma}^*_{k^*})$$

4. Удаляем l^*-й столбец матриц $\mathbf{S}$ и $\mathbf{T}$, так как моделей стало на одну меньше, и пересчитываем сходства s_{k^*l} и соответствующие им достигаемые уровни значимости t_{k^*l} для $l \neq k^*$.

$$s_{k^*l} = s(g_{k^*}(\mathbf{w}_{k^*}),\ g_l(\mathbf{w}_l)),$$

$$t_{k^*l} = \mathbb{P}(s(g_{k^*}(\mathbf{w}_{k^*}),\ g_l(\mathbf{w}_l)) < s_{k^*l} | \mathbf{w}_{k^*} = \mathbf{w}_l).$$

5. Переходим на шаг 1.

Метод последовательного объединения максимальных клик по наибольшему сходству. Этот метод основан на последовательном поиске наибольшего по числу моделей набора моделей такого, что все модели внутри набора являются статистически неразличимыми на уровне значимости α. Если имеется несколько наборов одинакового размера, выбирается тот, у которого сумма элементов подматрицы матрицы $\mathbf{T}$, соответствующей этому набору, минимальна. Если таких наборов несколько, выбирается произвольный. Затем модели внутри найденного набора объединяются в одну, для нее производится оптимизация и обучение. Такая идея приводит к следующему алгоритму.

1. Находим все клики попарно неразличимых моделей максимального размера
$$\tilde{\mathcal{K}} = \operatorname*{Arg\,max}_{\mathcal{K} \in 2^{\{1,\ldots,K\}}} |\mathcal{K}| \min_{k,\, l \in \mathcal{K}} [t_{kl} \geq \alpha].$$

2. Если для $\mathcal{K} \in \tilde{\mathcal{K}}$ $\min_{k,\, l \in \mathcal{K}} [t_{kl} \geq \alpha] = 0$, то есть в клике есть различимые модели, останавливаемся, поскольку построена $(s,\ \alpha)$ – адекватная многоуровневая модель. Иначе переходим на шаг 3.

3. Среди найденных клик находим клику с максимальной суммой достигаемых уровней значимости
$$\mathcal{K}^* = \arg\max_{\mathcal{K} \in \tilde{\mathcal{K}}} \sum_{k,\, l \in \mathcal{K}} t_{kl}.$$

4. Объединяем модели с индексами из $\mathcal{K}^*$ в одну и производим оптимизацию и обучение полученной модели, а также пересчет апостериорного распределения на вектор параметров объединенной модели. Пусть $k^* \in \mathcal{K}^*$.

$$\sqcup_{k \in \mathcal{K}^*} \mathcal{I}_k \to \mathcal{I}_{k^*},$$

$$\mathbf{A}^*_{k^*} = \arg\max_{\mathbf{A}_{k^*}} p(\mathbf{y}_{\mathcal{I}_{k^*}} | \mathbf{X}_{\mathcal{I}_{k^*}},\ \mathbf{A}_{k^*}),$$

$$\mathbf{w}^*_{k^*} = \arg\max_{\mathbf{w}_{k^*}} p(\mathbf{y}_{\mathcal{I}_{k^*}},\ \mathbf{w}_{k^*} | \mathbf{X}_{\mathcal{I}_{k^*}},\ \mathbf{A}^*_{k^*}),$$

$$\mathbf{\Sigma}^*_{k^*} = \left(\mathbf{X}^\intercal_{\mathcal{I}_{k^*}} \mathbf{R}_{k^*} \mathbf{X}_{\mathcal{I}_{k^*}} + \mathbf{A}^*_{k^*}\right)^{-1}, \text{где} \mathbf{R}_{k^*} = \operatorname{diag}(\sigma(\mathbf{w}^{*\intercal}_{k^*} \mathbf{x}_i)\sigma(-\mathbf{w}^{*\intercal}_{k^*} \mathbf{x}_i), i \in \mathcal{I}_{k^*}),$$

$$g_{k^*}(\mathbf{w}_{k^*}) = N(\mathbf{w}_{k^*} | \mathbf{w}^*_{k^*},\ \mathbf{\Sigma}^*_{k^*}).$$

5. Удаляем столбцы матриц $\mathbf{S}$ и $\mathbf{T}$ с номерами из $\mathcal{K}^* \setminus \{k^*\}$, так как моделей стало меньше, и пересчитываем сходства s_{k^*l} и соответствующие им достигаемые уровни значимости t_{k^*l} для $l \neq k^*$.

$$s_{k^*l} = s(g_{k^*}(\mathbf{w}_{k^*}),\ g_l(\mathbf{w}_l)),$$

$$t_{k^*l} = \mathbb{P}(s(g_{k^*}(\mathbf{w}_{k^*}),\ g_l(\mathbf{w}_l)) < s_{k^*l} | \mathbf{w}_{k^*} = \mathbf{w}_l).$$

6. Переходим на шаг 1.

Алгоритмы построения $(s,\ \alpha)$ – адекватных смесей моделей

Пусть имеется оптимальная обученная смесь моделей, заданная совместным правдоподобием

$$p(\mathbf{y},\ \boldsymbol{\pi},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K | \mathbf{X},\ \alpha,\ \mathbf{A}^*_1,\ \ldots,\ \mathbf{A}^*_K) =$$
$$\frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \prod_{k=1}^K \pi_k^{\alpha-1} \prod_{k=1}^K \frac{\sqrt{\det \mathbf{A}^*_l}}{(2\pi)^{n/2}} \exp\left(-\tfrac{1}{2}\mathbf{w}_l^\intercal \mathbf{A}^*_l \mathbf{w}_l\right) \prod_{i=1}^m \left(\sum_{l=1}^K \pi_l \sigma(y_i \mathbf{w}_l^\intercal \mathbf{x}_i)\right).$$

Пусть также $\boldsymbol{\pi}$, $\mathbf{w}^*_1$, $\ldots$, $\mathbf{w}^*_K$ есть оценки максимума апостериорной вероятности на вектор весов моделей, входящих в смесь и на векторы их параметров, полученные в результате обучения (**??**), а $\mathbf{A}^*_1$, $\ldots$, $\mathbf{A}^*_K$ есть оценки максимума обоснованности для ковариационных матриц (**??**). Обозначим $g_k(\mathbf{w}_k)$, $k = \overline{1,\ K}$ апостериорное распределение на вектор параметров модели k.

Несмотря на то, что в процессе обучения мультимодели происходит ее прореживание, то есть исключение избыточных моделей, ограничений на похожесть моделей в полученной смеси не накладывается, а потому смесь моделей может быть не $(s,\ \alpha)$ – адекватной. Опишем далее способы построения адекватной смеси моделей по данной оптимальной и обученной.

Пусть задана некоторая функция сходства s. Обозначим
$\mathbf{S} = \|s_{kl}(g_k(\mathbf{w}_k),\ g_l(\mathbf{w}_l))\|,\ k,\ l\ =\ \overline{1,\ K}$ матрицу значений попарных сходств

моделей, входящих в смесь моделей, а $\mathbf{T} = \|t_{kl}(g_k(\mathbf{w}_k), g_l(\mathbf{w}_l))\|, k, l = \overline{1,\ K}$ матрицу соответствующих достигаемых уровней значимости в условиях истинности гипотезы о совпадении моделей, то есть $t_{kl} = \mathbb{P}(s(g_k(\mathbf{w}_k), g_l(\mathbf{w}_l)) < s_{kl} | \mathbf{w}_k = \mathbf{w}_l)$.

Отметим, что случайность здесь происходит из того, что g_k и g_l есть апостериорные распределения на $\mathbf{w}_k$, $\mathbf{w}_l$ соответственно, полученные по выборке. Так как $\mathbf{y}$ есть случайные вектор, то и g_k, g_l случайны. Конкретный вид распределения значений функции сходства в условиях истинности гипотезы о совпадении моделей, то есть $F_s(x) = \mathbb{P}(s(g_k, g_l) < x | \mathbf{w}_k = \mathbf{w}_l)$, по которому рассчитываются уровни значимости t_{kl}, зависит от функции сходства s и считается вычисленным отдельно. Для предлагаемой функции сходства s-score для двух моделей из смеси это распределение дается теоремой **??**.

Далее рассматриваем матрицу достигаемых уровней значимости $\mathbf{T}$ и предложим несколько методов построения $(s,\ \alpha)$ – адекватной смеси моделей. Отметим, что если все достигаемые уровни значимости в матрице $\mathbf{T} = \|t_{kl}\|,\ k,\ l = \overline{1,\ K}$ не превосходят α, то есть $\forall\ k,\ l,\ k \neq l\ t_{kl} \leq \alpha$, то исходная обученная оптимальная смесь моделей уже является $(s,\ \alpha)$ – адекватной. Пусть далее это не так, то есть $\exists\, k,\ l,\ k \neq l\, t_{kl} > \alpha$. Рассмотрим несколько методов объединения моделей для построения $(s,\ \alpha)$ – адекватной мультимодели.

Метод последовательного парного объединения по наибольшему сходству. Этот метод основан на поиске двух наиболее близких друг к другу моделей, объединении их в одну оптимальную и вновь обученную. Затем производится пересчет элементов матрицы $\mathbf{T}$, соответствующих сходству объединенной модели с остальными. Итерации продолжаются до тех пор, пока $\exists\, k,\ l,\ k \neq l\ t_{kl} \geq \alpha$. Такая идея приводит к следующему алгоритму.

1. Находим $[k^*,\ l^*] = \arg\max_{k<l} t_{kl}$
2. Если $t_{k^*l^*} < \alpha$, останавливаемся. Построенная на данном шаге модель является $(s,\ \alpha)$ – адекватной. Иначе переходим на шаг 3.
3. Объединяем модели с номерами k^*, l^* и производим перенастройку смеси моделей в соответствии с алгоритмом совместного обучения и оптимизации (3.4.). Исключаем модель с номером l^*. В качестве начального приближения для $\boldsymbol{\pi}$, $\mathbf{w}_1$, …, $\mathbf{w}_k$, …, $\mathbf{w}_K$ берем следующие

$$\pi_{k^*} + \pi_{l^*} \to \pi_{k^*},\ 0 \to \pi_{l^*},\ \tfrac{\mathbf{w}_{k^*} + \mathbf{w}_{l^*}}{2} \to \mathbf{w}_{k^*},\ \mathbf{w}_k \to \mathbf{w}_k,\ k \neq k^*,\ l^*.$$

4. Удаляем l^*-й столбец матриц $\mathbf{S}$ и $\mathbf{T}$, так как моделей стало на одну меньше, и пересчитываем сходства s_{k^*l} и соответствующие им достигаемые уровни значимости t_{k^*l} для $l \neq k^*$.

$$s_{k^*l} = s(g_{k^*}(\mathbf{w}_{k^*}),\ g_l(\mathbf{w}_l)),$$

$$t_{k^*l} = \mathbb{P}(s(g_{k^*}(\mathbf{w}_{k^*}),\ g_l(\mathbf{w}_l)) < s_{k^*l} | \mathbf{w}_{k^*} = \mathbf{w}_l).$$

5. Переходим на шаг 1.

Метод последовательного объединения максимальных клик по наибольшему сходству. Этот метод основан на последовательном поиске наибольшего по числу моделей набора моделей такого, что все модели внутри набора являются статистически неразличимыми на уровне значимости α. Если имеется несколько наборов одинакового размера, выбирается тот, у которого сумма элементов подматрицы матрицы $\mathbf{T}$, соответствующей этому набору, минимальна. Если таких наборов несколько, выбирается произвольный. Затем модели внутри найденного набора объединяются в одну, для нее производится оптимизация и обучение. Такая идея приводит к следующему алгоритму.

1. Находим все клики попарно неразличимых моделей максимального размера
$$\tilde{\mathcal{K}} = \operatorname*{Arg\,max}_{\mathcal{K} \in 2^{\{1,\ldots,K\}}} |\mathcal{K}| \min_{k,\, l \in \mathcal{K}} [t_{kl} \geq \alpha].$$
2. Если для $\mathcal{K} \in \tilde{\mathcal{K}}$ $\min_{k,\, l \in \mathcal{K}} [t_{kl} \geq \alpha] = 0$, то есть в клике есть различимые модели, останавливаемся, поскольку построена $(s,\, \alpha)$ – адекватная смесь моделей. Иначе переходим на шаг 3.
3. Среди найденных клик находим клику с максимальной суммой достигаемых уровней значимости
$$\mathcal{K}^* = \arg\max_{\mathcal{K} \in \tilde{\mathcal{K}}} \sum_{k,\, l \in \mathcal{K}} t_{kl}.$$
4. Объединяем модели с индексами из $\mathcal{K}^*$ в одну и производим перенастройку смеси моделей в соответствии с алгоритмом совместного обучения и оптимизации (3.4.). Пусть $k^* \in \mathcal{K}^*$. Исключаем модели с номерами $l \in \mathcal{K}^* \setminus \{k^*\}$. В качестве начального приближения для $\boldsymbol{\pi}$, $\mathbf{w}_1, \ldots, \mathbf{w}_k, \ldots, \mathbf{w}_K$ берем следующие
$$\sum_{l \in \mathcal{K}^*} \pi_l \to \pi_{k^*},\ 0 \to \pi_l,\ l \in \mathcal{K} \setminus \{k^*\},\ \tfrac{1}{|\mathcal{K}^*|} \sum_{l \in \mathcal{K}^*} \mathbf{w}_l \to \mathbf{w}_{k^*},\ \mathbf{w}_k \to \mathbf{w}_k,\ k \notin \mathcal{K}^*.$$
5. Удаляем столбцы матриц $\mathbf{S}$ и $\mathbf{T}$ с номерами из $\mathcal{K}^* \setminus \{k^*\}$, так как моделей стало меньше, и пересчитываем сходства s_{k^*l} и соответствующие им достигаемые уровни значимости t_{k^*l} для $l \neq k^*$.
$$s_{k^*l} = s(g_{k^*}(\mathbf{w}_{k^*}),\ g_l(\mathbf{w}_l)),$$
$$t_{k^*l} = \mathbb{P}(s(g_{k^*}(\mathbf{w}_{k^*}),\ g_l(\mathbf{w}_l)) < s_{k^*l} | \mathbf{w}_{k^*} = \mathbf{w}_l).$$
6. Переходим на шаг 1.

Глава 5
Анализ прикладных задач

5.1. Иллюстрация вырожденности недиагональной оценки максимума обоснованности ковариационной матрицы параметров логистической модели

Проиллюстрируем результат теоремы об асимптотической вырожденности недиагональной оценки максимума обоснованности для ковариционной матрицы. Рассматриваем случай одной модели $K = 1$, признакового пространства размерности $n = 2$. В качестве истинного вектора параметров рассматриваем два случая $\mathbf{w}_1 = [1,\ 1]^\intercal$ и $\mathbf{w}_2 = [1,\ -1]^\intercal$. Варьируем число объектов в выборке, сгенерированных в соответствии с моделью логистической регрессии от 50 до 1000000 и оцениваем недиагональную ковариационную матрицу методом максимума обоснованности в соответствии с (2.5) с помощью аппроксимации Лапласа. Сэмплируем признаки $\mathbf{f}_1$, $\mathbf{f}_2$ независимо поэлементно из $N(0,\ 1)$. Пусть $\mathbf{X} = [\mathbf{f}_1,\ \mathbf{f}_2]$. Результаты эксперимента для случая некоррелированных признаков $\mathbf{f}_1$, $\mathbf{f}_2$ приведены в табл. 5.1, 5.2.

Сохраняя обозначения теоремы, имеем $\mathbf{A}^* = \begin{pmatrix} \sigma_1^{*2} & \kappa^*\sigma_1^*\sigma_2^* \\ \kappa^*\sigma_1^*\sigma_2^* & \sigma_2^{*2} \end{pmatrix}$. В обоих случаях, и когда истинные веса признаков имеют одинаковый знак, и когда истинные веса признаков имеют разный знак, при росте числа объектов наблюдается увеличение $\min(\sigma_1^*,\ \sigma_2^*)$ в согласии с теоремой. Более того, уже при $m = 10000$ в обоих случаях оцененная корреляция между веса признаков по модуля равна 1 с машинной точностью, а знак определяется как $-\text{sign}(w_1 w_2)$, что также находится в согласии с теоремой.

Рассмотрим теперь случай зависимых признаков. В качестве $\mathbf{X}$ возмем $\mathbf{X} = [\mathbf{f}_1,\ \frac{1}{\sqrt{2}}(\mathbf{f}_1 + \mathbf{f}_2)]$. Результаты эксперимента приведены в табл. 5.3, 5.4.

Для случая коррелированных признаков также наблюдаем, что при увеличении числа объектов растет $\min(\sigma_1^*, \sigma_2^*)$, а уже, начиная с $m = 100000$ для весов одного знака, и, начиная с $m = 10000$ для весов разных знаков, оценка максимума обоснованности для корреляции весов признаков с машинной точностью равна $-\text{sign}(w_1 w_2)$. Разная же скорость сходимости в двух рассматриваемых случаях объясняется коррелированностью признаков. Так было показано, что

Таблица 5.1: Иллюстрация вырожденности недиагональной оценки максимума обоснованности для ковариационной матрицы параметров логистической модели для случая параметров одного знака, $\mathbf{w} = \mathbf{w}_1 = [1,\ 1]^\intercal$.

Данные / m	50	100	1000	10^4	10^5	10^6
$\min(\sigma_1^*,\ \sigma_2^*)$	6.36	10.04	39.88	132.33	422.01	$1.35 \cdot 10^3$
κ^*	-0.9939	-0.998	-0.9998	-1	-1	-1

Таблица 5.2: Иллюстрация вырожденности недиагональной оценки максимума обоснованности для ковариационной матрицы параметров логистической модели для случая параметров разных знаков, $\mathbf{w} = \mathbf{w}_1 = [1, -1]^\intercal$.

Данные / m	50	100	1000	10^4	10^5	10^6
$\min(\sigma_1^*, \sigma_2^*)$	5.36	15.63	41.97	131.14	419.19	$1.34 \cdot 10^3$
κ^*	0.9765	0.9966	0.9997	1	1	1

Таблица 5.3: Иллюстрация вырожденности недиагональной оценки максимума обоснованности для ковариационной матрицы параметров логистической модели для случая параметров одного знака, $\mathbf{w} = \mathbf{w}_1 = [1, 1]^\intercal$ в случае коррелированных признаков.

Данные / m	50	100	1000	10^4	10^5	10^6
$\min(\sigma_1^*, \sigma_2^*)$	3.00	3.98	19.48	61.52	213.74	679.63
κ^*	-0.9683	-0.98	-0.9986	-0.9999	-1	-1

Таблица 5.4: Иллюстрация вырожденности недиагональной оценки максимума обоснованности для ковариационной матрицы параметров логистической модели для случая параметров разных знаков, $\mathbf{w} = \mathbf{w}_1 = [1, -1]^\intercal$ в случае коррелированных признаков.

Данные / m	50	100	1000	10^4	10^5	10^6
$\min(\sigma_1^*, \sigma_2^*)$	11.23	16.76	57.50	191.97	611.21	$1.93 \cdot 10^3$
κ^*	0.9990	0.9991	0.9999	1	1	1

теорема применима, если признаки не являются коллинеарными или вырожденными (см. границы применимости теоремы). Таким образом, в предельном случае, когда признаки идеально коррелированы, утверждаемых сходимостей не наблюдается, потому, видимо, имеет место разная скорость сходимости в промежуточных случаях.

5.2. Иллюстрация построения $(s,\ \alpha)$-адекватных многоуровневых моделей

Как уже указывалось, многоуровневые модели являются стандартным подходом, например, в кредитном скоринге [**?**, **?**]. Для учета неоднородностей данных выборку бьют на несколько совокупностей и для каждой из полученных усеченных выборок строят отдельную модель. При это разбиение производится либо с помощью кластеризации [**?**, 25, 26], либо путем деления по значениям некоторого признака [**?**], например, возраста или региона проживания заемщика. Отметим, что при этом нет ограничений на сходство моделей для разных подвыборок, а потому, так как кластеризация и разбиение на подвыборки осуществляется только на основании признакового описания, полученные модели для разных подвыборок могут быть статистически неразличимы. Кроме избыточности такой многоуровневой модели, она дает более низкое качество прогноза, поскольку для определения параметров неразличимых моделей вместо объединенной выборки используются ее части, что ведет к большей ошибке в определении параметров. Продемонстрируем далее этот эффект на синтетических и реальных данных.

Эксперимент с построением (s, α)-адекватных многоуровневых моделей на синтетических данных.

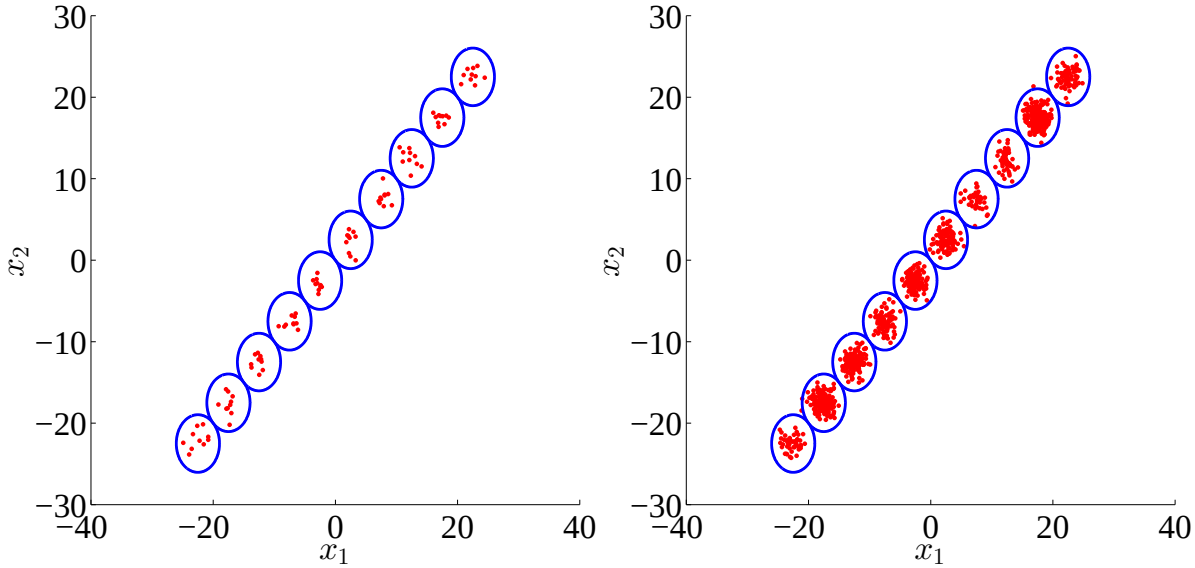
Рассмотрим признаковое пространство размерности $n = 2$. Генерируем объекты из $K = 10$ кластеров независимо следующим образом

$$\forall\, i \in \mathcal{I}_k\ \mathbf{x}_i \sim N([s_k,\ s_k]^\intercal,\ \mathbf{I}),$$

где $s_k = 2.5(2k-11)$ есть сдвиг k-го кластера, обеспечивающий расстояние, равное 5, между центрами кластеров. Размеры кластеров при этом определяются параметром δ_0, который указывает на отношение размера максимального кластера к размеру минимального. При $\delta = 1$ все кластеры генерируются одного размера. Иначе обозначим $\delta = \delta_0^{1/(K-1)}$ и генерируем случайную перестановку кластеров $\mathbf{p}$. Тогда для размеров кластеров имеем

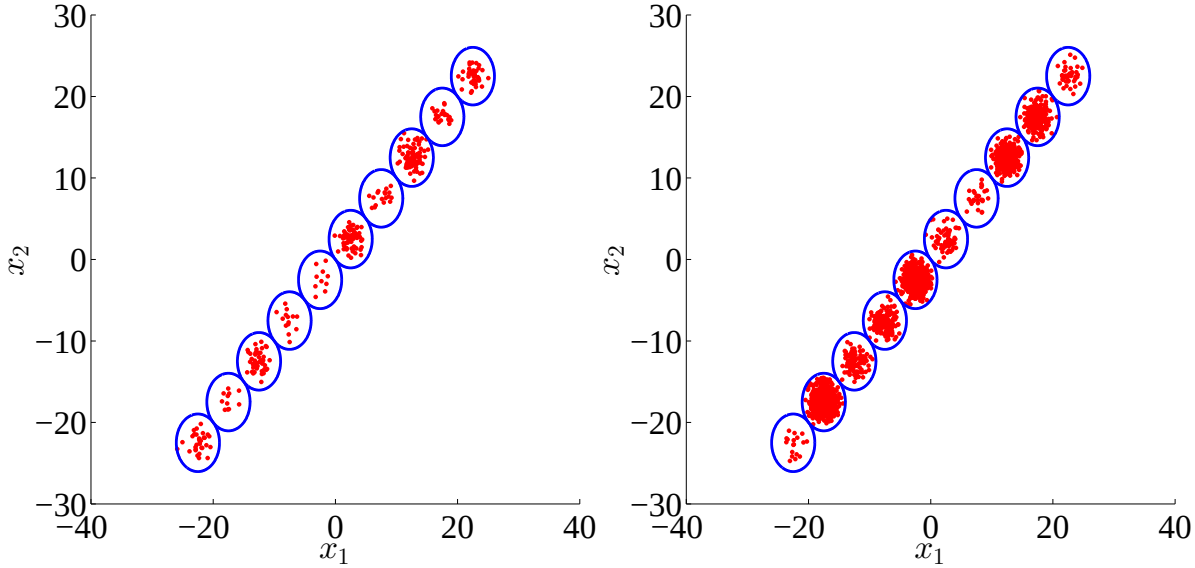
$$|\mathcal{I}_{p_k}| = \left\lfloor \max\left(10,\ \delta^{k-1}\frac{m(\delta-1)}{\delta^K - 1}\right)\right\rfloor.$$

Получающиеся выборки для разных значений m, δ_0 приведены на рис. 5.1. В зависимости от δ_0 в выборке либо имеются малые кластеры, для которых оценка параметров моделей будет наиболее неустойчивой, либо они отсутствуют.



(а) $m = 100$, $\delta_0 = 1$

(б) $m = 1000$, $\delta_0 = 5$



(в) $m = 350$, $\delta_0 = 10$

(г) $m = 2500$, $\delta_0 = 50$

Рис. 5.1: Примеры сгенерированной синтетической выборки с кластерной структурой для разных значений m и δ_0.

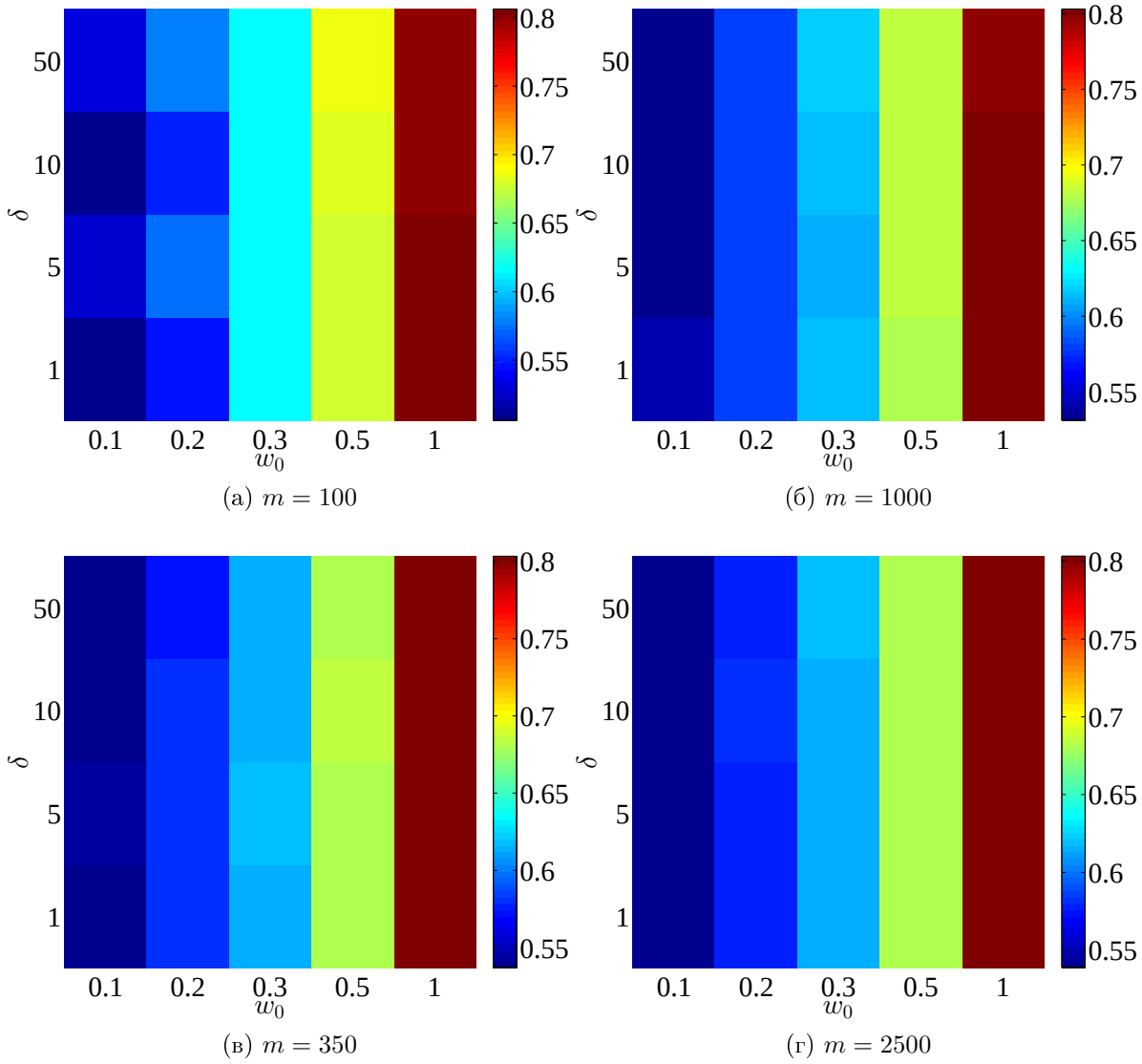
Проведем далее следующий эксперимент. Возьмем в качестве истинного вектора параметров $\mathbf{w} = w_0[1,\ 1]^\intercal$. Сгенерируем значения целевой переменной $\mathbf{y}$ в соответствии с моделью логистической регрессии (2.1). Отметим, что w_0 задает максимально достижимое качество для прогноза целевой переменной, поскольку при малом w_0 для каждого объекта вероятность принадлежать к каждому из классов близка к 0.5, а потому даже истинная гиперплоскость будет иметь близкое к 0.5 значение критерия качества AUC [48].

Построим на имеющихся данных сначала многоуровневую модель, состоящую из $K = 10$ моделей, каждую на своем кластере данных. Затем по полученной многоуровневой модели построим $(s,\ \alpha)$-адекватную мультимодель методом последовательного парного объединения по наибольшему сходству для предлагаемой функции сходства s-score на уровне значимости $\alpha = 0.05$. Генерацию и подсчет результатов произведем 10 раз. Полученные результаты осреднены и приведены на рис. 5.2.

На рис. 5.2 приведена зависимость средней AUC по кластерам на тестовой выборке для построенной $(s,\ \alpha)$ – адекватной многоуровневой модели от δ_0, отражающей степень различия кластеров по размеру, и w_0, определяющей максимально достижимое значение критерия качества AUC. Из рис. 5.2 заключаем, что для построенный $(s,\ \alpha)$ – адекватных многоуровневых моделей нет зависимости достигаемого качества от степени различия кластеров по размеру, что означает, что алгоритм построения $(s,\ \alpha)$ – адекватных многоуровневых моделей признает модели, построенные на малых кластерах неразличимыми с моделями для других кластеров, что приводит к их объединению. Действительно в эксперименте наблюдалось от 1 до 5 моделей в построенной $(s,\ \alpha)$ – адекватной мультимодели, причем в большинстве случаев все модели объединялись в одну. Отметим, однако, что для учета эффекта от множественного тестирования гипотез, при большом числе моделей K имеет смысл уменьшить уровень значимости α.

На рис. 5.3 приведем зависимость среднего значения минимально наблюдавшегося AUC по кластерам для построенной $(s,\ \alpha)$ – адекватной мультимодели при десяти повторениях эксперимента. За исключением нескольких выбросов для слабых моделей при малом w_0 при $m = 100$ зависимости качества предсказания для худшего по качеству кластера от степени различия размеров кластеров δ_0 не наблюдается. Рассмотрим теперь зависимость разности между средним качеством и средним худшим качеством по кластерам между построенной $(s,\ \alpha)$ – адекватной многоуровневой моделью и исходной многоуровневой моделью.

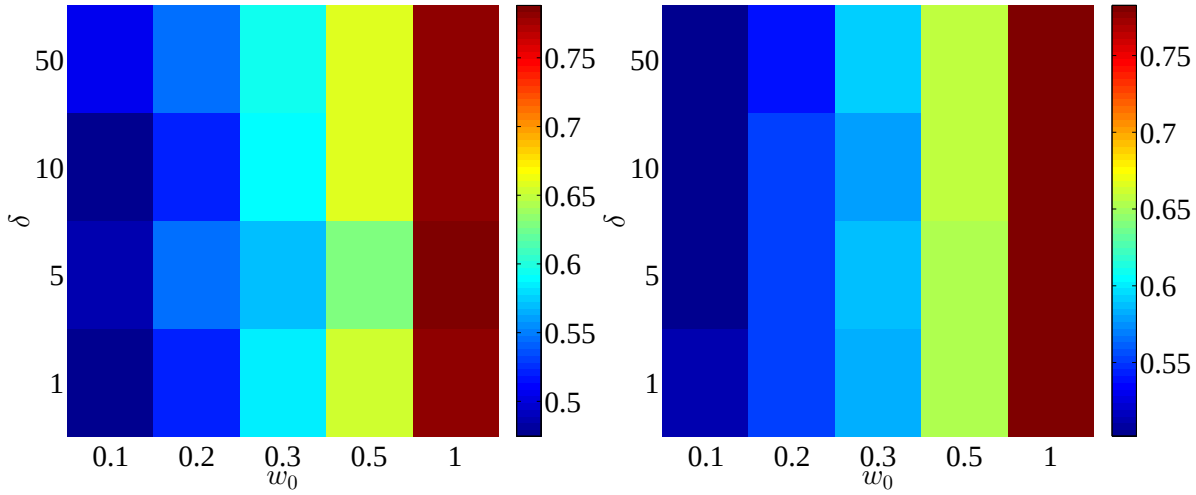
На рис. 5.4 и 5.5 приведена зависимость разности между средним качеством и минимальным по кластерам качеством для построенной $(s,\ \alpha)$ – адекватной многоуровневой моделью и исходной многоуровневой моделью соответственно. Из рис. **??** заключаем, что во все случаях построенная $(s,\ \alpha)$ – адекватная мультимодель оказывается не хуже исходной. Кроме того, во многих случаях она оказывается лучше по среднему и минимальному по кластерам каче-



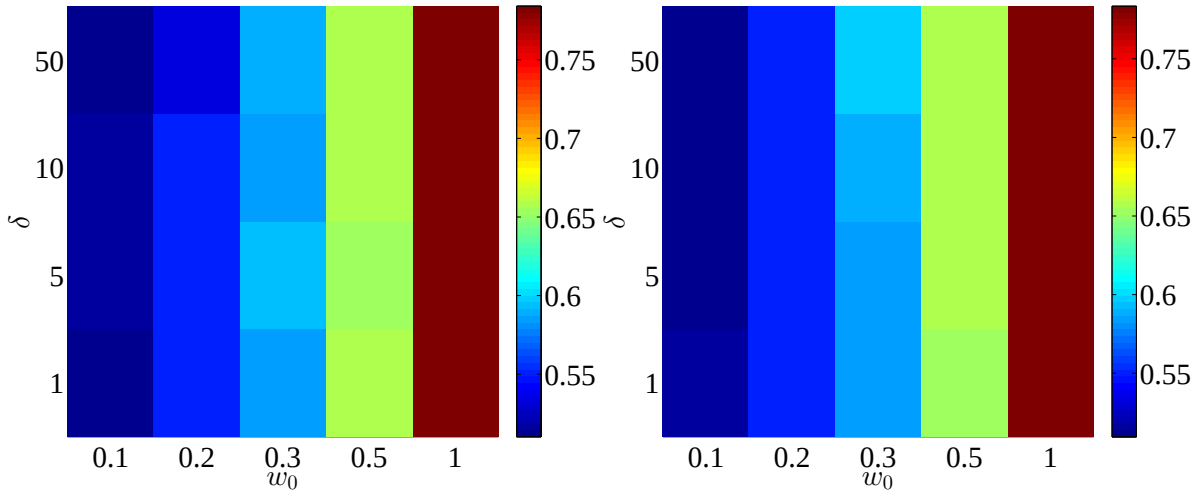
(а) $m = 100$ (б) $m = 1000$

(в) $m = 350$ (г) $m = 2500$

Рис. 5.2: Зависимость AUC от m, δ_0 и w_0 для построенной $(s,\ \alpha)$ – адекватной многоуровневой модели.

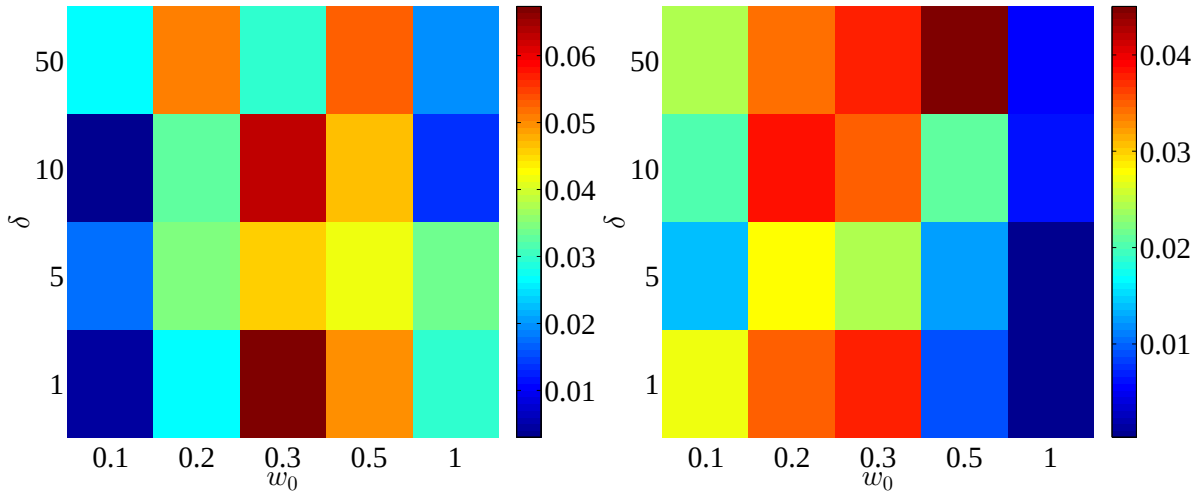


(а) $m = 100$ (б) $m = 1000$



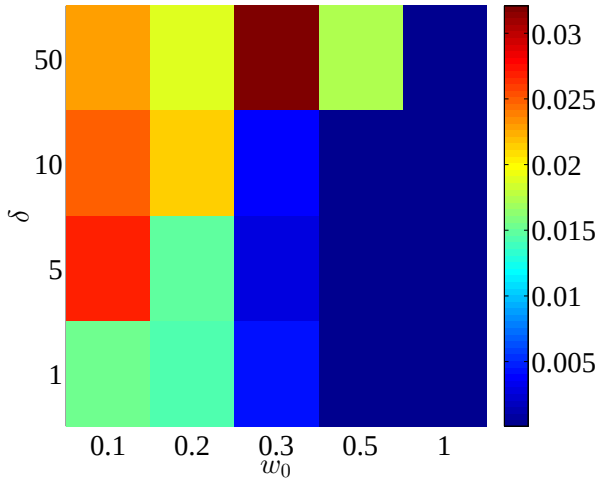
(в) $m = 350$ (г) $m = 2500$

Рис. 5.3: Зависимость минимального AUC по кластерам от m, δ_0 и w_0 для построенной $(s,\ \alpha)$ – адекватной многоуровневой модели.

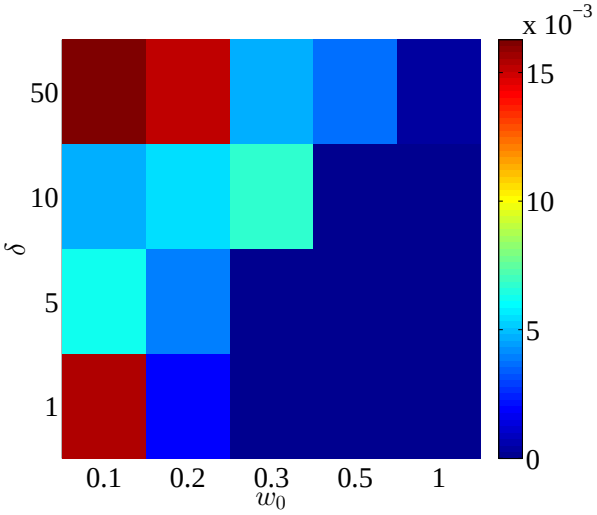


(а) $m = 100$

(б) $m = 1000$

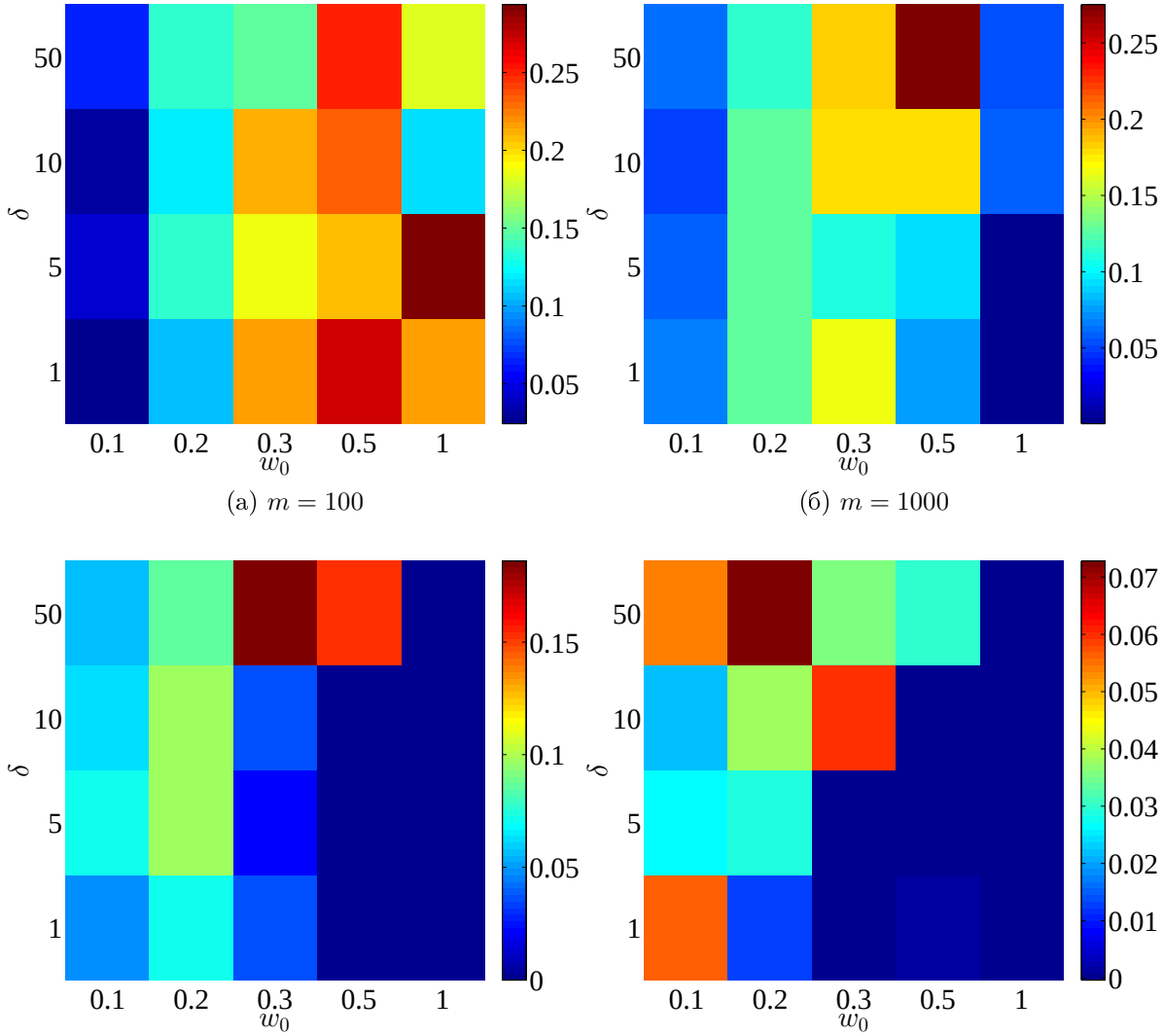


(в) $m = 350$



(г) $m = 2500$

Рис. 5.4: Зависимость разности AUC для построенной $(s,\ \alpha)$ – адекватной многоуровневой модели и исходной многоуровневой модели от m, δ_0 и w_0.



(а) $m = 100$ (б) $m = 1000$

(в) $m = 350$ (г) $m = 2500$

Рис. 5.5: Зависимость минимального AUC по кластерам для построенной (s, α) – адекватной многоуровневой модели и исходной многоуровневой модели от m, δ_0 и w_0.

Таблица 5.5: Сравнение исходной многоуровневой модели и построенной по ней $(s,\ \alpha)$ – адекватной многоуровневой модели для данных по немецким потребительским кредитам при $\alpha = 0.05$.

Данные/Группа	≤ 25	$(25, 45)$	≥ 45	$G_1 \cup G_2 \cup G_3$
AUC_{init}	0.7190	0.7658	0.8022	0.7663
AUC_{adeq}	**0.7230**	0.7655	**0.8052**	**0.7686**
t-статистика	12.16	-1.82	6.62	18.2
p-value	0	0.0344	0	0

ству по сравнению с исходной. Преимущество построенной $(s,\ \alpha)$ – адекватной многоуровневой модели тем выше, чем больше степень различия кластеров по размеру δ_0, чем слабее исходная модель, то есть чем меньше w_0, а также снижается при фиксированных δ_0 и w_0 при увеличении числа объектов в выборке. Отметим, однако, что кроме преимуществ, связанных с увеличением качества прогноза, полученная модель является адекватной, то есть содержит меньше моделей, чем исходная, и при этом все модели, в нее входящие, являются попарно статистически различимыми по построению, что повышает интерпретируемость.

Эксперимент с построением (s, α)-адекватных многоуровневых моделей на реальных данных по немецким потребительским кредитам. Протестируем теперь предлагаемый алгоритм построения $(s,\ \alpha)$ – адекватных многоуровневых моделей на реальных данных. В качестве выборки реальных данных используем выборку по немецким потребительским кредитам [43], состоящую из $m = 1000$ объектов. Для построения многоуровневой модели используем разбиение признака возраста $\mathbf{f}_a$ на три группы: $G_1 = \{(\mathbf{x},\ y) :\ x_a \leq 25\}$, $G_2 = \{(\mathbf{x},\ y) :\ x_a \in (25,\ 45)\}$, $G_3 = \{(\mathbf{x},\ y) :\ x_a \geq 45\}$ [**?**]. Размеры получившихся групп: 190, 605, 205 соответственно.

Сравним качество на кросс-валидации для исходной многоуровневой модели с приведенным разбиением на группы и построенной по ней $(s,\ \alpha)$ – адекватной многоуровневой модели, где в качестве функции сходства используется предлагаемая функция сходства, а уровень значимости равен $\alpha = 0.05$. Производилась генерация 5000 разбиений на обучение и контроль с размером обучения 0.7 от общего размера выборки с равномерным выбором объектов из каждой возрастной группы. Результаты сравнения приведены в табл. 5.5.

Из табл. 5.5 заключаем, что построение $(s,\ \alpha)$ – адекватной многоуровневой модели позволило увеличить качество примерно на 0.35% в терминах AUC для двух из трех групп без значимого ухудшения качества для оставшейся группы. Для объединения групп качество растет на 0.23%. Более того, указанное улучшение является статистически значимым с достигаемым уровнем значимости, равным 0 с точностью 10^{-10}. Рассматривая результаты построения (s, α) – адекватной модели, получаем, что в 36.3% случаях построение $(s,\ \alpha)$ – адекватной

Таблица 5.6: Сравнение исходной многоуровневой модели и построенной по ней $(s,\ \alpha)$ – адекватной многоуровневой модели для данных по немецким потребительским кредитам при $\alpha = 0.001$.

Данные/Группа	≤ 25	$(25, 45)$	≥ 45	$G_1 \cup G_2 \cup G_3$
AUC_{init}	0.7182	0.7661	0.8027	0.7664
AUC_{adeq}	**0.7285**	**0.7696**	**0.8146**	**0.7749**
t-статистика	24.30	16.93	20.43	45.39
p-value	0	0	0	0

многоуровневой модели производилось путем объединения групп объектов G_2 и G_3. В 37.5% случаев построенная исходная многоуровневая модель являлась $(s,\ \alpha)$ – адекватной на уровне значимости $\alpha = 0.05$. В 9.8% случаев построение $(s,\ \alpha)$ – адекватной многоуровневой модели производилось путем объединения всех групп. Отметим, что при получении этих результатов использовался достаточно высокий уровень ошибки первого рода $\alpha = 0.05$.

Рассмотрим теперь случай более низкой вероятности ошибки первого рода $\alpha = 0.001$, соответствующего требованию получить меньше моделей в мультимодели. Результаты сравнения приведены в табл. 5.6.

Из табл. 5.6 заключаем, что построение $(s,\ \alpha)$ – адекватной многоуровневой модели позволило увеличить качество значимо для всех трех групп, примерно на 1.1% в терминах AUC для двух из трех групп и на 0.35% для наиболее многочисленной группы. Для объединения групп качество растет на 0.85%. Более того, указанные улучшения являются статистически значимыми с достигаемым уровнем значимости, равным 10^{-63}. Рассматривая результаты построения $(s,\ \alpha)$ – адекватной модели, получаем, что в 61.3% случаях построение $(s,\ \alpha)$ – адекватной многоуровневой модели производилось путем объединения групп объектов G_2 и G_3. В 26.5% случаев построение $(s,\ \alpha)$ – адекватной многоуровневой модели производилось путем объединения всех групп. В 8% случаев построение $(s,\ \alpha)$ – адекватной многоуровневой модели производилось путем объединения групп объектов G_1 и G_2. В 3.6% случаев построенная исходная многоуровневая модель являлась $(s,\ \alpha)$ – адекватной на уровне значимости $\alpha = 0.001$. Отметим, что подобный результат показывает, что исходное разбиение на группы объектов для построения многоуровневой модели, видимо, не соответствует истинной зависимости между признаками и целевой переменной, а потому при построении $(s,\ \alpha)$ – адекватной мультимодели удалось улучшить качество.

Эксперимент с построением (s, α)-адекватных многоуровневых моделей на реальных данных по качеству белого вина. Протестируем теперь предлагаемый алгоритм построения $(s,\ \alpha)$ – адекватных многоуровневых моделей на реальных данных. В качестве выборки реальных данных используем выборку по качеству белого вина [47], состоящую из $m = 4898$ объектов.

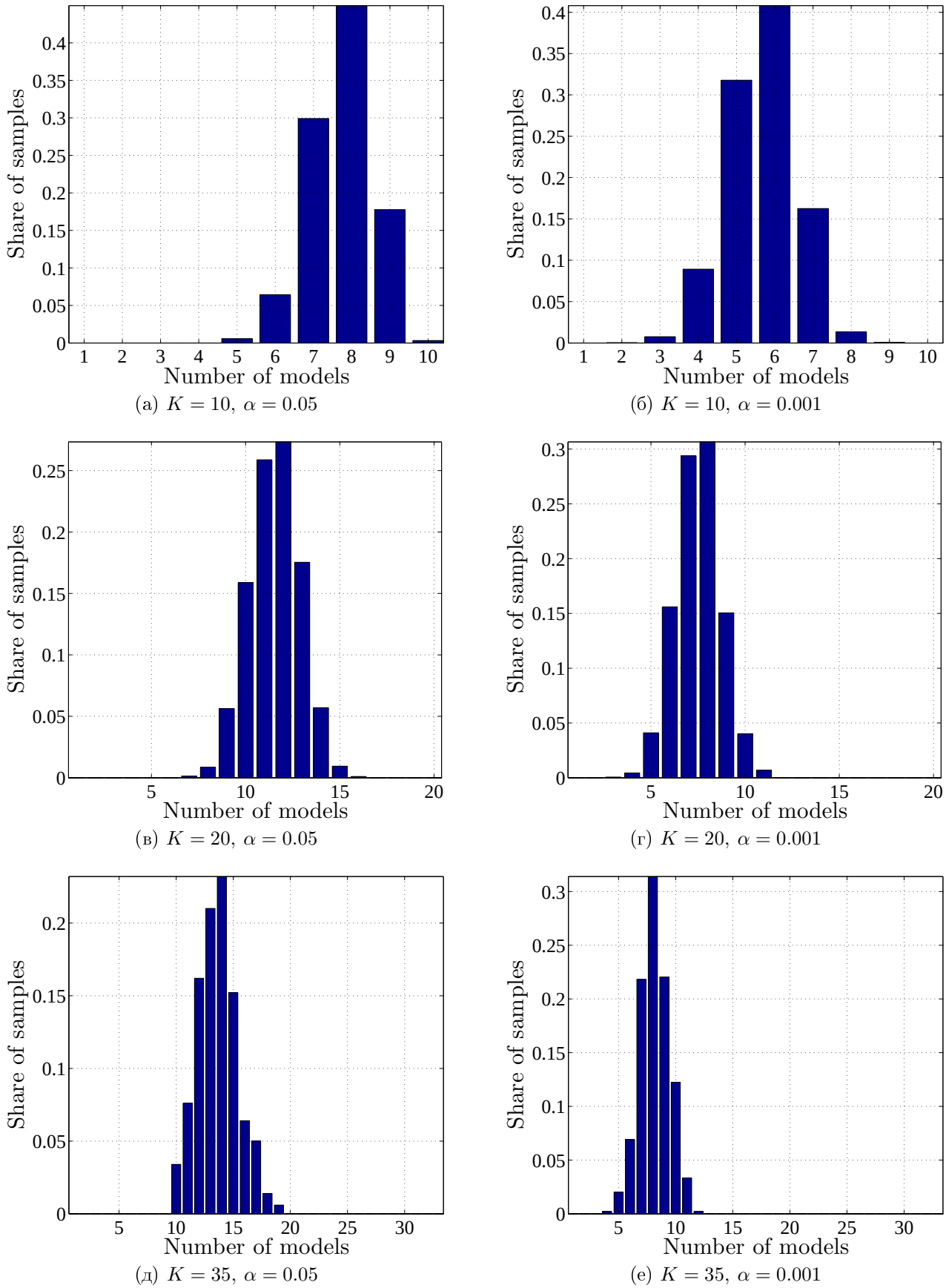
Таблица 5.7: Сравнение построенной (s, α) – адекватной многоуровневой модели с исходной и с одиночной логистической моделью.

Данные / $(K,\ \alpha)$	(10, 0.05)	(10, 0.001)	(20, 0.05)	(20, 0.001)	(35, 0.05)	(35, 0.0
AUC_{adeq}	0.8149	0.8139	0.8053	**0.8069**	**0.8028**	**0.80**
AUC_{init}	**0.8155**	**0.8155**	0.8054	0.8058	0.7992	0.799
AUC_s	0.8060	0.8061	0.8029	0.8032	0.8011	0.800
t_{init}	-20.57	-34.74	-1.13	15.75	42.52	53.6
t_s	106.29	98.60	24.43	41.60	14.42	33.7
$f_{init}, \%$	45.3	30.6	49.3	59.2	73.0	77.4
$f_s, \%$	92.6	91.7	64.3	72.4	59.2	71.0
$\overline{K}$	7.4	5.4	11.2	7.3	13.8	8.0

Для построения многоуровневой модели используем разбиение признака содержания свободного диоксида серы на $K = 10$ и $K = 20$ групп по квантилям распределения признака с одинаковым шагом, что дает группы приблизительно одинакового размера.

Будем производить 5000 генераций случайного разбиения выборки на обучение и тест, как и в для данных по немецким потребительским кредитам. Выбор объектов для обучения, как и ранее, равномерен по группам. Результаты сравнения предлагаемого метода построения $(s,\ \alpha)$ – адекватной многоуровневой модели с исходной многоуровневой моделью и с моделью, которая комбинирует все группы в одну приведены в табл. **??**. В табл. 5.7 AUC_{adeq}, AUC_{init}, AUC_s означают соответственно средние значения AUC по итерациям на всей тестовой выборке для построенной $(s,\ \alpha)$ – адекватной многоуровневой модели, исходной многоуровневой модели и одиночной логистической модели, а t_{init} и t_s значения t-статистики для гипотезы сравнения построенной $(s,\ \alpha)$ – адекватной многоуровневой модели против исходной и одиночной соответственно. f_{init} и f_s есть доля разбиени, для которых построенная многоуровневая модель оказалась лучше на тестовой выборке, чем исходная и одиночная соответственно. $\overline{K}$ показывает среднее число моделей в построенной $(s,\ \alpha)$ – адекватной многоуровневой модели. На рис. 5.6 приведены гистограммы распределения числа моделей в построенной оптимальной $(s,\ \alpha)$ – адекватной многоуровневой модели. Из табл. 5.7 заключаем, что большее качество достигается при α, для которого числом моделей в многоуровневой модели $K = 8$. Заметим, что результаты для единственной модели несколько отличаются при разных K. Это объясняется тем, что обучающая выборка на кросс-валидации выбирается равномерно по K группам, а потому при наличии неоднородности по признаку имеет статистически разный состав для разных K.

Отметим, что при $K = 10,\ 20$ и исходная многоуровневая модель, и построенная по ней $(s,\ \alpha)$ – адекватная работают лучше, чем одиночная логистическая модель, что указывает на наличие значимой неоднородности по признаку



(а) $K = 10$, $\alpha = 0.05$

(б) $K = 10$, $\alpha = 0.001$

(в) $K = 20$, $\alpha = 0.05$

(г) $K = 20$, $\alpha = 0.001$

(д) $K = 35$, $\alpha = 0.05$

(е) $K = 35$, $\alpha = 0.001$

Рис. 5.6: Гистограммы распределения числа моделей в построенной $(s,\ \alpha)$ – адекватной многоуровневой модели для разных значений K и α.

содержания свободного диоксида серы. Кроме того, при $K = 10$ построенная (s, α) – адекватная многоуровневая модель незначительно уступает исходной, в то время как при $K = 20$ и $K = 35$ наблюдается обратная ситуация. Превышение качества у исходной модели над построенной для $K = 10$ объясняется тем, что если модели в многоуровневой модели исходно разные, существует вероятность признать их одинаковыми, дающаяся вероятностью ошибки второго рода, что ухудшает построенную модель. При $K = 20$ и $K = 35$ наблюдается противоположная ситуация, поскольку количество моделей в исходной многоуровневой модели, видимо, значительно превышает требуемое для учета неоднородности по признаку. Таким образом, получаем, что при наличии неоднородности в данных построение (s, α) – адекватных многоуровневых моделей позволяет сократить число моделей в многоуровневой модели, что повышает интерпретируемости мультимодели, и одновременно значимо повысить качество прогноза. При отсутствии неоднородности наблюдается незначительное ухудшение качества прогноза из-за принятия одинаковых моделей за разные.

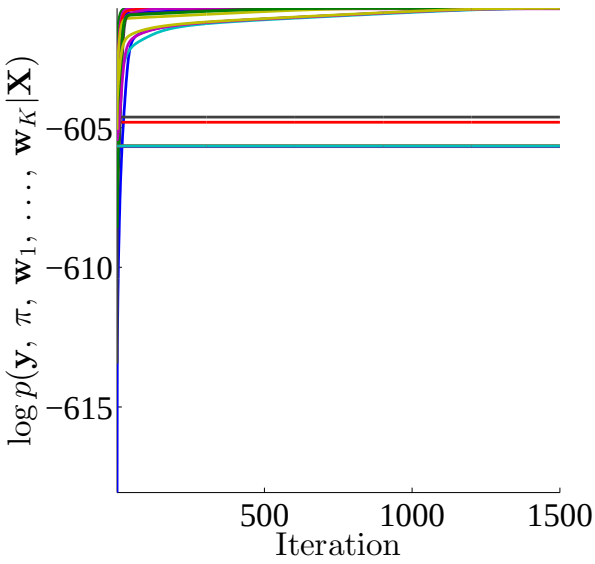
5.3. Иллюстрация построения (s, α)-адекватных смесей моделей

Приведем далее результаты экспериментов по построению (s, α) – адекватных смесей моделей на синтетических и реальных данных. Отметим, что число моделей в построенной смеси в отличие от многоуровневой модели, где это число задано заранее, для смеси моделейзадано лишь максимальное число моделей в смеси, а в силу прореживания смеси итоговая мультимодель может иметь меньше моделей. По этой причине для любого исходного числа моделей K возможно построить (s, α) – адекватную мультимодель для любого уровн значимости α. Для этого достаточно выбрать достаточно сильно прореживающее априорное распределение $p(\boldsymbol{\pi}|\alpha)$, что в итоге в смеси останется только одна модель, которая по определению является (s, α) – адекватной. Приведем далее результаты экспериментов на синтетических данных, показывающие, что обученные смеси моделей могут быть не (s, α) – адекватными.

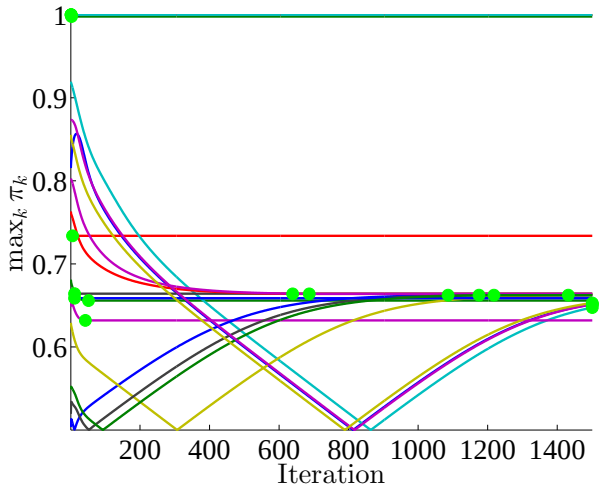
Иллюстрация многоэкстремальности и возможной неадекватности обученной смеси моделей. Задача обучения смеси моделей состоит в одновременном поиске весов моделей $\boldsymbol{\pi}$ и векторов их параметров $\mathbf{w}_1, \ldots, \mathbf{w}_K$ и является невыпуклой и имеет множество локальных экстремумов. По этой причине даже, если глобальный минимум обладает свойством адекватности, найденный локальный минимум может не быть адекватным.

Покажем сначала, что задача обучения смеси моделей является многоэкстремальной. Для этого сгенерируем выборку размера $m = 1000$ из признакового пространства размерности $n = 2$ из смеси $K = 2$ моделей с весами $\boldsymbol{\pi} = [0.6, 0.4]^\intercal$. Для восстановления смеси используем $K = 2$, $\alpha = 1$.

На рис. 5.7 приведены графики зависимости логарифма совместного правдоподобия и максимальной оценки вероятности модели $\max_k \pi_k$ для 20 запусков из разных начальных точек на протяжении 1500 итераций EM-алгоритма. Зелены-



(а) Зависимость совместного павдоподобия $\log p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi}|\mathbf{X})$ от номера итерации.



(б) Зависимость $\max_k \pi_k$ от номера итерации.

Рис. 5.7: Иллюстрация многоэкстремальности совместного правдоподобия смеси моделей.

ми точки на рис. 5.7б обозначен момент наступления сходимости для каждого из запусков. Отметим, что наряду с обученной моделью с наибольшим совместным правдоподобием, дающейся $\boldsymbol{\pi} = [0.6639,\ 0.3361]^\intercal$ и векторами параметров $\mathbf{w}_1 = [1.20,\ 1.28]^\intercal$, $\mathbf{w}_2 = [1.08,\ -1.32]^\intercal$. Отметим, что в окрестности этого максимума нашлись еще два, близких по значению совместного правдоподобия, но отличающихся по $\boldsymbol{\pi}$, $\mathbf{w}_1$, $\mathbf{w}_2$, следующего вида

- $\boldsymbol{\pi} = [0.6587,\ 0.3413]^\intercal$ и векторами параметров $\mathbf{w}_1 = [1.19,\ 0.41]^\intercal$, $\mathbf{w}_2 = [0.53,\ 0.17]^\intercal$,
- $\boldsymbol{\pi} = [0.6623,\ 0.3367]^\intercal$ и векторами параметров $\mathbf{w}_1 = [1.20,\ 1.28]^\intercal$, $\mathbf{w}_2 = [1.07,\ -1.31]^\intercal$.

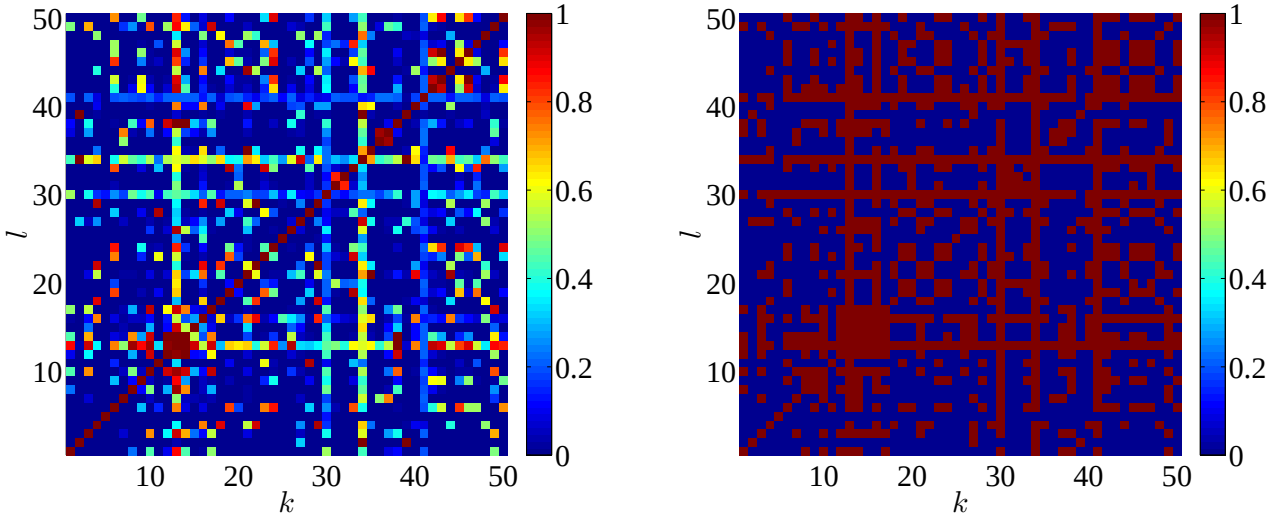
Вторый из этих максимумов очень близок к найденному решению, в то время как первый значительно отличается. Кроме этих двух максимумов есть еще несколько найденных локалных максимумов с значительно меньшим значением совместного правдоподобия.

- $\boldsymbol{\pi} = [1,\ 0]^\intercal$ и вектором параметров $\mathbf{w}_1 = [0.92,\ 0.31]^\intercal$,
- $\boldsymbol{\pi} = [0.9978,\ 0.0022]^\intercal$ и векторами параметров $\mathbf{w}_1 = [0.93,\ 0.32]^\intercal$, $\mathbf{w}_2 = [-0.24,\ -0.98]^\intercal$.

Таким образом, наряду с найденным решением, сходимость к которому наблюдается в широком диапазоне значений исходных параметров, существуют несколько локальных минимумов, что говорит о многоэкстремальности решаемой задачи.

Иллюстрация наадекватности обученной смеси моделей. Приведем далее пример, показывающий, что обученная смесь моделей может быть надекватной. Генерируем выборку размера $m = 5000$ в признаковом пространстве

размера $n = 2$ из смеси $K = 50$ моделей. При этом вектора параметров моделей генерируются так, что они достаточно разные, $\min_{k \neq l} \|\mathbf{w}_k - \mathbf{w}_l\| = 0.3334$. Обучим смесь моделей из $K = 50$ моделей на этих данных и сосчитаем уровни значимости для статистической различимости моделей, входящих в мультимодель. Матрица парных уровней значимости $\mathbf{T} = (t_{kl})$, $k, l = \overline{1, K}$ приведена на рис. 5.8.



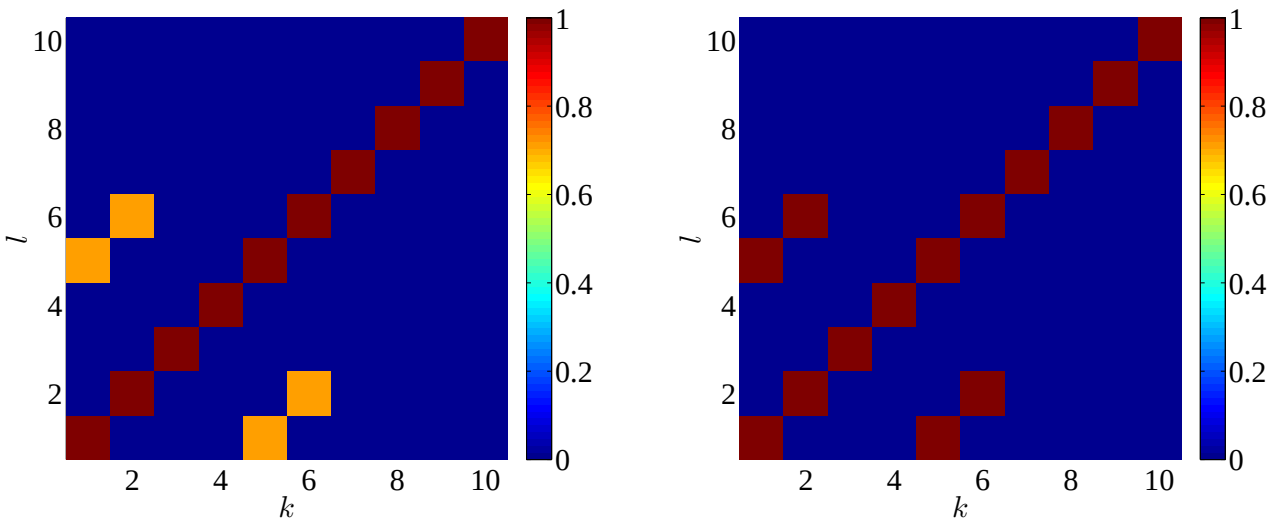
(а) $\mathbf{T}$ (б) $\tilde{\mathbf{T}} = (I(t_{kl} > 0.05))$, $k, l = \overline{1, K}$

Рис. 5.8: Матрица парных уровней значимости в условиях истинности гипотезы о совпадении моделей

Из рис. 5.8 заключаем, что в построенной смеси есть не (s, α) – различимые пары моделей на уровне значимости $\alpha = 0.05$. Таких пар 464, а для $\alpha = 0.01$ есть 717 пар не (s, α) – различимых моделей. Кроме того, в обученной смеси есть гораздо более близкие модели, чем в истинной модели $\min_{k \neq l} \|\hat{\mathbf{w}}_k - \hat{\mathbf{w}}_l\| = 0.0305$.

Сгенерируем теперь выборку того же размера $m = 5000$ в признаковом пространстве размера $n = 2$, но из смеси $K = 10$ моделей. При этом вектора параметров моделей генерируются так, что они достаточно разные, $\min_{k \neq l} \|\mathbf{w}_k - \mathbf{w}_l\| = 1.652$. Обучим смесь моделей из $K = 10$ моделей на этих данных и сосчитаем уровни значимости для статистической различимости моделей, входящих в мультимодель. Матрица парных уровней значимости $\mathbf{T} = (t_{kl})$, $k, l = \overline{1, K}$ приведена на рис. 5.9.

Из рис. 5.9 заключаем, что в построенной смеси есть не (s, α) – различимые пары моделей на уровне значимости $\alpha = 0.05$. Таких пар 2, а для $\alpha = 0.01$ есть 5 пар не (s, α) – различимых моделей. Кроме того, в обученной смеси есть гораздо более близкие модели, чем в истинной модели $\min_{k \neq l} \|\hat{\mathbf{w}}_k - \hat{\mathbf{w}}_l\| = 0.114$. Таким образом, приведенные примеры показывают, что обученная смесь моделей может не быть (s, α) – адекватной из-за отсутствия прямого контроля различимости моделей в смеси.



(а) $\mathbf{T}$ (б) $\tilde{\mathbf{T}} = (I(t_{kl} > 0.05))$, k, $l = \overline{1,\ K}$

Рис. 5.9: Матрица парных уровней значимости в условиях истинности гипотезы о совпадении моделей

5.4. Иллюстрация применения комбинирования признаков

Продемонстрируем работу предлагаемых алгоритмов комбинирования признаков на синтетических данных. Рассмотрим следующие наборов синтетических данных, содержащих повторяющиеся признаки.

1. $n_0 = 1$ фактор $\tilde{\mathbf{f}}^1$, имеющий вес w_0, $\tilde{\mathbf{f}}^1 \sim N(\mathbf{0}, \mathbf{I})$. Целевая переменная $\mathbf{y}$ генерируется в соответствии с моделью одиночной логистической регрессии (2.1) для $\mathbf{w} = w_0$, $\mathbf{X} = \tilde{\mathbf{f}}^1$. Наблюдаемая признаковая матрица $\mathbf{X}$ состоит из $n = 100$ зашумленных копий признака, $\mathbf{X} = \tilde{\mathbf{f}}^1\mathbf{e}_n^\intercal + \varepsilon_0\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon}_{ij} \sim N(0,\ 1)$, ε_0 задает уровень зашумления.
2. $n_0 = 2$ независимых фактора $\tilde{\mathbf{f}}^1$, $\tilde{\mathbf{f}}^2$, имеющих вес w_0, $\tilde{\mathbf{f}}^1$, $\tilde{\mathbf{f}}^2 \sim N(\mathbf{0}, \mathbf{I})$. Целевая переменная $\mathbf{y}$ генерируется в соответствии с моделью одиночной логистической регрессии (2.1) для $\mathbf{w} = [w_0,\ w_0]^\intercal$, $\mathbf{X} = [\tilde{\mathbf{f}}^1,\ \tilde{\mathbf{f}}^2]$. Наблюдаемая признаковая матрица $\mathbf{X}$ состоит из $n = 50$ зашумленных копий каждого признака, $\mathbf{X} = [\tilde{\mathbf{f}}^1\mathbf{e}_n^\intercal + \varepsilon_0\boldsymbol{\varepsilon}^1,\ \tilde{\mathbf{f}}^2\mathbf{e}_n^\intercal + \varepsilon_0\boldsymbol{\varepsilon}^2]$, $\boldsymbol{\varepsilon}^1_{ij}$, $\boldsymbol{\varepsilon}^2_{ij} \sim N(0,\ 1)$, ε_0 задает уровень зашумления.
3. $n_0 = 2$ независимых фактора $\tilde{\mathbf{f}}^1$, $\tilde{\mathbf{f}}^2$, имеющих веса w_0, $-w_0$, $\tilde{\mathbf{f}}^1$, $\tilde{\mathbf{f}}^2 \sim N(\mathbf{0}, \mathbf{I})$. Целевая переменная $\mathbf{y}$ генерируется в соответствии с моделью одиночной логистической регрессии (2.1) для $\mathbf{w} = [w_0,\ -w_0]^\intercal$, $\mathbf{X} = [\tilde{\mathbf{f}}^1,\ \tilde{\mathbf{f}}^2]$. Наблюдаемая признаковая матрица $\mathbf{X}$ состоит из $n = 50$ зашумленных копий $\mathbf{f}^1 = \beta\tilde{\mathbf{f}}^1 + \sqrt{1-\beta^2}\tilde{\mathbf{f}}^2$, $\mathbf{f}^2 = \beta\tilde{\mathbf{f}}^2 + \sqrt{1-\beta^2}\tilde{\mathbf{f}}^1$ смесей каждого признака, $\mathbf{X} = [\mathbf{f}^1\mathbf{e}_n^\intercal + \varepsilon_0\boldsymbol{\varepsilon}^1,\ \mathbf{f}^2\mathbf{e}_n^\intercal + \varepsilon_0\boldsymbol{\varepsilon}^2]$, $\boldsymbol{\varepsilon}^1_{ij}$, $\boldsymbol{\varepsilon}^2_{ij} \sim N(0,\ 1)$, ε_0 задает уровень зашумления. Здесь $\beta \in [0,\ 1/\sqrt{2}]$ задает уровень корреляции между признаками ($\beta = 0$ соответствует нулевой корреляции, а $\beta = 1/\sqrt{2}$ случаю единичной корреляции).

Случай, когда признаковая матрица состоит из $n = 100$ копий одного признака. В рассматриваемом случае оптимальной коибинацией в соответствии с (2.4) является комбинация признаков с одинаковым весом. При этом такой комбинации соответствует следующее представление комбинированного признака после нормировки к единичной дисперсии

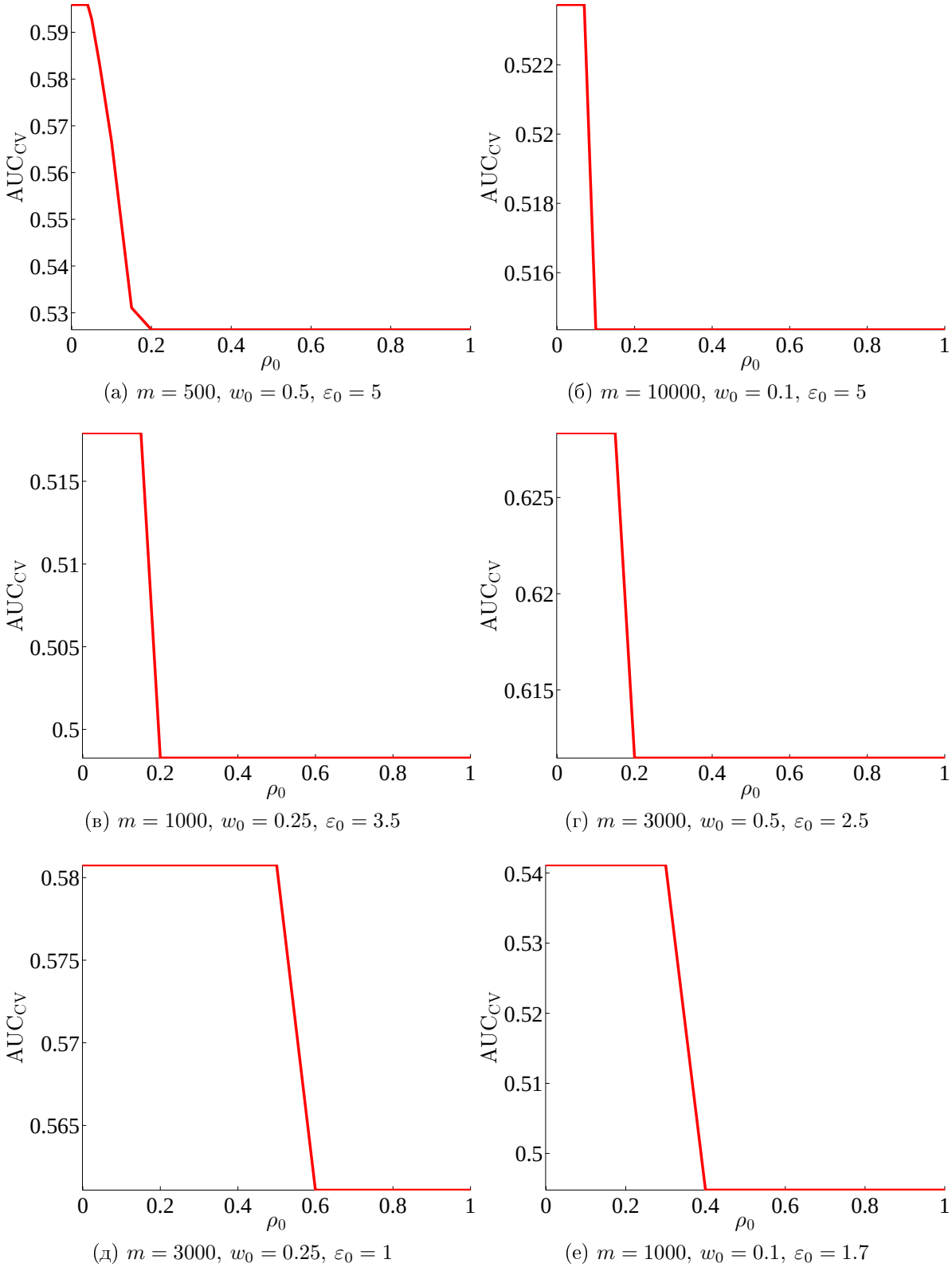
$$\mathbf{f}^* = \alpha\tilde{\mathbf{f}}^1 + \sqrt{1-\alpha^2}\boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\mathbf{0},\ \mathbf{I}),\ \alpha^2 = n/(n+\varepsilon_0^2).$$

Здесь $D_{\text{best}} = \alpha^2$ есть доля сигнала в оптимальной комбинации по дисперсии. Такое же представление имеет место и для других комбинаций, которые задаются весами обученной логистической модели после шкалирования к единичной дисперсии. Отметим, что отклонение комбинации от оптимальной, которое можно измерить по величине доли сигнала D влияет на качество прогноза. Далее в терминах доли сигнала и AUC на тестовой выборке сравнивам оптимальную комбинацию признаков $(D_{opt}, \text{AUC}_{opt})$, наиболее обоснованную модель, полученную с помощью логистической регрессии на исходных признаках $(D_{ev}, \text{AUC}_{ev})$, модель, построенную на одном признаке (D_s, AUC_s) и наиболее обоснованную обученную модель на комбинированных признаках $(D_{comb}, \text{AUC}_{comb})$ для разных размеров выборки m, значения w_0 и зашумленности ε_0.

Отметим, что для того, чтобы проводить комбинацию признаков с помощью предлагаемых методов, требуется задать границу ρ_0 по корреляции для комбинации. Если $\rho_0 = -1$, то все признаки комбинируется в один, а если $\rho_0 = 1$, то комбинирования не происходит. Так как значение ρ_0 неизвестно, оно определяется кросс-валидацией на обучающей выборке, для которой производится случайное разбиение на две подвыборки 100 раз. Приведем далее несколько графиков зависимости среднего качества на кросс-валидации от значения ρ_0 для разных значений m, w_0, ε_0 (см. рис. 5.10.

Зависимость AUC на кросс-валидации имеет ступенчатый характер (см. рис. 5.10). Оптимальной является комбинация всех признаков, а при малом параметре ρ_0 это и происходит, что ведет к качеству, близкому к оптимальному. При увеличении ρ_0 до некоторого значения по-прежнему происходит комбинация всех признаков в один, а потому качество не изменяется. При достаточно большом ρ_0 признаки используются отдельно, что ведет к более низкому качеству. Отметим, что граница перехода равна как раз $1/(1+\varepsilon_0^2)$, поскольку это и есть ожидаемое значение корреляции между зашумленными признаками. Более низкое качество при использовании признаков отдельно указывает на неоптимальность комбинации, которая получается при использовании исходных признаков в оптимальной обученной модели логистической регрессии.

Сравним далее оптимальную обученную логистическую модель на комбинированных признаках с оптимальной обученной логистической моделью на исходных признаках, с оптимальной обученной логистической моделью на одном признаке и с моделью с оптимальным комбинированием признаков для разных значений w_0, задающего максимальное качество модели. Используются



(а) $m = 500$, $w_0 = 0.5$, $\varepsilon_0 = 5$

(б) $m = 10000$, $w_0 = 0.1$, $\varepsilon_0 = 5$

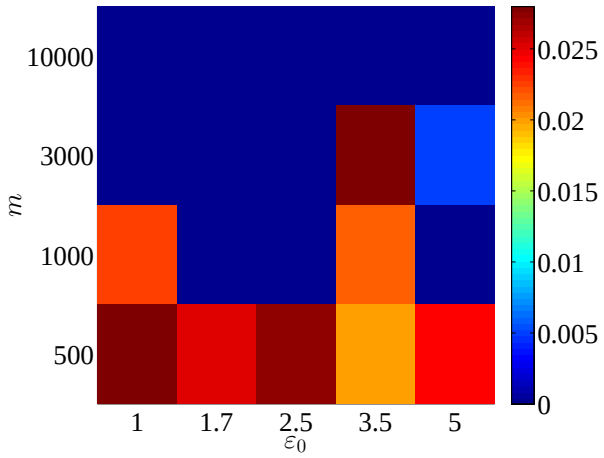
(в) $m = 1000$, $w_0 = 0.25$, $\varepsilon_0 = 3.5$

(г) $m = 3000$, $w_0 = 0.5$, $\varepsilon_0 = 2.5$

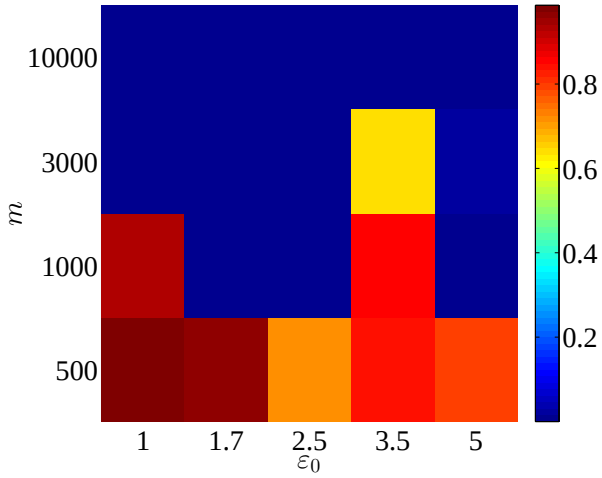
(д) $m = 3000$, $w_0 = 0.25$, $\varepsilon_0 = 1$

(е) $m = 1000$, $w_0 = 0.1$, $\varepsilon_0 = 1.7$

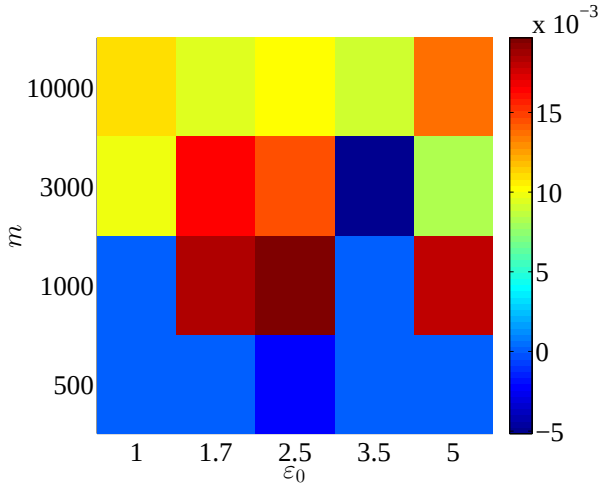
Рис. 5.10: Зависимость AUC на кросс-валидации от параметра отсечения по корреляции ρ_0 для разных m, w_0, ε_0.



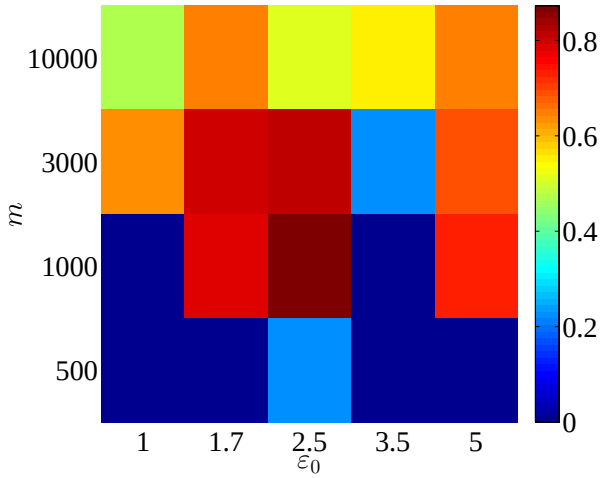
(а) Разница между AUC_{opt} и AUC_{comb}



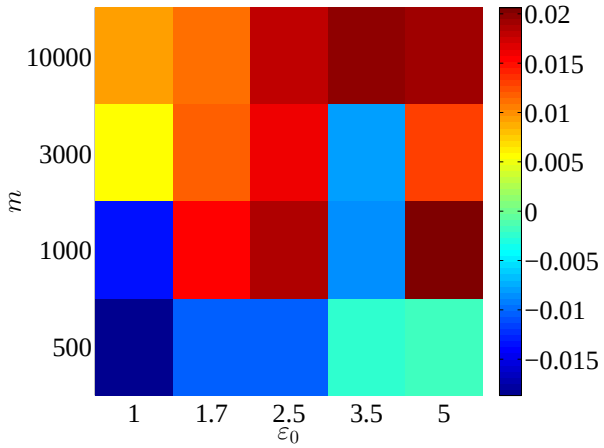
(б) Разница между D_{opt} и D_{comb}



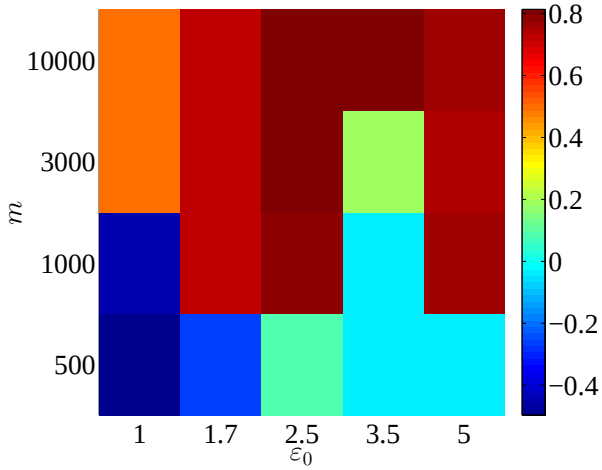
(в) Разница между AUC_{comb} и AUC_{ev}



(г) Разница между D_{comb} и D_{ev}

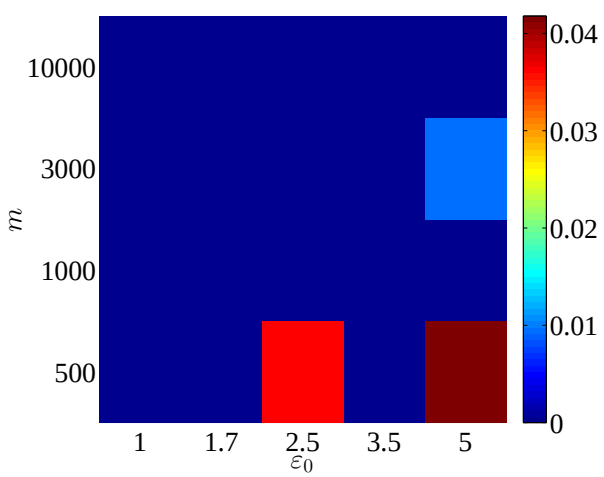


(д) Разница между AUC_{comb} и AUC_{s}

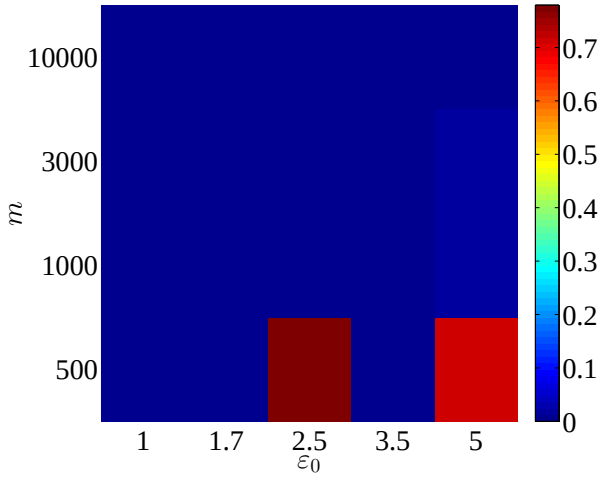


(е) Разница между D_{comb} и D_{s}

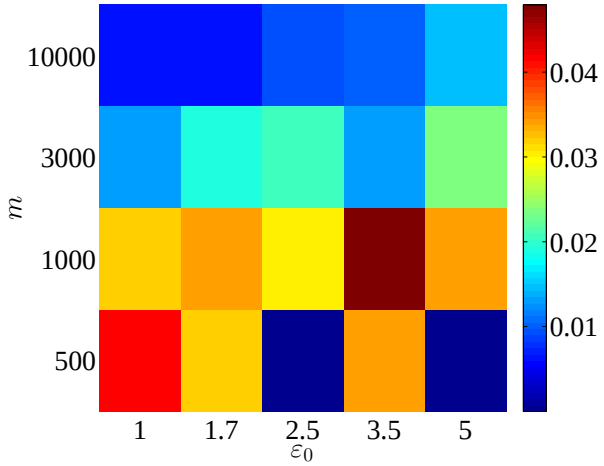
Рис. 5.11: Сравнение модели с комбинированными признаками с оптимальной моделью и моделью с исходными признаками для $w_0 = 0.1$



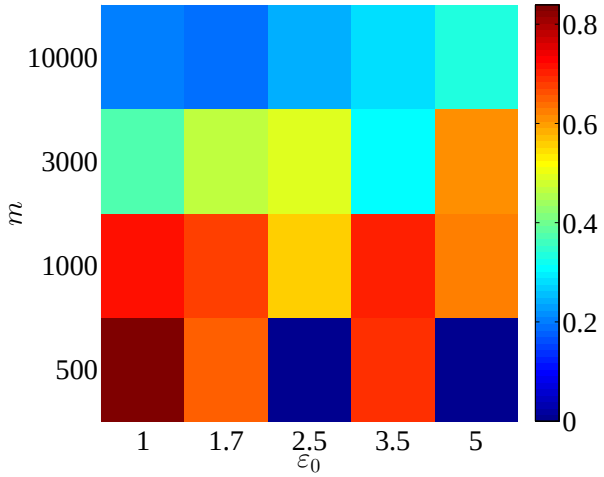
(а) Разница между AUC_{opt} и AUC_{comb}



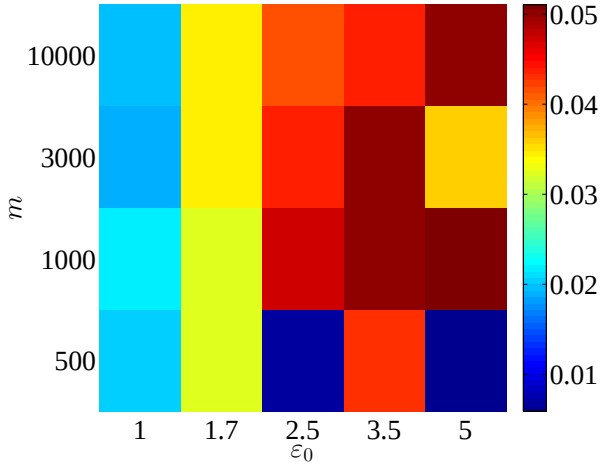
(б) Разница между D_{opt} и D_{comb}



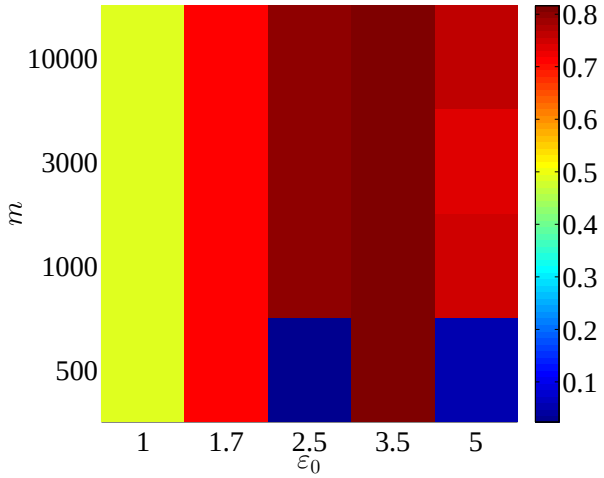
(в) Разница между AUC_{comb} и AUC_{ev}



(г) Разница между D_{comb} и D_{ev}

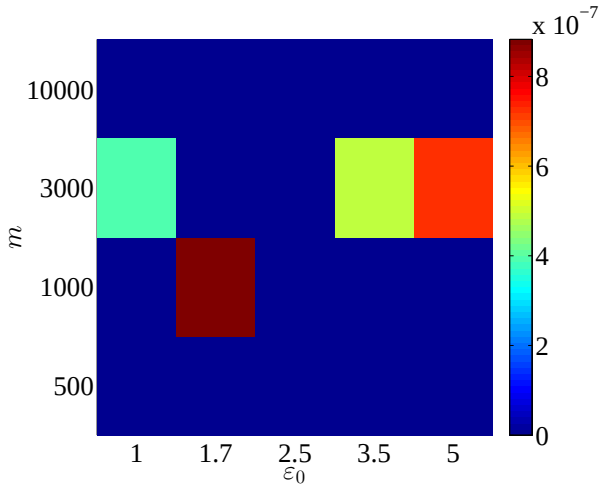


(д) Разница между AUC_{comb} и AUC_{s}

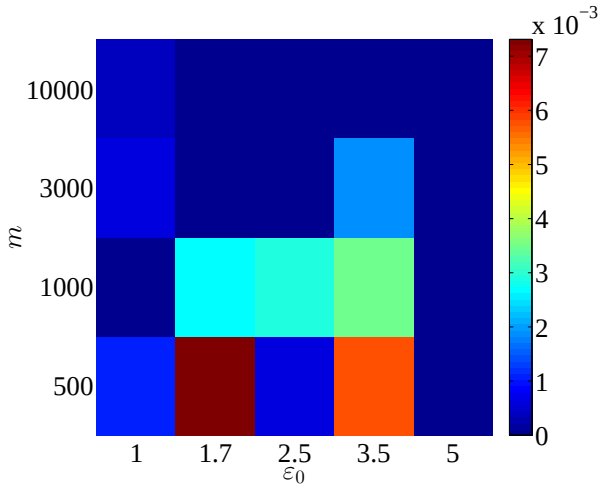


(е) Разница между D_{comb} и D_{s}

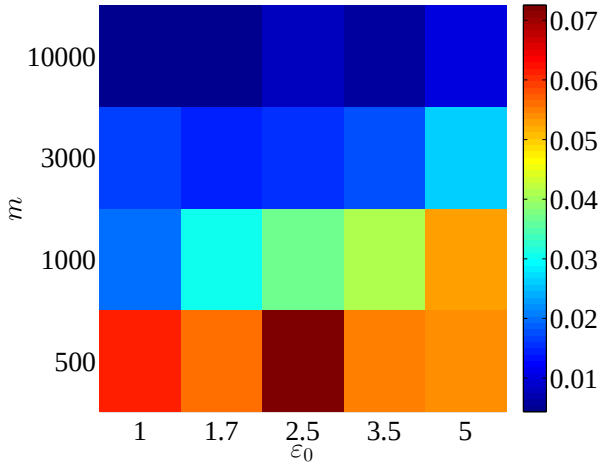
Рис. 5.12: Сравнение модели с комбинированными признаками с оптимальной моделью и моделью с исходными признаками для $w_0 = 0.25$



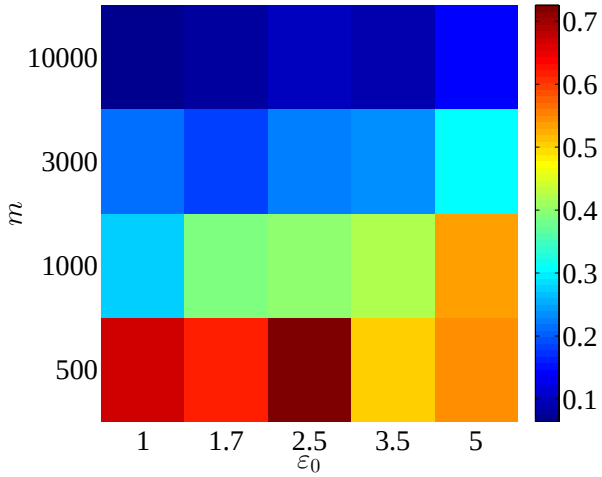
(а) Разница между AUC_{opt} и AUC_{comb}



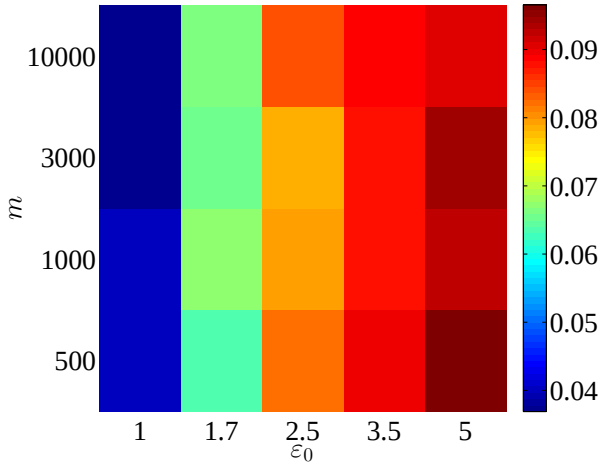
(б) Разница между D_{opt} и D_{comb}



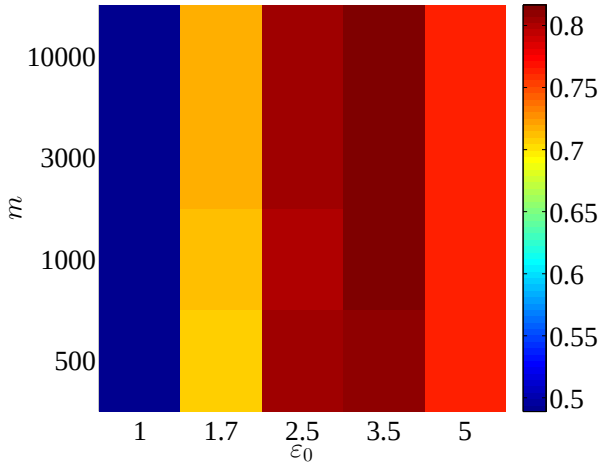
(в) Разница между AUC_{comb} и AUC_{ev}



(г) Разница между D_{comb} и D_{ev}



(д) Разница между AUC_{comb} и AUC_{s}



(е) Разница между D_{comb} и D_{s}

Рис. 5.13: Сравнение модели с комбинированными признаками с оптимальной моделью и моделью с исходными признаками для $w_0 = 0.5$

следующие значения параметров m, w_0, ε_0, $m \in \{500, 1000, 3000, 10000\}$, $w_0 \in \{0.1, 0.25, 0.5\}$, $\varepsilon_0 \in \{1, 1.7, 2.5, 3.5, 5\}$.

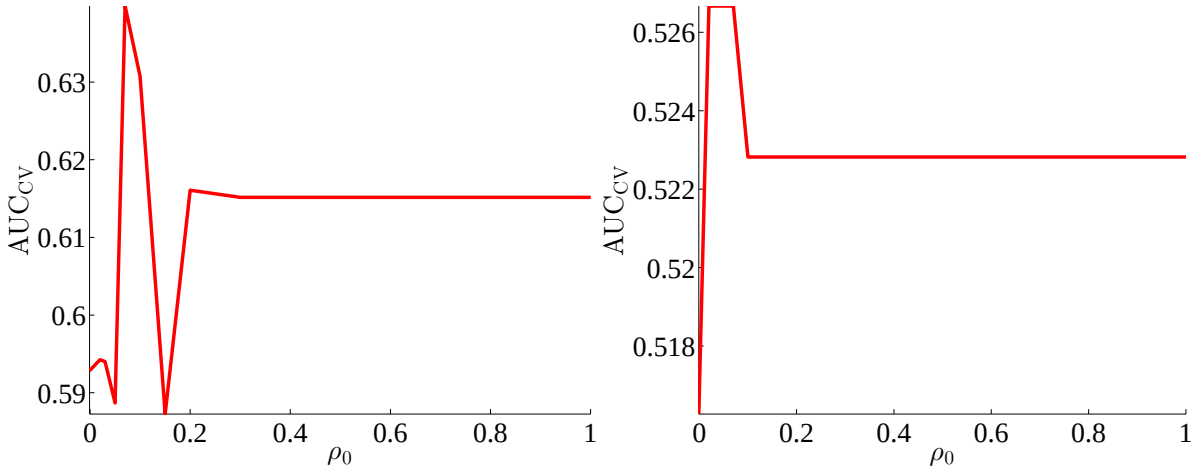
Результаты для модели с $w_0 = 0.1$, то есть с малым максимальным качеством приведены на рис. 5.11. Отметим, что модель с комбинированными признаками близка к оптимальной даже при большом уровн шума при $m \geq 1000$. При $m = 500$ из-за наличия шума, который влияет на оценки корреляций модель с комбинированными признаками уступает оптимальной, но по-прежнем не проигрывает оптимальной обученной модели, использующей все признаки. Отметим также, что и модель с комбинированными признаками, и оптимальная обученная модель, использующая все признаки работают лучше, чем модель, обученная после выбора одного признака за исключением случая малой выборки $m = 500$ для большого значения шума. Отметим, однако, что реализация отбора признаков в таких условиях затруднительна, поскольку математическое ожидание корреляции между признаками равно $1/(1 + \varepsilon_0^2)$, что близко к нулю при сильном шуме, а потому признаки не выглядят мультиколлинеарными.

Результаты для случае, когда $w_0 = 0.25$ и $w_0 = 0.5$ приведены на рис. 5.12и 5.13 соответственно. В этом случае максимальнаое качество для оптимальной модели выше и равно соответственно 0.57 и 0.63 против 0.52-0.53 для $w_0 = 0.1$. При $w_0 = 0.5$ уже даже для выборки размера $m = 500$ для всех значений шумов модель с комбинированными признаками почти совпадает (с точностью до 10^{-7} по доле сигнала и до 10^{-3} по AUC) с оптимальной моделью, значительно побеждая при этом оптимальную обученную модель на всех признаках и модель с отобранным признаком во всех случаях. Даже в случае большой выборки $m = 10000$ превышение AUC_{comb} над AUC_{ev} составляет 0.1. Отметим при этом, что преимущество больше для меньших выборок и при большей зашумленности ε_0. Для случая $w_0 = 0.25$ справедливо то же самое, за исключением случая малой выборки $m = 500$ В большой зашумленности ($\varepsilon_0 = 2.5, 5$), когда модель с комбинированными признаками побеждает оптимальную обученную на всех признаках и модель с отобранным признаком, но значительно проигрывает наилучшей возможной.

Случай, когда признаковая матрица состоит из $n = 50$ копий каждого из двух признаков.

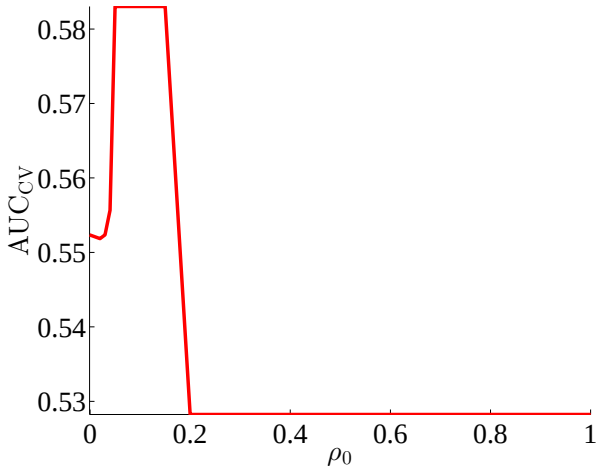
Как и в предыдущем случае, приведем сначала несколько графиков зависимости среднего качества на кросс-валидации от значения ρ_0 для разных значений m, w_0, ε_0 (см. рис. 5.10.

Зависимость AUC на кросс-валидации имеет ступенчатый характер (см. рис. 5.14). Оптимальным является набор признаков, состоящий из двух признаков, каждый из который является комбинацией $n = 50$ копий каждого из двух истинных признаков. При малом параметре ρ_0 все признаки комбинируются в один. Если бы признаки комбинировались с одинаковыми положительными весами, то такая комбинация была бы оптимальной. Однако знак в комбинации определяется корреляцией между признаками. Внутри одной группы копий ожидаемое значение корреляции равно $1/(1 + \varepsilon_0^2) > 0$, что позволяет даже при

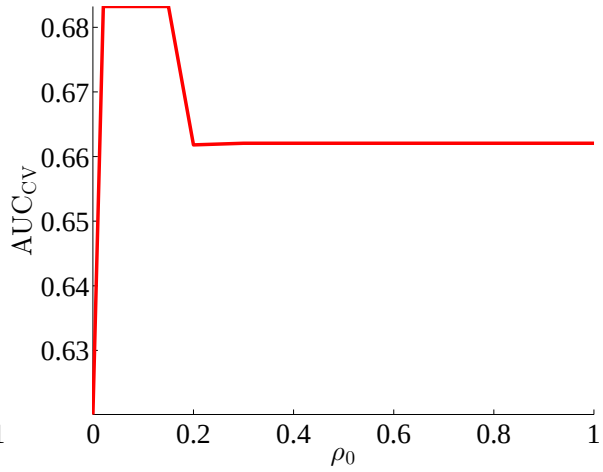


(а) $m = 500$, $w_0 = 0.5$, $\varepsilon_0 = 5$

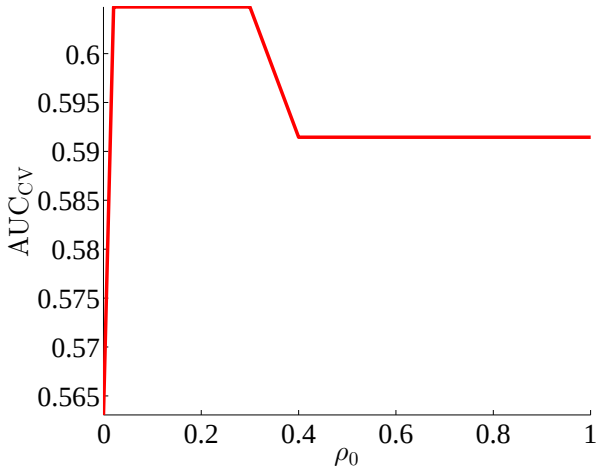
(б) $m = 10000$, $w_0 = 0.1$, $\varepsilon_0 = 5$



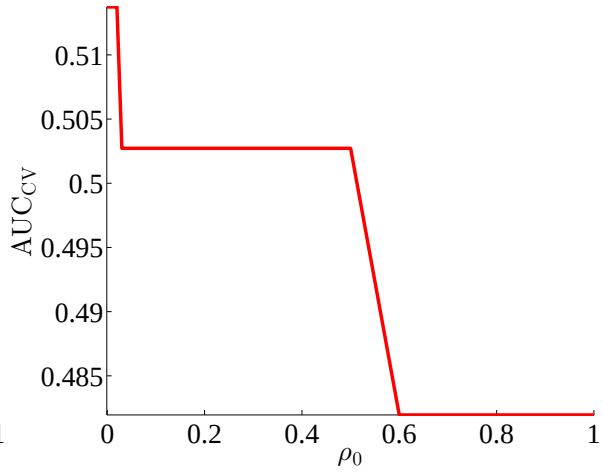
(в) $m = 1000$, $w_0 = 0.25$, $\varepsilon_0 = 3.5$



(г) $m = 3000$, $w_0 = 0.5$, $\varepsilon_0 = 2.5$



(д) $m = 3000$, $w_0 = 0.25$, $\varepsilon_0 = 1.7$



(е) $m = 1000$, $w_0 = 0.1$, $\varepsilon_0 = 1$

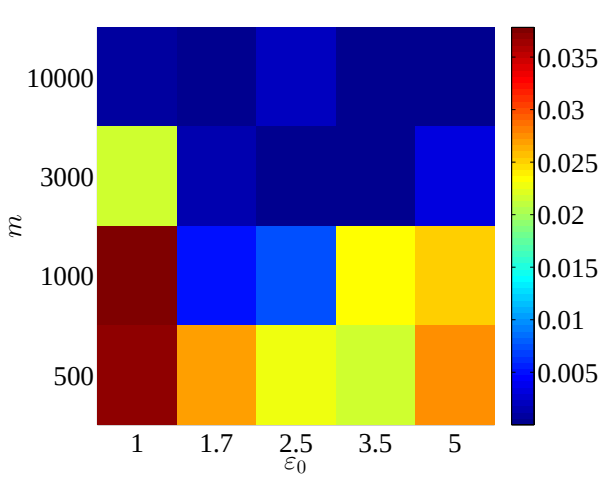
Рис. 5.14: Зависимость AUC на кросс-валидации от параметра отсечения по корреляции ρ_0 для разных m, w_0, ε_0.

большой зашумленности складывать признаки с правильным знаком. Однако между группами ожидаемая корреляция признаков равна нулю, а потому в среднем 25 из 100 признаков будут сложены с противоположным знаком, что ведет к неоптимальности и реальному учету только одного из факторов $\tilde{\mathbf{f}}^1$, $\tilde{\mathbf{f}}^2$. По мере возрастания ρ_0 признаки из разных групп копий начинают считаться низкокоррелированными, а признаки из одной группы по-прежнему объединяются. Этот участок и соответствует оптимальному комбинированию. При дальнейшем росте ρ_0 группы признаков разъединяются и признаки используются отдельно. Более низкое качество при использовании признаков отдельно указывает на неоптимальность комбинации, которая получается при использовании исходных признаков в оптимальной обученной модели логистической регрессии. Отметим, однако, что при малом числе объектов и при большой зашумленности указанный участок оптимального комбинирования может быть узок или не наблюдаться в силу зашумленности выборочных оценок корреляции.

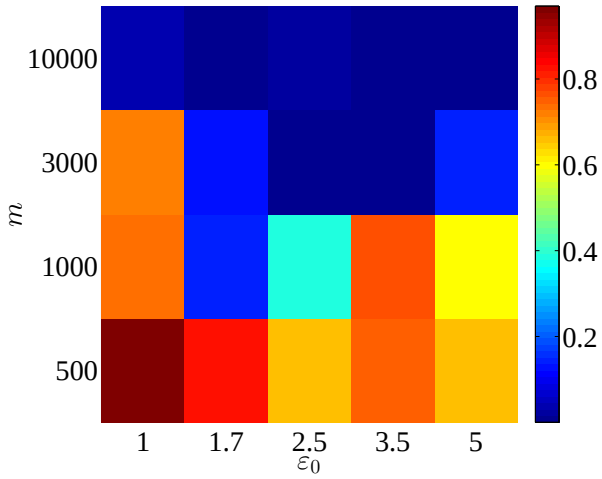
Сравним далее оптимальную обученную логистическую модель на комбинированных признаках с оптимальной обученной логистической моделью на исходных признаках, с оптимальной обученной логистической моделью на одном признаке и с моделью с оптимальным комбинированием признаков для разных значений w_0, задающего максимальное качество модели. Используются следующие значения параметров m, w_0, ε_0, $m \in \{500, 1000, 3000, 10000\}$, $w_0 \in \{0.1, 0.25, 0.5\}$, $\varepsilon_0 \in \{1, 1.7, 2.5, 3.5, 5\}$.

Результаты для модели с $w_0 = 0.1$, то есть с малым максимальным качеством приведены на рис. 5.15. Отметим, что модель с комбинированными признаками близка к оптимальной даже при большом уровне шума при $m = 10000$. При меньших размерах выборки m модель с комбинированными признаками слабее оптимальной в среднем на 0.015 по AUC, однако по-прежнему выигрывает у оптимальной обученной модели, использующей все признаки, во всех случаях, кроме двух, достигая схожих показателей качества для малых выборок $m = 500$. Отметим также, что и модель с комбинированными признаками, и оптимальная обученная модель, использующая все признаки работают лучше, чем модель, обученная после выбора одного признака за исключением одного случая малой выборки $m = 500$.

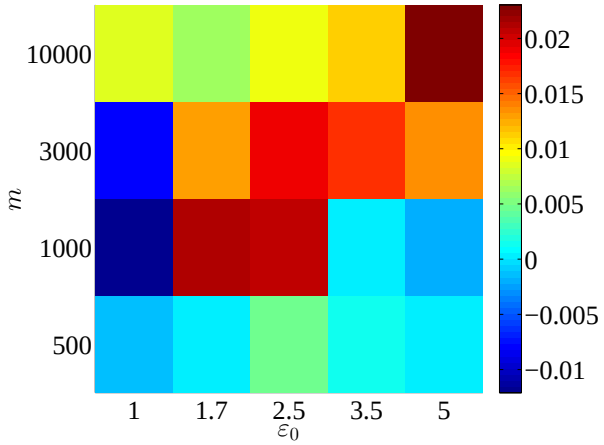
Результаты для случае, когда $w_0 = 0.25$ и $w_0 = 0.5$ приведены на рис. 5.16и 5.17 соответственно. В этом случае максимальное качество для оптимальной модели выше и равно соответственно 0.57 и 0.63 против 0.52-0.53 для $w_0 = 0.1$. При $w_0 = 0.5$ модель с комбинированными признаками для всех размеров выборки и для всех шумов, кроме наибольшего совпадает с оптимальной, будучи близкой к оптимальной при $\varepsilon_0 = 5$. Кроме того, модель с комбинированными признаками выигрывает у обученной оптимальной модели, использующей все признаки отдельно во всех случаях, кроме одного, и значительно опережают модель с отобраными признаками во вех случаях. Для $w = 0.25$ наблюдается схожий результат, но модель с комбинированными признаками совпадает с оптимальной лишь при $m = 10000$ и при малом шуме.



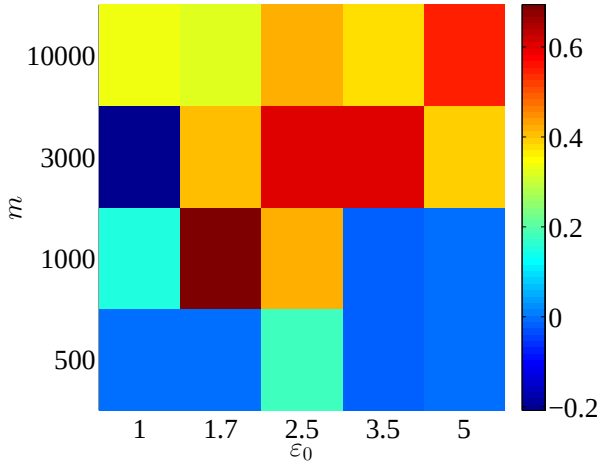
(а) Разница между AUC_{opt} и AUC_{comb}



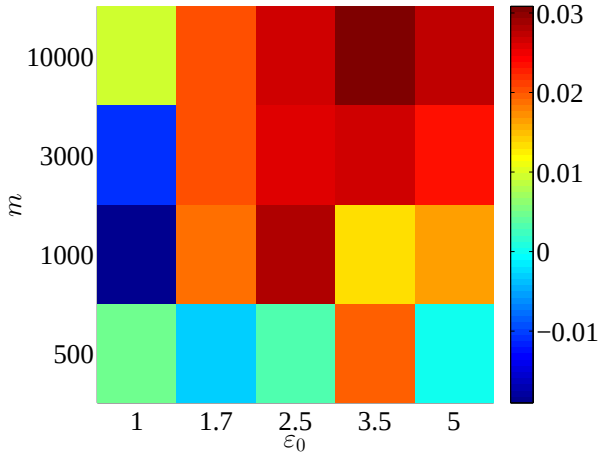
(б) Разница между D_{opt} и D_{comb}



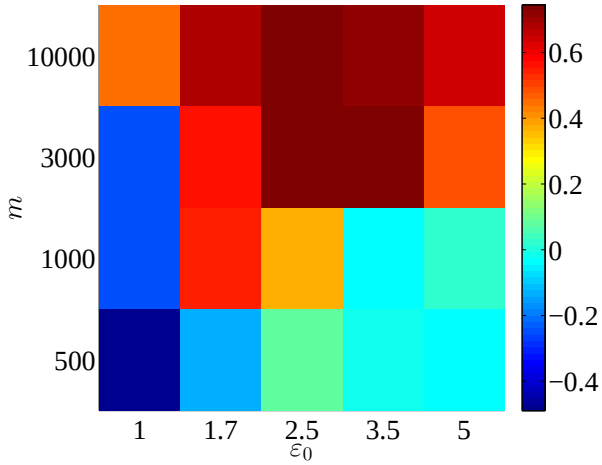
(в) Разница между AUC_{comb} и AUC_{ev}



(г) Разница между D_{comb} и D_{ev}

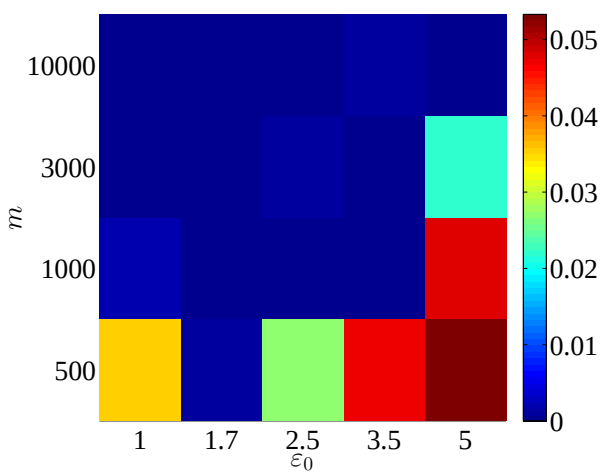


(д) Разница между AUC_{comb} и AUC_{s}

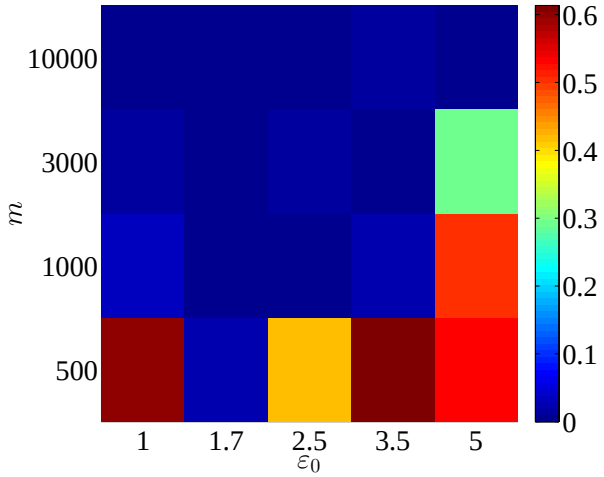


(е) Разница между D_{comb} и D_{s}

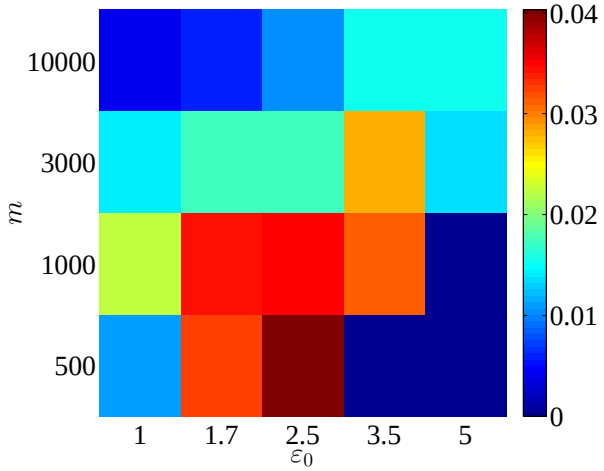
Рис. 5.15: Сравнение модели с комбинированными признаками с оптимальной моделью и моделью с исходными признаками для $w_0 = 0.1$



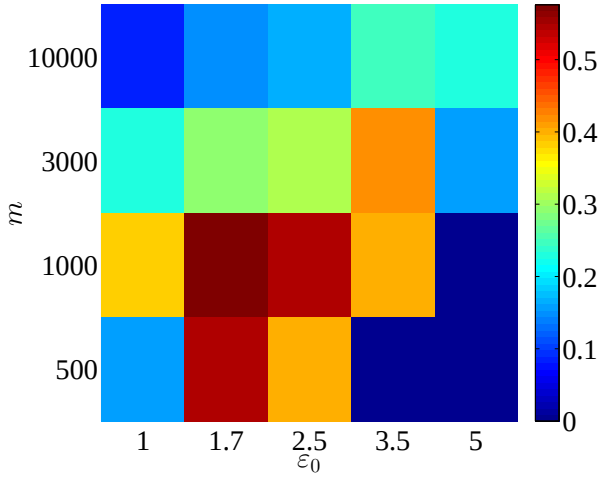
(а) Разница между AUC_{opt} и AUC_{comb}



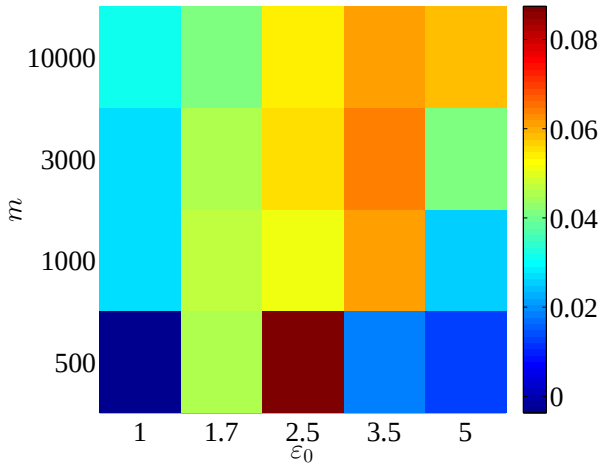
(б) Разница между D_{opt} и D_{comb}



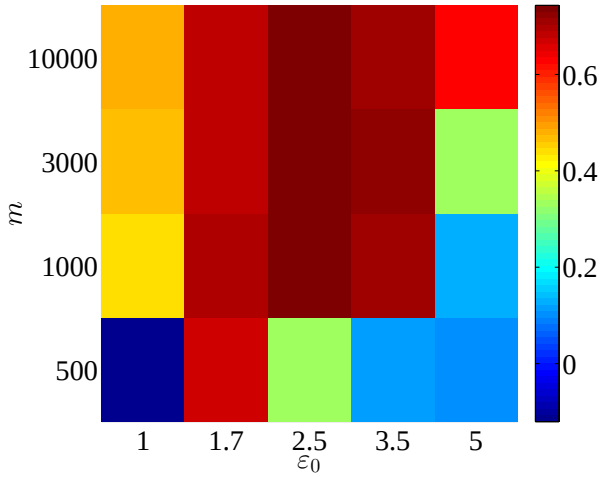
(в) Разница между AUC_{comb} и AUC_{ev}



(г) Разница между D_{comb} и D_{ev}

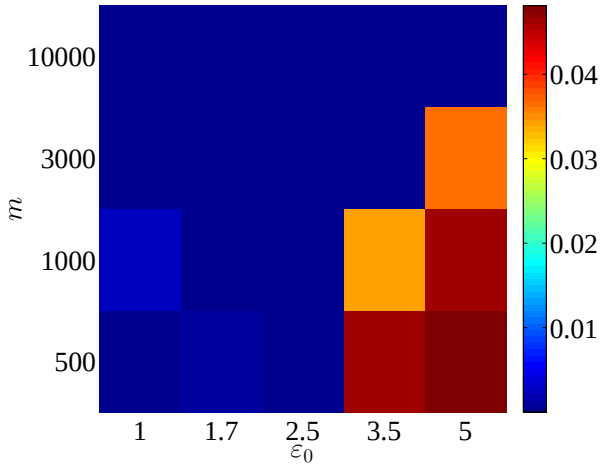


(д) Разница между AUC_{comb} и AUC_{s}

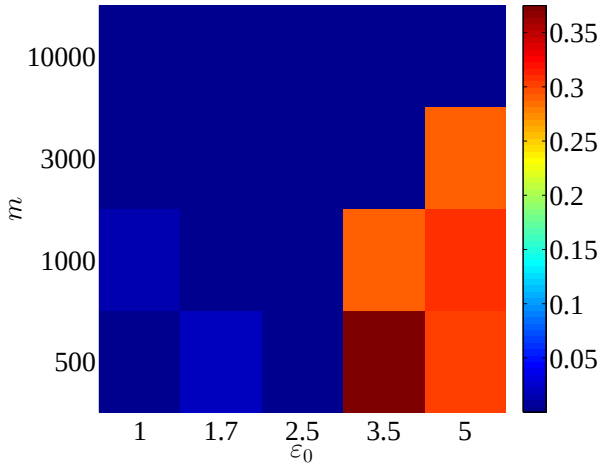


(е) Разница между D_{comb} и D_{s}

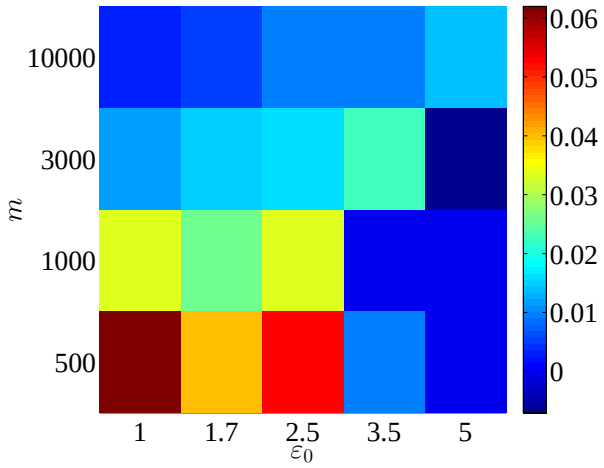
Рис. 5.16: Сравнение модели с комбинированными признаками с оптимальной моделью и моделью с исходными признаками для $w_0 = 0.25$



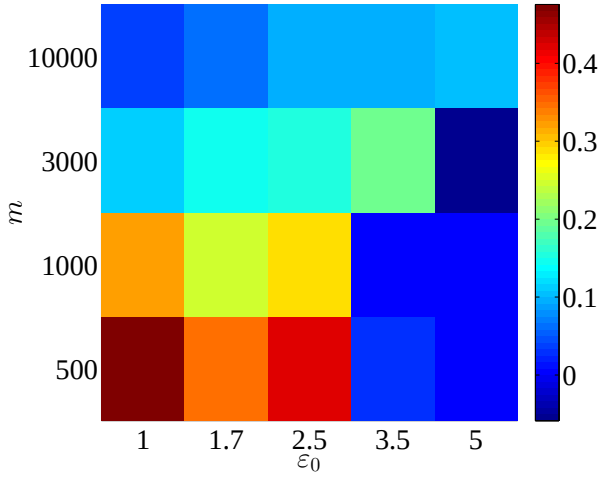
(а) Разница между AUC_{opt} и AUC_{comb}



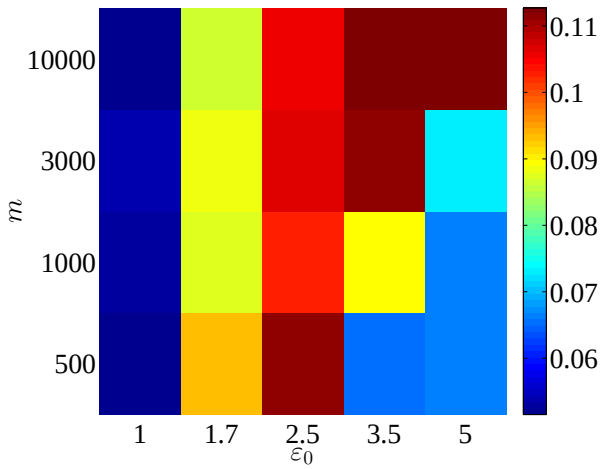
(б) Разница между D_{opt} и D_{comb}



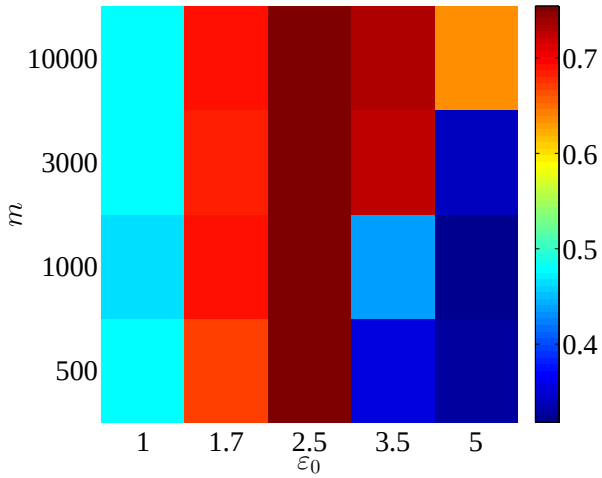
(в) Разница между AUC_{comb} и AUC_{ev}



(г) Разница между D_{comb} и D_{ev}



(д) Разница между AUC_{comb} и AUC_{s}



(е) Разница между D_{comb} и D_{s}

Рис. 5.17: Сравнение модели с комбинированными признаками с оптимальной моделью и моделью с исходными признаками для $w_0 = 0.5$

Отметим, что в предыдущем случае оптимальной стратегией было комбинирование всех имеющихс признаков, и потому хотя необходимость комбинирования и определяеся кросс-валидацией, модель, основанная на комбинировании признаков имеет преимущество. Однако в рассматриваемом случае такая стратегия приводит, как уже указывалось, к неоптимальному результату из-за нулевой ожидаемой корреляции признаков из разных групп. По этой причине правильное определение ρ_0 является определяющим для модели с комбинированными признаками. Полученные результаты показывают, что предлагаемый метод справляется с определением границы ρ_0 и построением комбинации признаков, близкой к оптимальной, даже в условиях сильной зашумленности. Более того полученные результаты показывают неоптимальность оценки вектора параметров модели в обученной оптимальной модели, поскольку часть признаков исключается из рассмотрения, хотя доля сигнала во всех из них одинакова, а часть признаков получают отрицательный вес. Таким образом, предлагаемый метод комбинирования признаков позволяет улучшить качество прогноза при наличии мультиколлинеарных признаков. При этом при их отсутствии в силу выбора ρ_0 на кросс-валидации, комбинирование не будет производиться, что и наблюдается во втором случае, когда комбинирование признаков из разных групп нежелательно.

5.5. Иллюстрация применения s-score для сравнения моделей

Покажем, как предлагаемую функцию сходства s-score можно использовать для сравнения моделей. Для примера рассмотрим случай признакового пространства размерности 2, то есть $k = 2$.

Пусть имеется две разных модели логистической регрессии, задаваемых векторами параметров $\mathbf{w}_1 \in \mathbb{R}^2$, $\mathbf{w}_2 \in \mathbb{R}^2$. Пусть для первой модели известна выборка размера n_1 $(\mathbf{X}_1, \mathbf{y}_1)$, а для второй – выборка размера n_2 $(\mathbf{X}_2, \mathbf{y}_2)$, причем выборки не пересекаются по объектам. Пусть также задан уровень значимости α для задачи сравнения моделей. В качестве априорного распределения используем на параметры $\mathbf{w}_1$ и $\mathbf{w}_2$ для обоих моделей используем равномерное псевдораспределение в $\mathbb{R}^2$. Тогда получим следующее совместное правдоподобие для каждой из моделей.

$$p(\mathbf{y}_k,\ \mathbf{w}_k|\mathbf{X}_k) = p(\mathbf{y}_k|\mathbf{X}_k,\ \mathbf{w}_k)p(\mathbf{w}_k) \propto p(\mathbf{y}_k|\mathbf{X}_k,\ \mathbf{w}_k),$$

где правдоподобие $p(\mathbf{y}_k|\mathbf{X}_k,\ \mathbf{w}_k)$ имеет вид

$$p(\mathbf{y}_k|\mathbf{X}_k,\ \mathbf{w}_k) = \prod_{i=1}^{n_k} \sigma(\mathbf{w}_k^\intercal \mathbf{x}_k^i),\ \text{где}\ \sigma(x) = \frac{1}{1+\exp(-x)}.$$

Пользуясь асимптотической нормальностью апостериорного распределения $p(\mathbf{w}_k|\mathbf{X}_k,\ \mathbf{y}_k)$ вектора параметров получаем следующую аппроксимацию

$$p(\mathbf{w}_k|\mathbf{X}_k,\ \mathbf{y}_k) \approx N(\mathbf{w}_k|\hat{\mathbf{w}}_k,\ \hat{\mathbf{\Sigma}}_k),$$

где

$$\hat{\mathbf{w}}_k = \arg\max_{\mathbf{w}} p(\mathbf{w}_k|\mathbf{X}_k,\ \mathbf{y}_k) = \arg\max_{\mathbf{w}} p(\mathbf{y}_k|\mathbf{X}_k,\ \mathbf{w}_k)$$

есть оценка максимума апостериорной вероятности для вектора параметров, а

$$\hat{\mathbf{\Sigma}}_k = \mathbf{H}_k^{-1},\ \mathbf{H}_k = -\frac{d^2(\log p(\mathbf{w}_k|\mathbf{X}_k,\ \mathbf{y}_k))}{d\mathbf{w}^2}|_{\mathbf{w}_k=\hat{\mathbf{w}}_k}.$$

В соответствии с методом сравнения моделей для пары нормальных распределений $g_1(\mathbf{w}_1) = N(\mathbf{w}_1|\hat{\mathbf{w}}_1,\ \hat{\mathbf{\Sigma}}_1)$ и $g_2(\mathbf{w}_2) = N(\mathbf{w}_2|\hat{\mathbf{w}}_2,\ \hat{\mathbf{\Sigma}}_2)$ получаем следующее выражение для функции сходства s-score

$$s(g_1,\ g_2) = (\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1)^\intercal \left(\hat{\mathbf{\Sigma}}_1 + \hat{\mathbf{\Sigma}}_2\right)^{-1} (\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1).$$

В соответствии с теоремой **??** для признакового пространства размерности $k = 2$ в условиях истинности гипотезы о совпадении моделей $s \sim U[0,\ 1]$. Тогда на уровне значимости α модели считаются разными, если

$$s(g_1,\ g_2) < \alpha$$

и неразличимыми в противном случае.

Проиллюстрируем далее зависимость распределения функции близости s-score от числа объектов n_1, n_2 в выборках и от близости различаемых моделей. В качестве примера рассмотрим $\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = 1$, а в качестве меры близости рассмотрим косинусную меру, то есть $\rho(\mathbf{w}_1,\ \mathbf{w}_2) = \mathbf{w}_1^\intercal \mathbf{w}_2$.

Для получения распределения функции $s(g_1,\, g_2)$ для фиксированных $\mathbf{w}_1$, $\mathbf{w}_2$ и $\mathbf{X}_1$, $\mathbf{X}_2$ будем многократно генерировать $\mathbf{y}_1$, $\mathbf{y}_2$ в соответствии с моделями

$$y_k^i \sim \mathrm{Be}\left(\sigma(\mathbf{w}_k^\intercal \mathbf{x}_k^i)\right).$$

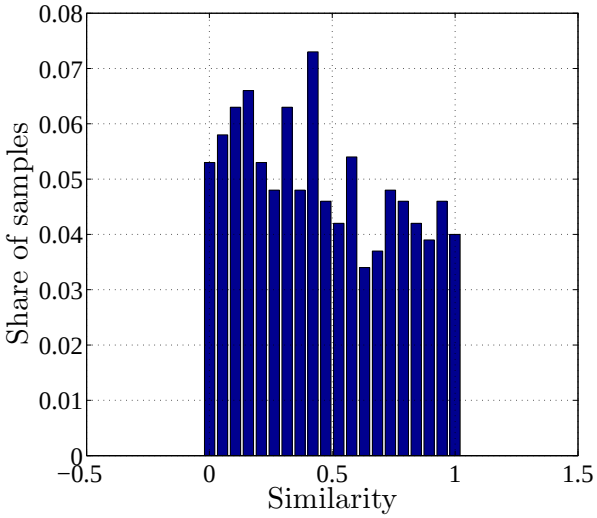
Полученные гистограммы распределения $s(g_1,\, g_2)$ для $n_1 = 10000$ в зависимости от размера второй выборки n_2 для случая, когда вектора параметров моделей $\mathbf{w}_1$, $\mathbf{w}_2$ имеют косинусное сходство $\rho = 0.95$ приведены на рис. 5.18. Отметим, что при увеличении размера второй выборки распределение функции сходства s-score $s(g_1,\, g_2)$ смещается к нулю и становится все менее близким к равномерному.

При малом n_2 ($n_2 = 30,\, 50$) распределение s-score близко к равномерному. Однако уже при $n_2 = 100\, \mathbb{P}(s < 0.05) \approx 0.15$, что в 3 раза превышает соответствующее значение для равномерного распределения. Вероятность ошибки второго рода $\mathbb{P}(H_0|H_1)$, то есть вероятность признать две модели совпадающими, хотя они являются различными, для $n_2 = 100$ тогда равна 0.85 при фиксированном уровне ошибки первого рода $\alpha = \mathbb{P}(H_1|H_0) = 0.05$. При $n_2 = 300$ она снижается до 0.48, а при $n_2 = 1000$ до 0.1.

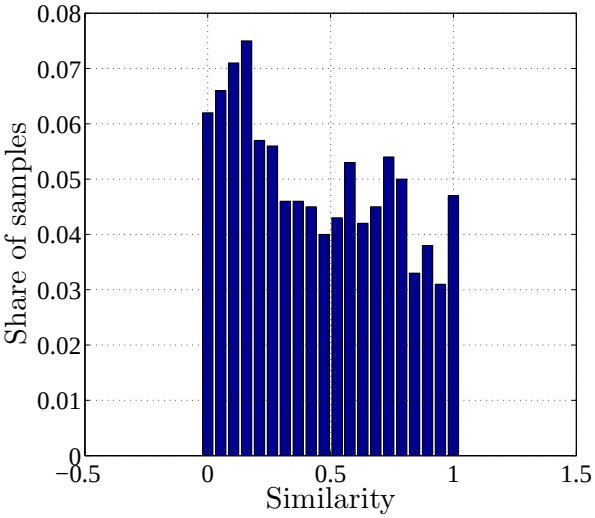
Рассмотрим теперь менее близкие модели. Иллюстрация для $\rho = 0.9$ приведена на рис. 5.19, а для случая $\rho = 0.5$ на рис. 5.20. Так для случая $\rho = 0.5$ уже при $n_2 = 100$ вероятность ошибки второго рода $\mathbb{P}(H_0|H_1) = 0.1$, что достигалось для $\rho = 0.95$ только при $n_2 = 1000$. Для случая $\rho = 0.5$ уже для $n_2 = 30$ распределение функции близости s-score отличается от равномерного. Так $\mathbb{P}(s < 0.05) > 0.3$, что дает уровень ошибки второго рода $\mathbb{P}(H_0|H_1) < 0.7$ уже для такого малого числа объектов.

Рассмотрим далее зависимость уровня ошибки второго рода $\mathbb{P}(H_0|H_1)$ при фиксированном уровне ошибки первого рода $\mathbb{P}(H_1|H_0) = \alpha = 0.05$ от косинусной близости рассматриваемых моделей для разного количества объектов во второй выборке n_2 при фиксированном $n_1 = 10000$ (см. рис. 5.21).

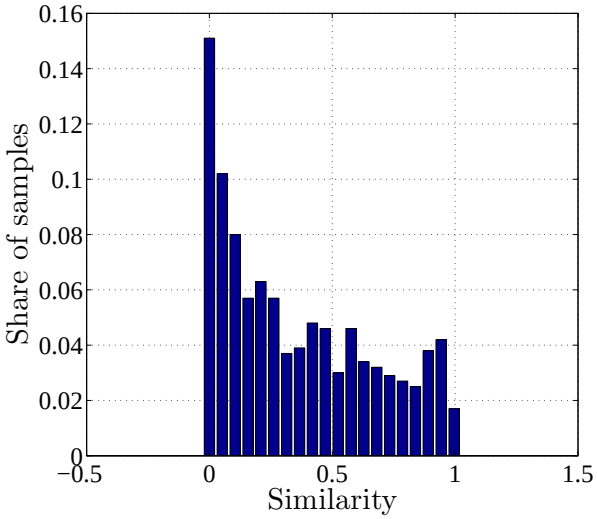
Из рис. 5.21 заключаем, что уровень ошибки второго рода растет при увеличении близости моделей и уменьшается при увеличении числа объектов во второй выборке n_2. Так для $n_2 = 30$ даже для двух моделей с $\rho = -0.5$ вероятность не различить эти модели $\mathbb{P}(H_0|H_1) = 0.1$, а при $\rho = 0.5$ эта вероятность достигает 0.7. Это означает, что если модели близки, то s-score не различает модели, а потому для предсказания классов для объектов из второй модели стоит использовать первую модель, которая, будучи смещенной, обладает тем не менее гораздо меньшей неопределенностью в оценке вектора параметров $\hat{\mathbf{w}}_1$.



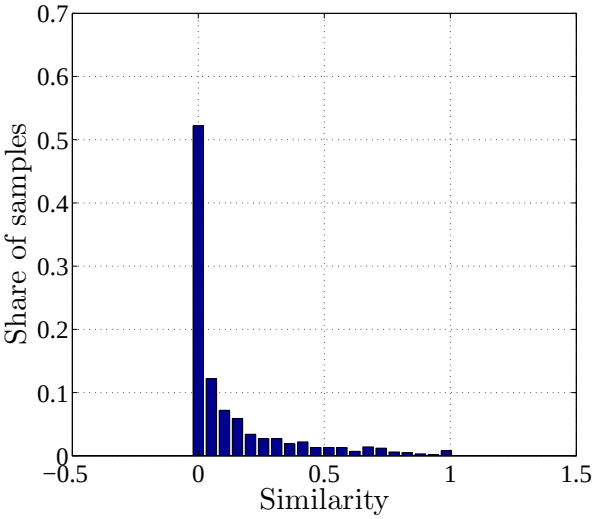
(а) $n_1 = 10000$, $n_2 = 30$



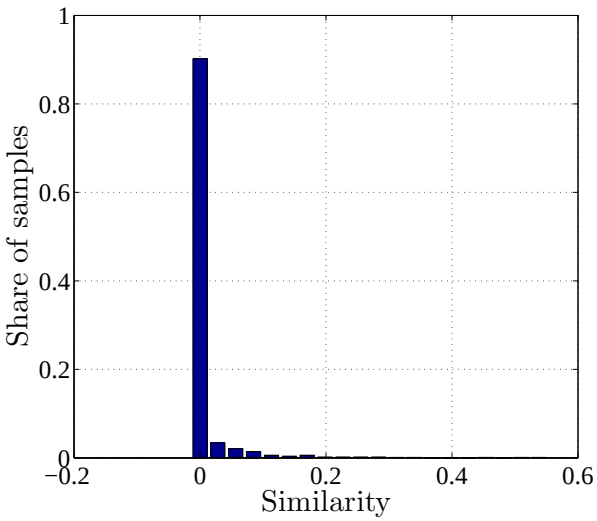
(б) $n_1 = 10000$, $n_2 = 50$



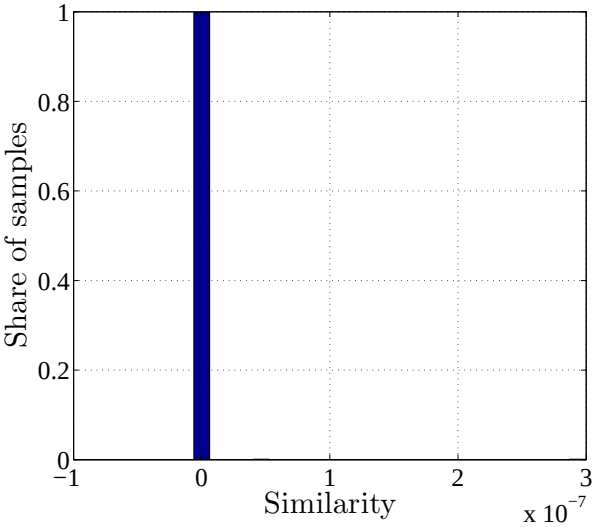
(в) $n_1 = 10000$, $n_2 = 100$



(г) $n_1 = 10000$, $n_2 = 300$

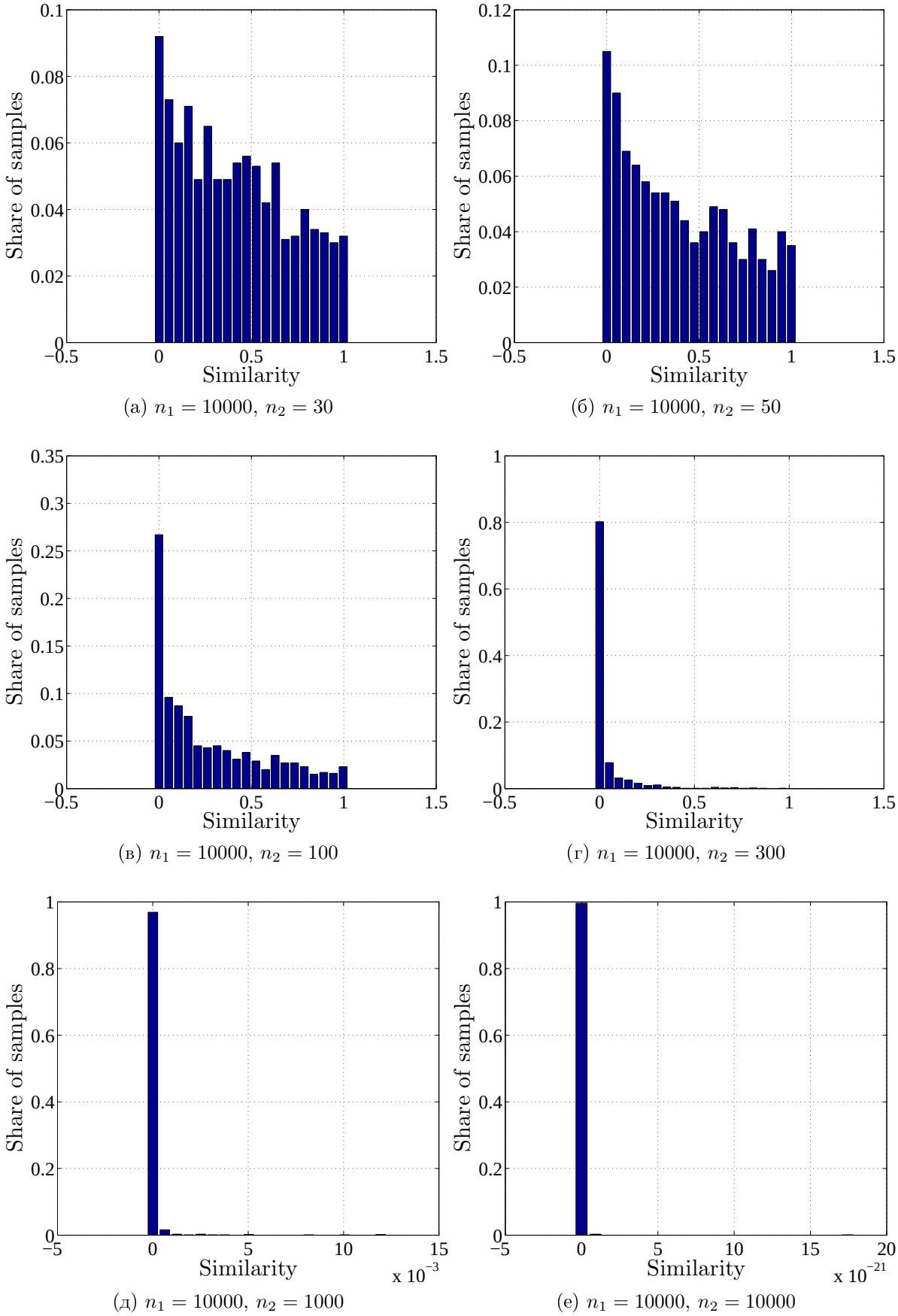


(д) $n_1 = 10000$, $n_2 = 1000$



(е) $n_1 = 10000$, $n_2 = 10000$

Рис. 5.18: Гистограмма распределения функции близости s-score $s(g_1, g_2)$ для $\rho = 0.95$ в зависимости от числа объектов во второй выборке n_2



(а) $n_1 = 10000$, $n_2 = 30$

(б) $n_1 = 10000$, $n_2 = 50$

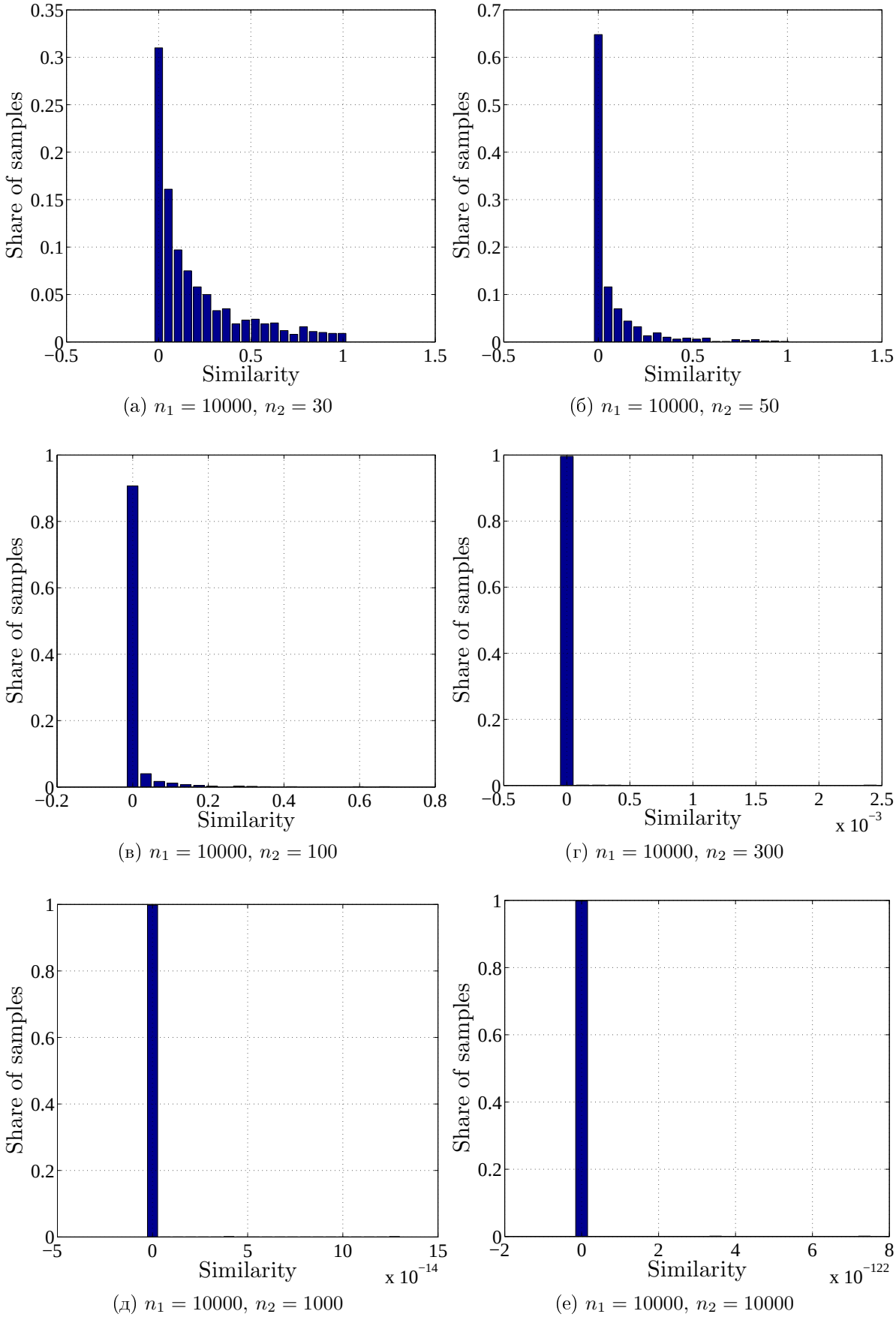
(в) $n_1 = 10000$, $n_2 = 100$

(г) $n_1 = 10000$, $n_2 = 300$

(д) $n_1 = 10000$, $n_2 = 1000$

(е) $n_1 = 10000$, $n_2 = 10000$

Рис. 5.19: Гистограмма распределения функции близости s-score $s(g_1,\ g_2)$ для $\rho = 0.9$ в зависимости от числа объектов во второй выборке n_2



(а) $n_1 = 10000$, $n_2 = 30$

(б) $n_1 = 10000$, $n_2 = 50$

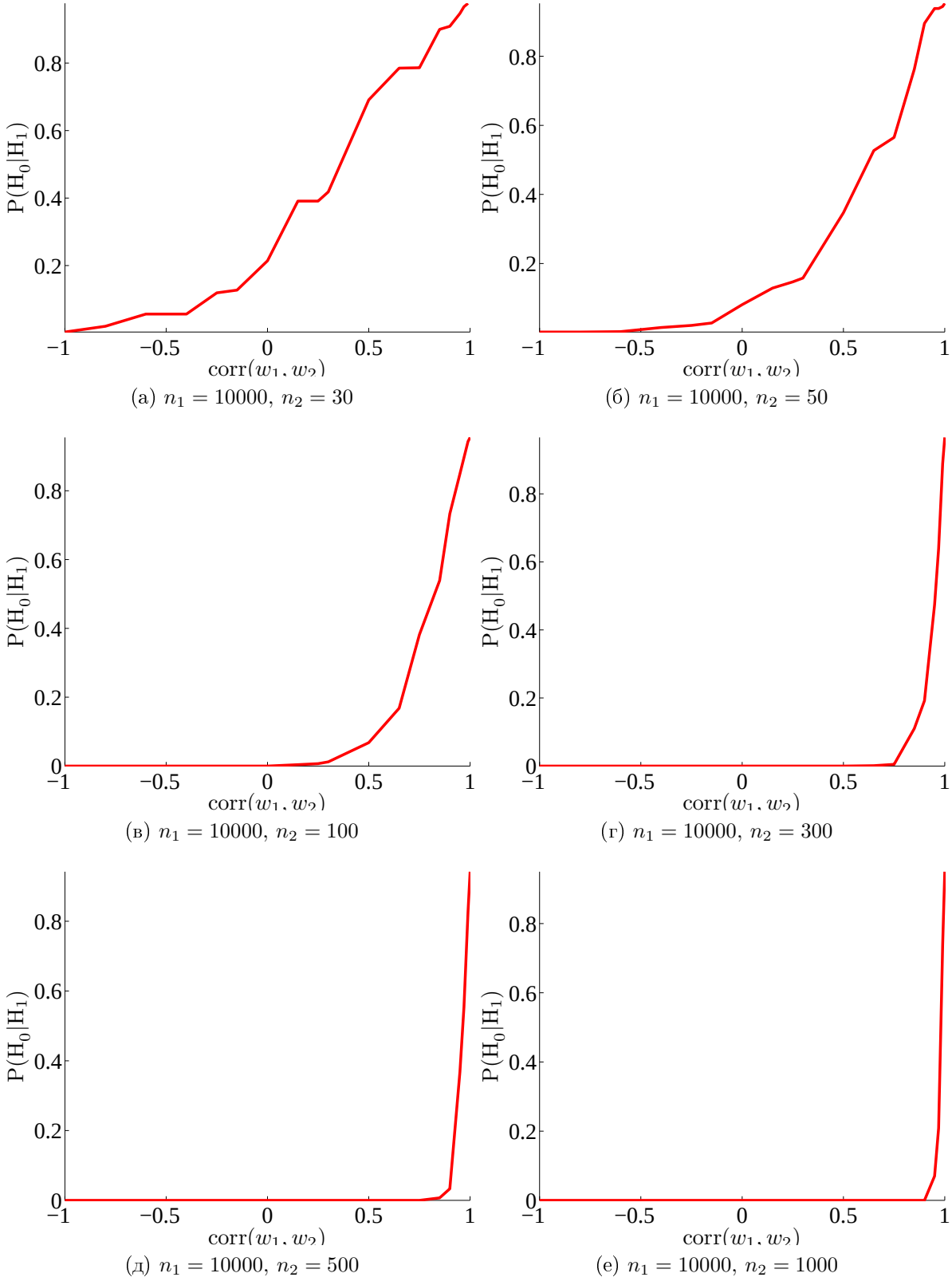
(в) $n_1 = 10000$, $n_2 = 100$

(г) $n_1 = 10000$, $n_2 = 300$

(д) $n_1 = 10000$, $n_2 = 1000$

(е) $n_1 = 10000$, $n_2 = 10000$

Рис. 5.20: Гистограмма распределения функции близости s-score $s(g_1,\ g_2)$ для $\rho = 0.5$ в зависимости от числа объектов во второй выборке n_2



(а) $n_1 = 10000$, $n_2 = 30$

(б) $n_1 = 10000$, $n_2 = 50$

(в) $n_1 = 10000$, $n_2 = 100$

(г) $n_1 = 10000$, $n_2 = 300$

(д) $n_1 = 10000$, $n_2 = 500$

(е) $n_1 = 10000$, $n_2 = 1000$

Рис. 5.21: Зависимость $\mathbb{P}(H_0|H_1)$ от косинусной близости между истинными параметрами двух моделей ρ для разных значений числа объектов во второй выборке n_2

Заключение

Основные результаты диссертационной работы заключаются в следующем.

В главе 1 введены основные понятия и определения. Рассмотрена задача двухклассовой классификации и даны определения жесткой и вероятностной модели двухклассовой классификации, а также определено решение задачи двухклассовой классификации. Приведены понятия смесей моделей и многоуровневых моделей, как типов мультимоделей для учета неоднородностей в данных. Рассмотрена модель логистической регрессии, а также задан вид априорных распределений на веса моделей в мультимодели и на вектора параметров моделей, использовавшиеся в работе.

В главе 2 принцип максимума обоснованности и понятие оптимальное модели на нем основанное. Для модели логистической регрессии рассмотрено два алгоритма максимизации обоснованности, основанных на двух разных приближениях: аппроксимации Лапласа для логарифма совместного распределения и вариационных нижних оценках на правдоподобие. Оценки максимума обоснованности для ковариационной матрицы вектора параметров модели получены в классе диагональных и общих неотрицательно определенных симметричных матриц. Доказано, однако, что недиагональная оценка ковариационной матрицы вектора параметров логистической модели является асимптотически вырожденной и не показывает наличие мультиколлинеарности между признаками. Получена схема оптимального комбинирования признаков при известной структуре зависимостей между ними и дисперсии шума. Для детектирования и учета мультиколлинеарности предложен метод комбинирования признаков.

В главе 3 рассмотрена задача обучения смесей моделей и многоуровневых моделей. Приведен вариационный EM-алгоритм для обучения смеси моделей при фиксированных значениях гиперпараметров. Предложен алгоритм совместного обучения и оптимизации смеси моделей, основанный на аппроксимации Лапласа и вариационном EM-алгоритме.

В главе 4 рассмотрена задача сравнения моделей как задача определения похожести апостериорных распределений параметров моделей. Введено определение корректной функции сходства, удовлетворяющей техническим требованиям, а также характеристическому требованию для данной задачи об неотличимости малоинформативного распределения от любого другого. Показано, что функции сходства, порожденные существующими расстояниями между распределениями, включая дивергенции Брегмана и f-дивергенции, не являются корректными. Предложена функция сходства, удовлетворяющая требованиям корректности, определенная в том числе и для пары распределений, определенных на разных носителях. Показано, что предложенная функция сходства обладает свойством монотонности. Получено асимптотическое распределение значений функции сходства в условиях истинности гипотезы о совпадении моделей. На основании полученного асимптотического распределения предложен метод статистического сравнения моделей. Введено понятие адекватной мульти-

модели, все модели в которой попарно статистически различимы. Предложены методы построения адекватных смесей моделей и многоуровневых моделей. Получены верхняя и нижняя оценки на максимальное число моделей в адекватной мультимодели.

В главе 5 произведен анализ свойств и границ применимости предлагаемых методов. Предлагаемые методы построения адекватных мультимоделей тестировались на синтетических данных, соответствующих разным условиям генерации выборки. Реализован программный комплекс, который позволяет классифицировать объекты на два класса путем построения, оптимизации и обучения мультимодели, ее прореживания для получения адекватной мультимодели. В рамках программного комплекса реализован модуль детектирования и комбинирования мультиколлинеарных признаков. Результаты классификации после комбинирования признаков сравнивались с результатами известных методов отбора признаков, основанными на максимизации обоснованности. Предлагаемый метод комбинирования признаков при наличии мультиколлинеарности показал лучшие результаты как на синтетических. так и на реальных данных. При отсутствии мультиколлинераности значимого ухудшения качества также не происходит. Результаты двухклассовой классификации с помощью построенных адекватных мультимоделей сравнивались с результатами классификации с помощью оптимальных обученных мультимоделей с и без комбинирования признаков на нескольких наборах данных из репозитория UCI. Предлагаемый метод построения адекватных мультимоделей позволяет не только построить более интерпретируемую мультимодель, но и в большинстве случаев дает лучшее качество классификации.

Список основных обозначений

Матрицы обозначены заглавными полужирными буквами, векторы — полужирными строчными буквами, множества — заглавными буквами.

$\mathbf{x}_i$ – вектор признакового описания i – го объекта выборки

$\mathbf{X}$ – матрица, содержащая признаковое описание объектов

$\mathbf{f}_j$ – вектор значений j – го признака

m – количество объектов в выборке

n – число признаков в признаковом описании

K – число моделей в мультимодели

$\mathbf{w}$ – вектор параметров модели

$\boldsymbol{\pi}$ – вектор весов моделей в смеси моделей

$\lambda_{\max}(\cdot)$ – максимальное собственное значение матрицы

$\lambda_{\min}(\cdot)$ – минимальное собственное значение матрицы

f_k – k-я модель в мультимодели

$N(\cdot|\mathbf{m}, \boldsymbol{\Sigma})$ – нормальное распределение со средним $\mathbf{m}$ и ковариационной матрицей $\boldsymbol{\Sigma}$

Ω_k – часть признакового пространства, где действует модель k в многоуровневой модели

$\mathcal{I}_k$ – множество индексов объектов выборки, которые лежат в области действия модели с номером k в многоуровневой модели

$\mathbf{F}$ – матрица значений факторов

$\mathbf{G}$ – матрица разложения наблюдаемых признаков по скрытым факторам

t_{kl} – достигаемый уровень значимости в условиях истинности гипотезы о совпадении моделей с номерами k и l

$\mathbf{T}$ – матрица парных достигаемых уровней значимости в условиях истинности гипотезы о совпадении моделей

$\sigma(\cdot)$ – сигма-функция

$\mathbb{R}$ — множество действительных чисел

$\mathbb{R}^+$ — множество неотрицательных действительных чисел

$\mathcal{M}_{diag}$ – множество диагональных матриц с неотрицательными элементами на диагонали

$\mathcal{M}_{full}$ – множество неотрицательно определенных симметричных мартриц

$D_{KL}(\cdot, \cdot)$ – дивергенция Кульбака-Лейблера

$D_{JS}(\cdot, \cdot)$ – расстояние Дженсона-Шеннона

$D_H(\cdot, \cdot)$ – расстояние Хеллингера

$D_B(\cdot, \cdot)$ – расстояние Бхаттачарайа

$D_F(\cdot, \cdot)$ – дивергенция Брегмана, порожденная функцией F

$\tilde{D}_F(\cdot, \cdot)$ – симметризованная дивергенция Брегмана, порожденная функцией F

$d_f(\cdot, \cdot)$ – f – дивергенция, порожденная функцией f

$s_{KL}(\cdot, \cdot)$ – функция сходства, порожденная дивергенцией Кульбака-Лейблера

$s_{JS}(\cdot, \cdot)$ – функция сходства, порожденная расстоянием Дженсона-Шеннона

$s_H(\cdot, \cdot)$ – функция сходства, порожденная расстоянием Хеллингера

$s_B(\cdot,\,\cdot)$ – функция сходства, порожденная расстоянием Бхаттачарайа
$s_F(\cdot,\,\cdot)$ – функция сходства, порожденная дивергенцией Брегмана
$\tilde{s}_F(\cdot,\,\cdot)$ – функция сходства, порожденная симметризованной дивергенцией Брегмана
$s_f(\cdot,\,\cdot)$ – функция сходства, порожденная f – дивергенцией
$s(\cdot,\,\cdot)$ – предлагаемая функция сходства s-score
$I_{KL}(g_2|g_1)$ – KL-информативность g_2 относительно g_1
$s_{KL}(\cdot|\cdot)$ – несимметричное KL-сходство
$s_{tr}(\cdot,\,\cdot)$ – тривиальная функция сходства
$g|_A(\cdot)$ – сужение распределения g на множество A
$|\,\cdot\,|$ — число элементов в множестве для счетного множества, объем множества для несчетного множества
$[i = j]$ — индикаторная функция
$\boldsymbol{O}$ – нулевая квадратная матрица
$\mathbf{e}_n$ – вектор размера n, содержащий единицы

Список иллюстраций

Список таблиц

Литература

1. *Стрижов В. В.* Функция ошибки в задачах восстановления регрессии // Заводская лаборатория, 2013. Т. 79. №. 5. С. 65–73.
2. *Diop A. et al.* Maximum likelihood estimation in the logistic regression model with a cure fraction // Electronic Journal of Statistics, 2011. Vol. 5. Pp. 460–483.
3. *Fahrmeir L., Kaufmann H.* Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models // The Annals of Statistics, 1985. Pp. 342–368.
4. *Nurunnabi A. A. M. et al.* Outlier Detection in Logistic Regression // Multidisciplinary Computational Intelligence Techniques: Applications in Business, Engineering, and Medicine: Applications in Business, Engineering, and Medicine, 2012. P. 257.
5. *Pregibon D.* Logistic regression diagnostics // The Annals of Statistics, 1981. Pp. 705–724.
6. *Breiman L.* Bagging predictors // Machine learning, 1996. Vol. 24. No. 2. Pp. 123–140.
7. *Katrutsa A. M., Strijov V. V.* Stress test procedure for feature selection algorithms // Chemometrics and Intelligent Laboratory Systems, 2015. Vol. 142. Pp. 172–183.
8. *Лужбин А. А.* К вопросу о разработке статистических моделей вероятности дефолта в условиях дефицита данных // Известия Санкт-Петербургского университета экономики и финансов, 2013. Т. 84. №. 6. С. 114–117.
9. *Yuksel S. E., Wilson J. N., Gader P. D.* Twenty years of mixture of experts // Neural Networks and Learning Systems, IEEE Transactions on, 2012. Vol. 23. No. 8. Pp. 1177–1193.
10. *Margineantu D. D., Dietterich T. G.* Pruning adaptive boosting // ICML, 1997. Vol. 97. Pp. 211–218.
11. *Dietterich T. G.* An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization // Machine learning, 2000. Vol. 40. No. 2. Pp. 139–157.
12. *Martınez-Munoz G., Suárez A.* Aggregation ordering in bagging // Proc. of the IASTED International Conference on Artificial Intelligence and Applications, 2004. Pp. 258–263.
13. *Dai Q., Han X.* An efficient ordering-based ensemble pruning algorithm via dynamic programming // Applied Intelligence, 2015. Pp. 1–15.
14. *Martínez-Muñoz G., Suárez A.* Pruning in ordered bagging ensembles // Proceedings of the 23rd international conference on Machine learning, ACM, 2006. Pp. 609–616.

15. *Martínez-Muñoz G., Suárez A.* Using boosting to prune bagging ensembles // Pattern Recognition Letters, 2007. Vol. 28. No. 1. Pp. 156–165.
16. *Martinez-Muoz G., Hernández-Lobato D., Suarez A.* An analysis of ensemble pruning techniques based on ordered aggregation // Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2009. Vol. 31. No. 2. Pp. 245–259.
17. *Bakker B., Heskes T.* Clustering ensembles of neural network models // Neural networks, 2003. Vol. 16. No. 2. Pp. 261–269.
18. *Giacinto G., Roli F.* An approach to the automatic design of multiple classifier systems // Pattern recognition letters, 2001. Vol. 22. No. 1. Pp. 25–33.
19. Zhou Z. H., Tang W. Selective ensemble of decision trees // Springer Berlin Heidelberg: Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 2003. Pp. 476–483.
20. Zhou Z. H., Wu J., Tang W. Ensembling neural networks: many could be better than all // Artificial intelligence, 2002. Vol. 137. No. 1. Pp. 239–263.
21. *Bishop C. M.* Pattern recognition and machine learning. // Springer, 2006.
22. *Bishop C. M., Nasrabadi N. M.* Pattern recognition and machine learning. // Journal of electronic imaging, 2007. Vol. 16. No. 4.
23. *Verlinde P., Cholet G.* Comparing decision fusion paradigms using k-NN based classifiers, decision trees and logistic regression in a multi-modal identity verification application // Proc. Int. Conf. Audio and Video-Based Biometric Person Authentication (AVBPA),1999. Pp. 188–193.
24. *Gelman A., Hill J.* Data analysis using regression and multilevel/hierarchical models // Cambridge University Press, 2006.
25. *Yap B. W., Ong S. H., Husain N. H. M.* Using data mining to improve assessment of credit worthiness via credit scoring models // Expert Systems with Applications, 2011. Vol. 38. No. 10. Pp. 13274–13283.
26. *Zakrzewska D.* On integrating unsupervised and supervised classification for credit risk evaluation // Information technology and control, 2015. Vol. 36. No. 1. Pp. 98–102.
27. *Hsieh N. C.* Hybrid mining approach in the design of credit scoring models // Expert Systems with Applications, 2005. Vol. 28. No. 4. Pp. 655–665.
28. *Harris T.* Credit scoring using the clustered support vector machine // Expert Systems with Applications, 2015. Vol. 42. No. 2. Pp. 741–750.
29. *Palmer J. et al.* Variational EM algorithms for non-Gaussian latent variable models // Advances in neural information processing systems. – 2005. – C. 1059-1066.
30. *Freund Y., Schapire R. E.* A desicion-theoretic generalization of on-line learning and an application to boosting // Springer Berlin Heidelberg: Computational learning theory, 1995. Pp. 23–37.

31. *Friedman J. et al.* Additive logistic regression: a statistical view of boosting // The annals of statistics, 2000. Vol. 28. No. 2. Pp. 337–407.
32. *Oh I. S., Lee J. S., Moon B. R.* Hybrid genetic algorithms for feature selection. // IEEE transactions on pattern analysis and machine intelligence, 2004. Vol. 26. No. 11. Pp. 1424–1437.
33. *Siddiqi N.* Credit risk scorecards: developing and implementing intelligent credit scoring // Wiley, 2006.
34. *Hosmer D. W., Lemeshow S.* Applied logistic regression // A Wiley-Interscience Publication, 2000.
35. *Khalili A.* An Overview of the New Feature Selection Methods in Finite Mixture of Regression Models // Journal of Iranian Statistical Society, 2011. Vol. 10. No. 2. Pp. 201–235.
36. *Chandrashekar G., Sahin F.* A survey on feature selection methods // Computers & Electrical Engineering, 2014. Vol. 40. No. 1. Pp. 16–28.
37. *Rodriguez-Lujan I. et al.* Quadratic programming feature selection // Journal of Machine Learning Research, 2010. Vol. 11. Pp. 1491–1516.
38. *Gheyas I. A., Smith L. S.* Feature subset selection in large dimensionality domains // Pattern recognition, 2010. Vol. 43. No. 1. Pp. 5–13.
39. *Леонтьева Л. Н.* Последовательный выбор признаков при восстановлении регрессии // Машинное обучение и анализ данных, 2012. Т. 1. № 3. С. 335–346.
40. *Oreski S., Oreski G.* Genetic algorithm-based heuristic for feature selection in credit risk assessment // Expert systems with applications, 2014. Vol. 41. No. 4. Pp. 2052–2064.
41. *Vergara J. R., Estévez P. A.* A review of feature selection methods based on mutual information // Neural Computing and Applications, 2014. Vol. 24. No. 1. Pp. 175–186.
42. *Motrenko A., Strijov V., Weber G. W.* Bayesian sample size estimation for logistic regression.
43. Данные по немецким потребительским кредитам. http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/, 2000.
44. Данные по венгерским потребительским кредитам. http://svn.code.sf.net/p/mlalgorithms/code/Aduenko2013BsThesis/data/cleared_
45. Данные конкурса Интернет-математика 2009. http://imat2009.yandex.ru/.
46. Данные по сердечным заболеваниям в Южной Африке. http://svn.code.sf.net/p/mlalgorithms/code/Aduenko2013BsThesis/data/SAHD.cs
47. Данные по качеству белого вина. http://archive.ics.uci.edu/ml/datasets/Wine+Qu
48. *Ling C. X., Huang J., Zhang H.* AUC: a statistically consistent and more discriminating measure than accuracy // International joint Conference on artificial intelligence, 2003. Vol. 18. Pp. 519–526.

49. *Van den Noortgate W., De Boeck P., Meulders M.* Cross-classification multilevel logistic models in psychometrics // Journal of Educational and Behavioral Statistics, 2003. Vol. 28. No. 4. Pp. 369–386.
50. Frigyik B. A., Srivastava S., Gupta M. R. Functional Bregman divergence and Bayesian estimation of distributions // IEEE Transactions on Information Theory, 2008. Vol. 54. No. 11. Pp. 5130–5139.
51. Petz D. Bregman divergence as relative operator entropy // Acta Mathematica Hungarica, 2007. Vol. 116. No. 1–2. Pp. 127–131.
52. Zhang Z. et al. Similarity search on bregman divergence: Towards non-metric indexing // Proceedings of the VLDB Endowment, 2009. Vol. 2. No. 1. Pp. 13–24.
53. Basseville M. Divergence measures for statistical data processing—An annotated bibliography // Signal Processing, 2013. Vol. 93. No. 4. Pp. 621–633.
54. Veyrat-Charvillon N., Standaert F. X. Mutual information analysis: how, when and why? // Cryptographic Hardware and Embedded Systems-CHES 2009. Springer Berlin Heidelberg, 2009. Pp. 429–443.
55. Kailath T. The divergence and Bhattacharyya distance measures in signal selection // IEEE transactions on communication technology, 1967. Vol. 15. No. 1. Pp. 52–60.
56. Weinstein E., Feder M., Oppenheim A. V. Sequential algorithms for parameter estimation based on the Kullback-Leibler information measure // IEEE Transactions on Acoustics, Speech, and Signal Processing, 1990. Vol. 38. No. 9. Pp. 1652–1654.
57. Wang C., Chang H. H., Boughton K. A. Kullback–Leibler information and its applications in multi-dimensional adaptive testing // Psychometrika, 2011. Vol. 76. No. 1. Pp. 13–39.
58. *Andersen E. B.* Asymptotic properties of conditional maximum-likelihood estimators // Journal of the Royal Statistical Society. Series B (Methodological), 1970. Pp. 283–301.
59. *Strasser H.* The asymptotic equivalence of Bayes and maximum likelihood estimation // Journal of Multivariate Analysis, 1975. Vol. 5. No. 2. Pp. 206–226.
60. *Fisher T. J., Sun X.* Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix // Computational Statistics & Data Analysis, 2011. Vol. 55. No. 5. Pp. 1909–1918.
61. *Huang J. Z. et al.* Covariance matrix selection and estimation via penalised normal likelihood // Biometrika, 2006. Vol. 93. No. 1. Pp. 85–98.
62. *Ledoit O., Wolf M.* A well-conditioned estimator for large-dimensional covariance matrices // Journal of multivariate analysis, 2004. Vol. 88. No. 2. Pp. 365–411.
63. *Liechty J. C., Liechty M. W., Müller P.* Bayesian correlation estimation // Biometrika, 2004. Vol. 91. No. 1. Pp. 1–14.

64. *A. A. Aduenko , V. V. Strijov* Multimodelling and Object Selection for Banking Credit Scoring // 20th Conference of the International Federation of Operational Research Societies. — Barcelona: 2014. — P. 136.
65. *A. A. Aduenko , V. V. Strijov* Multimodelling and Model Selection in Bank Credit Scoring // 27th European Conference for Operational Research. — Glasgow: 2015. — P. 273.
66. *А. А. Адуенко, В. В. Стрижов* Анализ пространства параметров в задачах выбора мультимоделей // Математические методы распознавания образов ММРО-17. Тезисы докладов 17-й Всероссийской конференции с международным участием. — г. Светлогорск, Калининградская область: Торус пресс, 2015. С. 10–11.
67. *А. А. Адуенко, В. В. Стрижов* Анализ пространства параметров в задачах выбора мультимоделей // Интеллектуализация обработки информации ИОИ-2016. Тезисы докладов 11-й Международной конференции. — Москва, Россия-Барселона, Испания: Торус пресс, 2016. С. 10–11.
68. *Адуенко А.А.* Выбор признаков и шаговая логистическая регрессия для задачи кредитного скоринга // Машинное обучение и анализ данных, 2012. № 3. С. 279-291.
69. *А. А. Адуенко, А. А. Кузьмин, В. В. Стрижов* Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Известия ТулГУ, 2012. № 3. С. 119-131.
70. *А. А. Адуенко, В. В. Стрижов* Алгоритм оптимального расположения названий коллекции документов // Программная инженерия, 2013. № 3. С. 21–25.
71. *А. В. Иванова, А. А. Адуенко, В. В. Стрижов* Алгоритм построения логических правил при разметке текстов // Программная инженерия, 2013. № 6. С. 41–47.
72. *А. А. Адуенко, Н. И. Амелькин* О предельных движениях волчка с внутренней диссипацией в однородном поле тяжести // Труды МФТИ, 2013. № 18(2). С. 126-133.
73. *А. А. Кузьмин, А. А. Адуенко, В. В. Стрижов* Тематическая классификация тезисов крупной конференции с использованием экспертной модели // Информационные технологии, 2014. № 6. С. 22-26.
74. *А. А. Адуенко, Н. И. Амелькин* Асимптотические свойства движений тяжелого волчка с внутренней диссипацией // ПММ, 2014. Т. 78. Вып. 1. С. 13-28.
75. *А. А. Адуенко, В. В. Стрижов* Совместный выбор объектов и признаков в задачах многоклассовой классификации коллекции документов // Инфокоммуникационные технологии, 2014. № 1. С. 47–53.
76. *А. А. Адуенко, А. С. Василейский, А. И. Карелов, И. А. Рейер, К. В. Рудаков, В. В. Стрижов* Алгоритмы выделения и совмещения устойчивых отражателей на спутниковых снимках // Компьютерная оптика, 2015. Т. 39. Вып. 4. С. 622–630.

77. *А. А. Адуенко, Н. И. Амелькин* О резонансных вращениях маятника с вибрирующим подвесом // ПММ. 2015. Т. 79. Вып. 6. С. 756–767.
78. *Moerbeek M., Van Breukelen G. J. P., Berger M. P. F.* Optimal experimental designs for multilevel logistic models // Journal of the Royal Statistical Society: Series D (The Statistician), 2001. Vol. 50. No. 1. Pp. 17–30.
79. *Grün B., Leisch F.* Fitting finite mixtures of generalized linear regressions in R // Computational Statistics & Data Analysis, 2007. Vol. 51. No. 11. Pp. 5247–5252.
80. *Ge Y., Jiang W.* On consistency of Bayesian inference with mixtures of logistic regression // Neural Computation, 2006. Vol. 18. No. 1. Pp. 224–243.
81. *Muthén B., Shedden K.* Finite mixture modeling with mixture outcomes using the EM algorithm // Biometrics, 1999. Vol. 55. No. 2. Pp. 463–469.
82. *Follmann D. A., Lambert D.* Identifiability of finite mixtures of logistic regression models // Journal of Statistical Planning and Inference, 1991. Vol. 27. No. 3. Pp. 375–381.
83. *Paleologo G., Elisseeff A., Antonini G.* Subagging for credit scoring models // European Journal of Operational Research, 2010. Vol. 201. No. 2. Pp. 490–499.
84. *Gibbs M. N., MacKay D. J. C.* Variational Gaussian process classifiers // IEEE Transactions on Neural Networks, 2000. Vol. 11. No. 6. Pp. 1458–1464.
85. *Blei D. M., Kucukelbir A., McAuliffe J. D.* Variational inference: A review for statisticians // arXiv preprint arXiv:1601.00670, 2016.
86. *Bonilla E. V., Steinberg D., Reid A.* Extended and unscented kitchen sinks // International Conference on Machine Learning. – 2016.
87. *Boyd S., Vandenberghe L.* Convex optimization. Cambridge university press, 2004.
88. *Ширяев А. Н.* Вероятность. Элементарная теория вероятностей. Математические основания. Предельная теорема: Учебник для студ. вузов. // М.: МЦНМО. – 2004.
89. *MacKay D. J. C.* Bayesian methods for adaptive models // California Institute of Technology, 1992.
90. *Mackay D. J, C.* The evidence framework applied to classification networks // Neural computation, 1992. Vol. 4. No. 5. Pp. 720-736.
91. *Morales-Enciso S., Branke J.* Tracking global optima in dynamic environments with efficient global optimization // European Journal of Operational Research, 2015. Vol. 242. No. 3. Pp. 744–755.
92. *Tolles J., Meurer W. J.* Logistic regression: relating patient characteristics to outcomes // Jama, 2016. Vol. 316. No. 5. Pp. 533–534.
93. *Bagley S. C., White H., Golomb B. A.* Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain // Journal of clinical epidemiology, 2001. Vol 54. No. 10. Pp. 979–985.

94. *Supriya B. N. et al.* Twitter Sentiment Analysis Using Binary Classification Technique // International Conference on Nature of Computation and Communication. Springer International Publishing, 2016. Pp. 391–396.
95. *Joshi B. et al.* On binary reduction of large-scale multiclass classification problems // International Symposium on Intelligent Data Analysis. Springer International Publishing, 2015. Pp. 132–144.
96. *Tax D. M. J., Duin R. P. W.* Using two-class classifiers for multiclass classification // Proceedings of 16th IEEE International Conference on Pattern Recognition, 2002. Vol. 2. Pp. 124–127.
97. *Liu Y., Zheng Y. F.* One-against-all multi-class SVM classification using reliability measures // Proceedings. IEEE International Joint Conference on Neural Networks, 2005. Vol. 2. Pp. 849–854.
98. *Rifkin R., Klautau A.* In defense of one-vs-all classification // Journal of machine learning research, 2004. Vol. 5. Pp. 101-141.
99. *Berkhin P.* A survey of clustering data mining techniques // Grouping multidimensional data. Springer Berlin Heidelberg, 2006. Pp. 25–71.
100. *Loh W. Y., Shih Y. S.* Split selection methods for classification trees // Statistica sinica, 1997. Pp. 815–840.
101. *Chamroukhi F.* Piecewise regression mixture for simultaneous functional data clustering and optimal segmentation // arXiv preprint arXiv:1312.6974, 2013.
102. *Herzenstein M., Andrews R. L., Dholakia U., Lyandres E.* The democratization of personal consumer loans? Determinants of success in online peer-to-peer lending communities // Boston University School of Management Research Paper, 2008. No. 2009-14.
103. *Hoffman M. D. et al.* Stochastic variational inference // Journal of Machine Learning Research, 2013. Vol. 14. No. 1. Pp. 1303–1347.
104. *Wang C., Blei D. M.* Variational inference in nonconjugate models // Journal of Machine Learning Research, 2013. Vol. 14. Pp. 1005-1031.
105. *Jiang S. et al.* An improved K-nearest-neighbor algorithm for text categorization // Expert Systems with Applications, 2012. Vol. 39. No. 1. Pp. 1503–1509.