Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Адуенко Александр Александрович

Топологический анализ пространства параметров в задачах выбора мультимоделей

010958 — Прикладная информатика

Выпускная квалификационная работа магистра

Научный руководитель:
д. ф.-м. н. Стрижов Вадим Викторович

Москва
2015

Содержание

Аннотация

Банковские скоринговые модели используются для оценки вероятности дефолта заемщика по кредитному опроснику. Для построения скоринговой модели нужно выбрать набор информативных объектов (клиентских записей), который позволит получить несмещенную оценку параметров модели. Предлагается алгоритм отбора объектов, основанный на анализе ковариационной матрицы оценок параметров. Вводится специфичность объекта, которая показывает, является ли объект выбросом. Для случая плохо определенной ковариационной матрицы предлагается эмпирическая специфичность. Алгоритм проиллюстрирована на четырех наборах данных из репозитория UCI: данные по немецким потребительским кредитам, данные по сердечным заболеваниям в Южной Африке, данные по качеству вина и данные по расположению белка в клетке. Вычислительный эксперимент показывает статистическую значимость полученного улучшения качества на тестовой выборке для всех четырех наборов данных. Предлагаемый метод сравнивается с тремя широко используемыми методами фильтрации выбросов: остатки, основанные на дисперсии, остатки Пирсона и байесовы остатки,–на четырех наборах данных из репозитория UCI, а также на синтетических данных, имеющих кластеризованные и некластеризованные выбросы. Метод демонстрирует приемлемое качество классификации даже для данных, имеющих до 30–40% выбросов. Данные зачастую могут быть кластеризованы, что означает, что зависимость между признаками и вероятностью дефолта может отличаться для разных кластеров. Для учета неоднородности данных предлагается использовать мультимодели и многоуровневые модели. Для отбора объектов для мультимоделей и многоуровневых моделей предлагается обобщение метода, основанного на специфичности. Для решения задачи отбора признаков рассматривается применение обоснованности модели вместе с методом Белсли.

Ключевые слова: *потребительский кредит, кредитный скоринг, вероятность дефолта, мультимодели, многоуровневые модели, отбор объектов и признаков, фильтрация выбросов.*

Введение

Актуальность темы. Задача кредитного скоринга становится все более актуальной вместе с распространением и широким использованием разного рода кредитов, особенно потребительских. Если кредитные риски крупных компаний оцениваются международными рейтинговыми агенствами на основании публичной отчетности, а потому решение о выдаче кредита и процентной ставке принять относительно легко, то никаких общепризнанных данных о «кредитном рейтинге» отдельного заемщика нет. А поэтому решение о выдаче кредита и ставке принимается либо экспертным путем, либо с помощью некоторой скоринговой системы. Под скоринговой системой подразумевается автоматизированная система, которая по предоставленным заемщиком данным оценивает вероятность дефолта по кредиту.

В последние годы нарастающую популярность приобретают сервисы равноправного (peer-to-peer) кредитования. Такие сервисы позволяют связать заемщика и кредитора напрямую без использования посредника в лице банков. Обороты такого рода кредитования существенно выросли с момента старта первых платформ и составляют уже несколько миллиардов долларов в год [1]. Сервис равноправного кредитования представляет из себя платформу, в которой может зарегистрироваться любой пользователь и ссудить кому-либо деньги, либо подать заявку на заем. Один из наиболее популярных сервисов в Великобритании, занимающийся равноправным кредитованием Zopa, обработал за время своего существования заемы на общую сумму более миллиарда долларов [1]. В Великобритании существуют и другие крупные подобные сервисы: Rates Settler, Funding Circle и другие. Однако подобные сервисы популярны и в других странах: LendingClub и Prosper в Соединенных Штатах Америки, БезБанка и ФинГуру в России и другие. В традиционном потребительском кредитовании заемщик предоставляет банку информацию о своем финансовом положении, а банк использует эту информацию для принятия решения о выдаче кредита либо экспертным путем, либо с помощью скориноговой модели. При этом клиент банка, имеющий в банке депозит, никак не связан с клиентом, подавшим заявку на кредит, а риски невозврата берет на себя банк. При этом процент по заему обычно меняется в достаточно узких границах [2]. В связи с отсутствием посредника в лице банка в равноправном кредитовании требуется оценивать риски каждого заема в отдельности и в соответствии с этими рисками определять приемлемый процент по заему с таким риском. Некоторые платформы предлагают собственные скоринговые системы для оценки риска каждого заема (например, российская ФинГуру [3]). От качества этих скоринговых систем во многом может зависеть популярность сервиса.

Отметим, что при создании скоринговых систем возникает несколько основных проблем. Во-первых, требуется выделить информацию, на основании которой скоринговая система будет оценивать заявку заемщика. Список потенциальных вопросов очень широк. Однако большая часть собранных данных не коррелирует с текущей платежеспособностью, да и правдивость ответов на многие сложно проверить. В случае с равноправным кредитованием последнее замечание имеет особенную важность. В связи с этим возникает проблема отбора важных данных — признаков, которые с одной стороны коррелируют с платежеспособностью, а с другой стороны — легко проверить на подлинность. Так в случае с равноправным кредитованием такими признаками может служить кредитная история заемщика в рамках платформы.

Второй проблемой является выбор информативного множества клиентов. Банковские клиентские базы содержат миллионы записей, часть из которых имеют ошибки. Невозврат кредита может быть обусловлен не какими-то причинами, имевшимися на момент получения кредита, а чем-то произошедшим после, что было трудно или невозможно предсказать. Данные таких клиентов не нужно использовать при построении скоринговой модели, поскольку они ухудшат прогноз [4].

Третьей проблемой является неоднородность данных. Для регионов с равномерно высокими доходами у населения можно предположить, что уровень дохода не столь важен для возврата кредита, как в регионах с низким средним доходом и высоким имущественным расслоением. В работе решается задача разделения объектов на однородные совокупности, определения их числа и построения своей модели для каждой совокупности. Учет неоднородности порождает новую проблему. Те объекты, которые считались выбросами при использовании одной модели логистической регрессии, могут перестать быть таковыми после ее усложнения и рассмотрения, например, мультимоделей и многоуровневых моделей. То же может произойти и с признаками: часть признаков, которые были нерелевантны для всей совокупности объектов, могут быть коррелированы с платежеспособностью для некоторого подмножества объектов. Для равноправного кредитования данная проблема также актуальна. Например, если сервис, который работал в рамках одной страны, становится международным, возможно, что скоринговая модель, работавшая в рамках одной страны, может иметь низкое качество прогнозов для клиентов из другой страны.

Альтернативой скоринговой системе является использование экспертов для принятия решений по кредитам. Ясно, что в случае с равноправным кредитованием использование экспертов затруднено, поскольку информация о заемщике лимитирована, а личной встречи с ним не происходит. В случае с банковскими кредитами применение экспертов (банковских работников) оправдано, однако может оказаться слишком затратным. За первые три месяца текущего года по данным ЦБ РФ было выдано более триллиона рублей кредитов [5]. Общий накопленный объем кредитов превышает 10 триллионов рублей, а просроченная задолженность превышает 750 миллиардов. Поэтому улучшение показателя возврата даже на несколько процентов позволит избежать убытков на десятки миллиардов рублей. Внедрение скоринговых систем, которые автоматически проводят отбор объектов и информативных признаков, а также строят модели данных, в банковском кредитовании оправдано.

Цель работы. Построить алгоритм классификации объектов, определяющий требуемое число моделей для описания данных и их параметры, а также проводящий фильтрацию выбросов и отбор информативных признаков.

Методы исследований. При построении алгоритма использовались методы оценки характеристик распределения по выборке, проверки гипотез о виде распределения, методы обработки категориальных признаков. Для программной реализации разработанного алгоритма использовалась среда MATLAB.

Научная новизна.

- Разработан алгоритм отбора признаков, основанный на применении обоснованности модели для оценки полной ковариационной матрицы параметров.

- Разработан алгоритм отбора выбросов для модели логистической регрессии.
- Разработано обобщение алгоритма отбора выбросов на случай мультимоделей и многоуровневых моделей.

Практическая ценность. Разработан программный модуль, который

- фильтрует выбросы для логистической модели, многоуровневых моделей и смесей логистических моделей;
- отбирает признаки в логистической модели;
- строит многоуровневые модели или смесь логистических моделей;
- определяет требуемое количество моделей;
- визуализирует результаты.

Положения, выносимые на защиту:

- Итеративный алгоритм оценки полной ковариационной матрицы параметров логистической регрессии в множестве всех ковариационных матриц соответствующего размера.
- Алгоритм отбора признаков, основанный на применении предложенной модификации метода Белсли к оценке ковариационной матрицы параметров, полученной с помощью максимизации обоснованности модели.
- Алгоритм отбора выбросов с использованием ковариационной матрицы оценок параметров для многоуровневых моделей и смесей логистических моделей.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных методов на реальных и синтетических данных; 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. Доклад: Multimodelling and Object Selection for Banking Credit Scoring.
2. 57я научная конференция МФТИ с международным участием, 2014 г. Доклад: “Мультимоделирование при построении моделей в задачах банковского скоринга”.

Работа поддержана грантами Российского фонда фундаментальных исследований:

1. 14-07-31264, Российский фонд фундаментальных исследований в рамках гранта “Развитие методов визуализации иерархических тематических моделей”,

2. 14-07-31205, Российский фонд фундаментальных исследований в рамках гранта "Развитие теории выбора мультимоделей в задачах прогнозирования и классификации"

Публикации по теме дипломной работы. Основные результаты по теме диплома изложены в следующих изданиях:

1. Адуенко А. А. Выбор признаков и шаговая логистическая регрессия для задачи кредитного скоринга // Машинное обучение и анализ данных, 2012. № 3. С. 279–291.

2. Адуенко А. А., Кузьмин А. А., Стрижов В. В. Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Известия ТулГУ, 2012. № 3. С. 119–131.

3. Адуенко А. А., Стрижов В. В. Алгоритм оптимального расположения названий коллекции документов // Программная инженерия, 2013. № 3. С. 21–25.

4. Иванова А. В., Адуенко А. А., Стрижов В. В. Алгоритм построения логических правил при разметке текстов // Программная инженерия, 2013. № 6. С. 41–45.

5. Адуенко А. А., Стрижов В. В. Совместный выбор объектов и признаков в задачах многоклассовой классификации коллекции документов // Инфокоммуникационные технологии, 2014. № 1. С. 47–53.

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.

Результаты квалификационной работы магистра были применены для классификации на два класса для данных по немецким [6] потребительским кредитам, заболеваниям сердца в Южной Африке [7], данным по качеству вина [8] и данным по локализации белка в клетке [9], а также для сгенерированных синтетических данных, содержащих кластеризованные и некластеризованные выбросы и синтетических данных, имеющих известную зависимость между признаками.

Обзор литературы. Рассматривается задача банковского кредитного скоринга. Скоринговая карта позволяет оценить вероятность дефолта заемщика, основываясь на ответах в кредитной анкете. Имея оценку вероятности дефолта, банк решает, одобрить ли заявку на кредит. Скоринговые карты создаются, исходя из кредитных историй [10]. Базы кредитных историй содержат миллионы записей, часть из которых могут быть повреждены и содержать неверные данные. Поэтому чтобы построить скоринговую карту высокого качества, которая не потребует изменений на протяжении нескольких лет, требуется выбрать набор клиентских записей из базы кредитных историй, которые будут использованы для настройки модели. Предсказательная сила построенной модели может существенно отличаться для разных наборов клиентских записей, использованных для ее настройки.

Две наиболее часто используемые функции связи в моделях с бинарным ответом есть логит и пробит функции [12]. Однако для моделей, в которых переменная

отклика одномерная, существенной разницы от использования этих двух функций не выявлено [12, 13]. В работе используется логит или сигмоидная функция связи в выражении для вероятности дефолта заемщика [14, 15]. Оценки параметров моделей должны удовлетворять следующим требованиям: быть несмещенными и должны позволять делать качественные и стабильные к небольшим изменениям выборки прогнозы. Предполагается, что смещение оценок параметров модели зависит от наличия выбросов в наборе клиентских записей, выбранном для оценки параметров модели. Поэтому первой задачей, которая рассматривается в рамка данной работы является задача отбора объектов или фильтрации выбросов.

Существуют прямые и непрямые методы обнаружения выбросов [16]. Прямые методы используют пошаговые add-delete процедуры для обнаружения и фильтрации выбросов.Непрямые методы используют специально построенные функции для определения вероятности принадлежности каждого объекта множеству выбросов. Предлагаемый в работе метод относится к методам второго класса. В статье [17] рассматриваются некластеризованные выбросы и вводится функция расстояния между элементами выборки. Результат анализа дендрограммы кластеризации дает сценарий фильтрации выбросов. В работе [18] в качестве функции расстояния предлагается использовать расстояние Махалонобиса. В работе [19] предлагается использовать расстояния Махалонобиса, но на уровне кластеров. Filzmoser и др. [20] разработали метод обнаружения выбросов в выборке, в которой объекты содержат большое число признаков.

В данной работе вводится функция специфичности для каждого объекта выборки. Вычисление функции специфичности требует наличия ковариационной матрицы параметров логистической модели [21]. В качестве ковариационной матрицы параметров используется ковариационная матрица апостериорного распределения параметров при использовании равномерного априорного псевдораспределения на параметры логистической модели. Известно, что наличие выбросов в выборке существенно влияет на оценки регрессионных параметров [22]. Кроме того, удаление объекта, который не является выбросом, обычно не изменяет существенно полученные оценки. Эти свойства и лежат в основе построенной функции специфичности, используемой для фильтрации выбросов. Так как свойства могут не выполнены для выборов, содержащих большие кластеры выбросов, метод стоит использовать с осторожностью, если априори неизвестно отсутствие таких кластеров. Кроме того, предполагается, что число объектов значительно больше, чем число признаков, поскольку в противном случае оценка максимума правдоподобия для параметров модели вырождается. Отметим, что типичная кредитная анкета содержит 10-100 полей, а базы кредитных истории–миллионы записей. Поэтому указанное свойство выполнено.

В работе признаковые описания объектов считаются заданными и неслучайными. Значения откликов для разных объектов считаются независимыми. Случай зависимых откликов рассмотрен в [23, 24]. Используя введенную функции специфичности в качестве набора клиентских записей, по которым настраиваются параметры модели, выбираются наименее специфичные объекты. Статистическая значимость улучшения качества после отбора выбросов в терминах AUC [25] показана с помощью сэмплирования AUC. Нормальность полученного эмпирического распределения проверялась с помощью критерия Шапиро-Уилка [26].

Второй важной задачей является отбор информативных признаков, то есть множества полей анкеты, которые имеют значимую статистическую связь с вероятно-

стью дефолта заемщика [14, 15]. Для отбора признаков можно использовать, например, генетические алгоритмым [27, 28], которые позволяют в значительной мере сократить перебор возможных наборов признаков. Идею генетических алгоритмов можно применить и для отбора объектов или совместного отбора объектов и признаков. Однако так как число объектов в задаче кредитного скоринга значительно превышает число признаков, такое рассмотрение оказывается возможным только для выборок очень малого размера. Схожим подходом можно считать шаговые алгоритмы отбора признаков [23]. Другим подходом к отбору признаков является подход, основанный на оценке информативности признаков [10, 11], при котором в модели оставляют некоторое количество наиболее информативных признаков.

Отметим, что задача отбора признаков может решаться как отдельно, так и одновременно с задачей фильтрации выбросов. Примером отдельного решения задачи отбора признаков является применение генетического алгоритма для отбора признаков, а затем применение метода классификации с отбором объектов к полученной выборке с меньшим числом признаков [29]. Одновременное решение задачи отбора признаков и фильтрации выбросов может быть реализовано, например, путем изменения оптимизационной задачи так, чтобы целевая функция и ограничения учитывали также требование отбора информативных признаков [30–32].

Так в качестве указанного изменения целевой функции для отбора информативных признаков можно использовать регуляризацию [33]. Регуляризация позволяет уменьшить вероятность переобучения. На практике одним из свидетельств переобучения считают большие значения нормы вектора параметров по сравнению с характерным диапазоном изменения целевой переменной. По этой причине регуляризация зачастую состоит в введении в целевую функцию штрафа за норму вектора параметров модели [34]. Два часто использующихся штрафа есть квадратичный штраф [34] и l_1 норма вектора параметров модели [35]. Заметим, что и l_1 штраф, и квадратичный штраф являются выпуклыми, а потому если целевая функция до регуляризации была выпуклой, то она останется таковой и после регуляризации, что позволяет применять эффективные методы оптимизации [36], хотя в случае с l_1 штрафом требуется учитывать недифференцируемость в нуле [37]. Отметим, однако, что l_1 штраф является значительно более поощряющим разреженность, то есть уменьшающим число используемых признаков [38, 39] Часто используют также невыпуклые штрафы, поскольку они могут лучше поощрять разреженность полученного вектора параметров [40]. В работе [33] рассматривается также совместное применение непрерывных штрафов за рост нормы вектора параметров и дискретных, связанных с числом параметров и объектов в модели. Стоит однако отметить, что использование дискретных штрафов создает дополнительные трудности в оптимизации [37].

Подходом, который является двойственным к регуляризации (приводит к тем же целевым функциям), является подход, основанный на введении априорного распределения на вектор параметров модели [41,42]. Так регуляризации с помощью квадратичного штрафа приводит к той же задаче оптимизации, что и введение априорного нормального распределения на параметры модели, а l_1 штраф приводит к той же задаче, что и введение априорного распределения Лапласа. Преимуществом такого подхода является то, что возможно получить не только точечную оценку параметров модели, но и апостериорное распределение параметров модели. С помощью апостериорного распределения можно, например, построить доверительную область значения параметров и убрать из модели те признаки, веса которых отличаются от нуля

незначимо. Однако, как и в случае с регуляризацией, существует произвол в выборе коэффициента регуляризации, то есть характерной дисперсии априорного распределения. В зависимости от значения этого коэффициента значимыми могут получаться разные наборы признаков.

Выбор коэффициента регуляризации можно сделать, например, с помощью кросс-валидации [43]. Другим способом является использование обоснованности модели [14, 44]. В случае со скалярным параметром методы дают схожие результаты [**?**]. Однако обоснованность модели можно использовать и непосредственно для отбора признаков [14,44,46]. Для этого в качестве априорного распределения, например, используют нормальное с диагональной матрицей ковариации, вообще говоря, с разными дисперсиями. Однако такой метод не позволяет учитывать зависимости между признаками. По этой причине в работе предлагается оценивать полную ковариационную матрицу. Полученную ковариационную матрицу используем для отбора признаков с помощью предложенной модификации метода Белсли для линейной регрессии [42, 47].

Третьей важной проблемой является проблема учета неоднородности данных. Для решения этой проблемы можно использовать кластеризацию, а затем строить модель для каждого кластера в отдельности [48]. Также вместо кластеризации можно использовать параметрическую зависимость между вероятностью принадлежности объекта признакового пространства и его признаковым описанием. Такой подход используется, например, в смессях экспертов [14,49]. Также существуют методы построения иерархических моделей [49,50]. В данной работе предлагается использовать многоуровневые модели [51, 52] и смеси логистических моделей [14, 53]. Необходимое число моделей для описания данных можно, например, выбирать с помощью пошагового добавления моделей [47]. В данной работе предлагается использовать для этой цели симметричное априорное распределение Дирихле, поощряющее разреженность. Выбор параметра распределения Дирихле предлагается выполнять с помощью кросс-валидации [43].

1 Постановка задачи

Задана выборка $D = \{(\mathbf{x}_i, y_i)\}$, элементы которой индексированы множеством $\mathcal{I} = \{1, \ldots, m\} \ni i$. Представим элементы $\mathbf{x}_i$ в виде матрицы $\mathbf{X} = [\mathbf{x}_1^\intercal, \ldots, \mathbf{x}_m^\intercal]^\intercal \in \mathbb{R}^{m \times n}$, где m–число элементов выборки, n–количество признаков. Элементы $y_i \in \{0, 1\}$ представим в виде вектора ответов $\mathbf{y} = [y_1, \ldots, y_m]^\intercal$. Здесь число 0 означает, что заемщик кредит вернул, а 1 — не вернул, то есть произошел дефолт заемщика. Для индексации признаков введем обозначение $\mathcal{J} = \{1, \ldots, n\}$.

Предполагается, что каждый элемент y_i есть реализация бернуллевской случайной величины $Y_i \sim Be(p_i)$, $\{Y_i\}_{i=1}^m$ независимы в совокупности. При этом вероятность невозврата кредита p_i заемщиком определяется с помощью модели логистической регрессии

$$p_i = f(\mathbf{x}_i, \mathbf{w}) = f(\mathbf{x}_i^\intercal \mathbf{w}) = \frac{1}{1 + \exp\left(-\mathbf{x}_i^\intercal \mathbf{w}\right)}. \tag{1}$$

Здесь $\mathbf{w} \in \mathcal{W} = \mathbb{R}^n$ вектор параметров модели. Для нахождения оценки $\hat{\mathbf{w}}$ вектора параметров $\mathbf{w}$ будет использоваться метод наибольшего правдоподобия (2).

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}\in\mathbb{R}^n} L(\mathbf{y}|\mathbf{X},\ \mathbf{w}), \tag{2}$$

где $L(\mathbf{y}|\mathbf{X},\ \mathbf{w})$ функция правдоподобия данных, а $\hat{\mathbf{w}}$ – оценка максимума правдоподобия для параметров модели $\mathbf{w}$.

Введем обозначения $\mathbf{X}(\mathcal{B},\ \mathcal{A})$ и $\mathbf{y}(\mathcal{B})$, соответствующие усеченной матрице объект-признак $\mathbf{X}(\mathcal{B},\ \mathcal{A}) = [x_{ij}]$, $i \in \mathcal{B} \subseteq \mathcal{I}$, $j \in \mathcal{A} \subseteq \mathcal{J}$ и усеченному вектору ответов $\mathbf{y}(\mathcal{B}) = [y_i]$, $i \in \mathcal{B} \subseteq \mathcal{I}$. Тогда задачу отбора объектов и признаков можно записать в виде

$$[\mathcal{B},\ \mathcal{A}] = \arg\max_{\mathcal{B}\in\mathcal{S},\ \mathcal{A}\in\mathcal{J}} L(\mathbf{y}(\mathcal{T})|\mathbf{X}(\mathcal{T},\ \mathcal{A}),\ \hat{\mathbf{w}}), \tag{3}$$

где

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}\in\mathbb{R}^{|\mathcal{A}|}} L(\mathbf{y}(\mathcal{B})|\mathbf{X}(\mathcal{B},\ \mathcal{A}),\ \mathbf{w}). \tag{4}$$

1.1 Функция правдоподобия и ее свойства

Запишем функцию правдоподобия данных для одной модели и покажем, что она является вогнутой, а потому имеет единственный максимум. В силу предположения о независимости в совокупности случайных величини $\{Y_i\}_{i=1}^m$ для функции правдоподобия имеем выражение

$$L(\mathbf{y}(\mathcal{B})|\mathbf{X}(\mathcal{B},\ \mathcal{A}),\ \mathbf{w}) = \prod_{t\in\mathcal{B}} f(\mathbf{x}_t^\intercal(\mathcal{A})\mathbf{w})^{y_t}\left(1 - f(\mathbf{x}_t^\intercal(\mathcal{A})\mathbf{w})\right)^{1-y_t}. \tag{5}$$

Рассмотрим функцию

$$l(\mathbf{y}(\mathcal{B})|\mathbf{X}(\mathcal{B},\ \mathcal{A}),\ \mathbf{w}) = -\ln L(\mathbf{w}|\mathbf{X}(\mathcal{B},\ \mathcal{A}),\ \mathbf{y}(\mathcal{B})). \tag{6}$$

Далее для удобства записи опустим $\mathcal{B}$ и $\mathcal{A}$ в выражении (6), введем обозначение $r = |\mathcal{B}|$ и без ограничения общности предположим $\mathcal{B} = \{1,\ \ldots,\ r\}$, тогда

$$l(\mathbf{y}|\mathbf{X},\ \mathbf{w}) = -\sum_{t=1}^{r} y_t \ln f(\mathbf{x}_t^\intercal\mathbf{w}) + (1 - y_t)\ln(1 - f(-\mathbf{x}_t^\intercal\mathbf{w})). \tag{7}$$

Продифференцируем (7) по $\mathbf{w}$ с учетом $f'(x) = f(x)\cdot f(-x)$, получим

$$\frac{\partial l(\mathbf{w})}{\partial \mathbf{w}} = -\sum_{t=1}^{r} \mathbf{x}_t(y_t - f(\mathbf{x}_t^\intercal\mathbf{w})) = \mathbf{X}^\intercal(\mathbf{f} - \mathbf{y}). \tag{8}$$

Найдем гессиан $l(\mathbf{w}|\mathbf{X},\ \mathbf{y})$ и покажем, что он положительно определен, откуда и получим выпуклость $l(\mathbf{w}|\mathbf{X},\ \mathbf{y})$, а в силу монотонности преобразования $g(x) = -\ln(x)$, получим вогнутость $L(\mathbf{w}|\mathbf{X},\ \mathbf{y})$ и единственность максимума функции правдоподобия.

$$\mathbf{H} = \frac{\partial^2 l(\mathbf{w})}{\partial \mathbf{w}^2} = \sum_{t=1}^{r} \mathbf{x}_t f(\mathbf{x}_t^\intercal\mathbf{w}) f(-\mathbf{x}_t^\intercal\mathbf{w})\mathbf{x}_t^\intercal = \mathbf{X}^\intercal\mathbf{R}\mathbf{X}, \tag{9}$$

где $\mathbf{R}$ — диагональная матрица с элементами $f(\mathbf{x}_t^\intercal\mathbf{w})f(-\mathbf{x}_t^\intercal\mathbf{w}) > 0$ на диагонали. Рассмотрим произвольный вектор $\mathbf{u} \neq \mathbf{0}$.

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

откуда $\mathbf{H}$ положительно определенная, то есть $l(\mathbf{w})$ выпуклая, а потому имеет единственный минимум. Перейдем к алгоритму нахождения этого минимума.

1.2 Оценка параметров скоринговой модели

В случае линейной регрессии существует явное выражение для оценки наибольшего правдоподобия, поскольку в этом случае функция $l(\mathbf{w})$ квадратичная. В силу нелинейности сигмоидной функции получить явное выражение для $\hat{\mathbf{w}}$ не удается, но можно предложить итерационную процедуру, основанную на методе Ньютона-Рафсона для нахождения оценки наибольшего правдоподобия $\hat{\mathbf{w}}$ для вектора параметров логистической модели.

На начальном шаге задается вектор $\hat{\mathbf{w}}_0$ оценок параметров. На каждом следующем шаге вычисляется новое приближение $\hat{\mathbf{w}}_i$ к оценке наибольшего правдоподобия $\hat{\mathbf{w}}$ по формуле

$$\hat{\mathbf{w}}_i = \hat{\mathbf{w}}_{i-1} - \mathbf{H}^{-1}\nabla l(\hat{\mathbf{w}}_{i-1}) = \hat{\mathbf{w}}_{i-1} - \left(\mathbf{X}^\intercal\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^\intercal(\mathbf{f}-\mathbf{y}). \tag{10}$$

Вычисления по формуле (10) продолжаем до тех пор, пока изменение нормы вектора $\mathbf{w}$ не перестанет быть значительным.

2 Выбор оптимального множества объектов и признаков

2.1 Алгоритм устойчивого отбора признаков

До сих пор предполагалось, что вектор параметров модели $\mathbf{w}$ неизвестен, но фиксирован. Введем далее априорное распределение на вектор параметров $\mathbf{w}$. Тогда совместное правдоподобие приобретает вид

$$p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) = p(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w}|\mathbf{A}). \tag{11}$$

В (11) предполагается независимость $\mathbf{w}$ от признаков, а также $\mathbf{A}$ обозначены параметры априорного распределения на $\mathbf{w}$. Конкретный вид распределения и смысл параметров будет приведен в дальнейшем. Тогда в качестве вектора параметров модели выбираем моду апостериорного распределения на $\mathbf{w}$, то есть

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X},\ \mathbf{y},\ \mathbf{A}) = \arg\max_{\mathbf{w}} \frac{p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A})}{\int p(\mathbf{y},\ \mathbf{w}'|\mathbf{X},\ \mathbf{A})d\mathbf{w}'} = \arg\max_{\mathbf{w}} p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}). \tag{12}$$

Заметим, что оценка на вектор параметров $\hat{\mathbf{w}}$ из (15) в точности совпадает с оценкой $\hat{\mathbf{w}}$ из (2), если в качестве априорного $p(\mathbf{w}|\mathbf{A})$ ввести равномерное псевдораспределение на $\mathbf{w}$. Рассмотрим сначала такое равномерное априорное псевдораспределение и получим оценку ковариационной матрицы параметров в этом случае.

Случай равномерного априорного псевдораспределения. В качестве оценки ковариационной матрицы используем ковариационную матрицу апостериорного распределения параметров. Заметим, что из свойств оценки максимума правдоподобия для вектора параметров $\hat{\mathbf{w}}$ имеем $\nabla l(\hat{\mathbf{w}}) = 0$, откуда

$$\ln\frac{p(\mathbf{w}|\mathbf{X},\ \mathbf{y})}{p(\hat{\mathbf{w}}|\mathbf{X},\ \mathbf{y})} \approx -\frac{1}{2}(\mathbf{w}-\hat{\mathbf{w}})^\intercal\mathbf{H}(\mathbf{w}-\hat{\mathbf{w}}). \tag{13}$$

Пользуясь локально нормальной аппроксимацией апостериорного распределения параметров $\mathbf{w}$, получим, что $\mathbf{w} \sim N(\hat{\mathbf{w}}, \mathbf{H}^{-1})$. Таким образом, оценкой ковариационной матрицы параметров является матрица $\mathbf{H}^{-1}$.

Случай нормального априорного распределения. Рассмотрим теперь случай более содержательного априорного распределения на параметры $\mathbf{w}$. В качестве такого априорного распредедения используем нормальное распределение

$$\mathbf{w} \sim N(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1}).$$

При таком априорном распределении имеем, что если $a_{jj} = \infty$, то признак с номером j избыточен, поскольку априори его вес $w_j = 0$. Это свойство можно использовать для отбора признаков. Также всю матрицу $\mathbf{A}^{-1}$ можно использовать для отбора признаков, например, с помощью модификации метода Белсли, изложенной ниже.

Однако априори матрица $\mathbf{A}$ неизвестна, так как ее наличие априори фактически означало бы знание структуры зависимостей параметров модели, что не выполняется для реальных данных. Поэтом матрицу $\mathbf{A}$, как и вектор параметров $\mathbf{w}$ требуется оценить. Прямым способом оценки можно назвать следующий

$$[\hat{\mathbf{w}},\ \hat{\mathbf{A}}] = \arg\max_{\mathbf{w},\,\mathbf{A}} p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}).$$

Однако данный способ не дает содержательной оценки матрицы $\mathbf{A}$, так как, взяв $a_{jj} = \infty$ для всех признаков, получим, что $p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A})$ принимает значение, равное ∞ при $\mathbf{w} = \mathbf{0}$ и значение 0 иначе. Таким образом, оптимальные значения $\hat{\mathbf{w}}$, $\hat{\mathbf{A}}$ тривиальны и не зависят от выборки, что не является содержательным. По этой причине для нахождения оценки $\hat{\mathbf{A}}$ предлагается использовать максимизацию обоснованности [44]. Опишем соответвующую задачу.

$$\hat{\mathbf{A}} = \arg\max_{\mathbf{A}\in\mathcal{M}} p(\mathbf{y}|\mathbf{X},\ \mathbf{A}) = \arg\max_{\mathbf{A}\in\mathcal{M}} \int p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A})d\mathbf{w}, \tag{14}$$

где $\mathcal{M}$ – некоторое подмножество неотрицательно определенных матриц размера $m \times m$. Величина правдоподобия $p(\mathbf{y}|\mathbf{X},\ \mathbf{A})$ называется обоснованностью и показывает, насколько в среднем рассматриваемая модель соответствует данным.

Отметим, что интеграл в выражении (14) аналитически не считается, поскольку правдоподобие $p(\mathbf{y}|\mathbf{X},\ \mathbf{w})$ 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. Далее рассмотрим два различных подхода к аппроксимации интеграла в выражении (14) и нахождения матрицы $\mathbf{A}$.

Оценка ковариационной матрицы с помощью аппроксимации Лапласа Перепишем интеграл в выражении (14) в следующем виде

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

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

$$\mathbf{w}_{MP} = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{X},\ \mathbf{y},\ \mathbf{A}) = \arg\max_{\mathbf{w}} \frac{p(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w}|\mathbf{A})}{\int p(\mathbf{y}|\mathbf{X},\ \mathbf{w}')p(\mathbf{w}'|\mathbf{A})d\mathbf{w}'}. \tag{15}$$

Тогда из (15) 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 $q(\mathbf{w})$.

$$q(\mathbf{w}) \approx q(\mathbf{w}_{MP}) - \frac{1}{2}(\mathbf{w}-\mathbf{w}_{MP})^{\intercal}\mathbf{H}^{-1}(\mathbf{w}-\mathbf{w}_{MP}), \text{ где } \mathbf{H}^{-1} = -\nabla^2 q(\mathbf{w})|_{\mathbf{w}=\mathbf{w}_{MP}}. \quad (16)$$

Используем (16) для приближения $I(\mathbf{A})$, получим

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

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

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

$$\log I(\mathbf{A}) = \log p(\mathbf{y}|\mathbf{X},\, \mathbf{w}_{MP}) + \frac{1}{2}\log\det \mathbf{A} - \frac{1}{2}\mathbf{w}_{MP}^{\intercal}\mathbf{A}\mathbf{w}_{MP} - \frac{1}{2}\log\det \mathbf{H}^{-1}. \quad (17)$$

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

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

было использовано выражение (9) для гессиана функции правдоподобия и нормальность априорного распределения $p(\mathbf{w}|\mathbf{A})$. Диагональная весовая матрица $\mathbf{R}$ также определяется в (9). Здесь лишь отметим, что 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

$$\mathbf{A} = \arg\min_{\mathbf{A}\in\mathcal{M}} \left[-\log p(\mathbf{y}|\mathbf{X},\, \mathbf{w}_{MP}) - \frac{1}{2}\log\det \mathbf{A} + \frac{1}{2}\mathbf{w}_{MP}^{\intercal}\mathbf{A}\mathbf{w}_{MP} + \frac{1}{2}\log\det(\mathbf{X}^{\intercal}\mathbf{R}\mathbf{X} + \mathbf{A})\right]. \quad (18)$$

Отметим, что в (18) оценка максимума апостериорной вероятности для параметров $\mathbf{w}_{MP}$ зависит от матрицы $\mathbf{A}$, причем через решение задачи оптимизации, аналогичной нахождению оценки максимума правдоподобия на параметры (10). Поэтому получение полной производной по $\mathbf{A}$ затруднено. Потому для решения задачи (18) 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 далее два множества: диагональные матрицы и матрицы общего вида.

Опишем теперь итерационный алгоритм решения задачи (18). Выбираем начальное приближение для матрицы $\mathbf{A}$. Далее по очереди выполняем итерации по $\mathbf{w}_{MP}$ при фиксированной $\mathbf{A}$ с предыдущей итерации и по $\mathbf{A}$ при фиксированном $\mathbf{w}_{MP}$ с предыдущей итерации.

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

При фиксированной матрице $\mathbf{A}$ задача (18) эквивалентна следующей задаче нахождения $\mathbf{w}_{MP}$

$$\tilde{l}(\mathbf{w}) = -\log p(\mathbf{y}|\mathbf{X},\, \mathbf{w}) + \frac{1}{2}\mathbf{w}^\intercal \mathbf{A}\mathbf{w} \to \min_{\mathbf{w}}.$$

Для решения этой задачи, как уже отмечалось, можно использовать метод Ньютона-Рафсона (10), добавив члены от квадратичного слагаемого в градиент и гессиан минимизируемой функции

$$\frac{\partial \tilde{l}(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{A}\mathbf{w} + \mathbf{X}^\intercal(\mathbf{f} - \mathbf{y}),$$

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

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

При фиксированном $\mathbf{w}_{MP}$ задача (18) эквивалентна следующей задаче нахождения $\mathbf{A}$

$$l(\mathbf{A}) = -\frac{1}{2}\log\det\mathbf{A} + \frac{1}{2}\mathbf{w}_{MP}^\intercal \mathbf{A}\mathbf{w}_{MP} + \frac{1}{2}\log\det(\mathbf{X}^\intercal\mathbf{R}\mathbf{X} + \mathbf{A}) \to \min_{\mathbf{A}\in\mathcal{M}}.$$

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

Случай, когда множество допустимых матриц $\mathcal{M}$ есть множество диагональных матриц с положительными элементами на диагонали размера $n \times n$.

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

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

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

$$\alpha_j^{new} = \frac{1 - \alpha_j^{old} H_{jj}^{old}}{w_j^2}. \tag{19}$$

Случай, когда $\mathcal{M}$ есть множество всех симметричных положительно определенных матриц размера $n \times n$.

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

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

$$\mathbf{A}^{new} = \left(\mathbf{w}_{MP}\mathbf{w}_{MP}^\intercal + (\mathbf{X}^\intercal\mathbf{R}\mathbf{X} + \mathbf{A}^{old})^{-1}\right)^{-1}. \tag{20}$$

Вариационная оценка матрицы ковариаций Рассмотрим второй метод нахождения оценки матрицы $\mathbf{A}$, то есть приближенного решения задачи (14). Этот метод основан на построении вариационной нижней оценки к функции правдоподобия [14]. Далее приведем определение.

Определение.

Вариационной оценкой функции $f(x)$ называется функция $g(x, \xi)$, такая что:

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

Для сигмоиды существует вариационная нижняя оценка вида [14]

$$\sigma(x) \geq \sigma(\xi) \exp\{(x - \xi)/2 - \lambda(\xi)(x^2 - \xi^2)\}, \tag{21}$$

где

$$\lambda(\xi) = \frac{1}{2\xi}\left[\sigma(\xi) - \frac{1}{2}\right]. \tag{22}$$

Воспользуемся ей для оценки интеграла в (14). Преобразуем вектор ответов, заменив метки классов с {0, 1} на {-1, 1}:

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 \\ -1 \\ -1 \\ \vdots \\ 1 \end{pmatrix}.$$

Тогда

$$P(y_i = t_i \mid \mathbf{X}, \mathbf{w}) = \sigma(t_i \mathbf{w}^\intercal \mathbf{x}_i), \quad t_i \in \{-1, 1\},\ i = 1, \ldots, n.$$

Правдоподобие $p(\mathbf{y}|\mathbf{X},\ \mathbf{w})$ имеет вид

$$p(\mathbf{y}|\mathbf{X},\ \mathbf{w}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i,\ \mathbf{w}) = \prod_{i=1}^{n} \sigma(y_i \mathbf{w}^\intercal \mathbf{x}_i). \tag{23}$$

Воспользуемся вариационной нижней оценкой (21) для сигмоиды и получим вариационную нижнюю оценку для функции правдоподобия (23)

$$p(\mathbf{y}|\mathbf{X},\ \mathbf{w}) \geq \prod_{i=1}^{n} \sigma(\xi_i) \exp\left[\frac{y_i \mathbf{w}^\intercal \mathbf{x}_i - \xi_i}{2} - \lambda(\xi_i)((\mathbf{w}^\intercal \mathbf{x}_i)^2 - \xi_i^2)\right]. \tag{24}$$

В (24) введен вектор вариационных параметров $\boldsymbol{\xi}$ размера $n \times 1$. Пользуясь вариационной нижней оценкой для правдоподобия (24), получим нижнюю оценку (уже не вариационную) для обоснованности модели $p(\mathbf{y}|\mathbf{X},\ \mathbf{A})$

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

$$\int \prod_{i=1}^{n} \sigma(\xi_i) \exp\left[\frac{y_i \mathbf{w}^\intercal \mathbf{x}_i - \xi_i}{2} - \lambda(\xi_i)((\mathbf{w}^\intercal \mathbf{x}_i)^2 - \xi_i^2)\right] \frac{\sqrt{\det \mathbf{A}}}{(2\pi)^{m/2}} \exp(-1/2 \mathbf{w}^\intercal \mathbf{A} \mathbf{w}) d\mathbf{w}. \tag{25}$$

Выпишем выражение для логарифма подинтегрального выражения $\log I(\mathbf{w})$ в (25)

$$\log I(\mathbf{w}) = -\frac{1}{2}\mathbf{w}^\intercal\mathbf{A}\mathbf{w}+\frac{1}{2}\log\det\mathbf{A}+\sum_{i=1}^{n}\log\sigma(\xi_i)+\sum_{i=1}^{n}\frac{y_i\mathbf{w}^\intercal\mathbf{x}_i-\xi_i}{2}-\lambda(\xi_i)\left[(\mathbf{w}^\intercal\mathbf{x}_i)^2-\xi_i^2\right]= \\ -\frac{1}{2}(\mathbf{w}-\mathbf{w}_0)^\intercal\mathbf{A}'(\mathbf{w}-\mathbf{w}_0)+C.$$

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

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

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

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

$$\log p(\mathbf{y}|\mathbf{X},\mathbf{A}) \geq \frac{1}{2}\log\det\mathbf{A}-\frac{1}{2}\log\det\mathbf{A}'+\sum_{i=1}^{n}\log\sigma(\xi_i)-\frac{1}{2}\sum_{i=1}^{n}\xi_i+\sum_{i=1}^{n}\lambda(\xi_i)\xi_i^2+\frac{1}{2}\mathbf{v}^\intercal\mathbf{A}'^{-1}\mathbf{v} = b(\mathbf{A},\boldsymbol{\xi}). \tag{26}$$

Заменим теперь задачу максимизации обоснованности (14) задачей максимизации нижней оценки на ее логарифма

$$[\mathbf{A},\boldsymbol{\xi}] = \arg\max_{\boldsymbol{\xi},\,\mathbf{A}\in\mathcal{M}}\left[\frac{1}{2}\log\det\mathbf{A}-\frac{1}{2}\log\det\mathbf{A}'+\sum_{i=1}^{n}\log\sigma(\xi_i)-\frac{1}{2}\sum_{i=1}^{n}\xi_i+\sum_{i=1}^{n}\lambda(\xi_i)\xi_i^2+\frac{1}{2}\mathbf{v}^\intercal\mathbf{A}'^{-1}\mathbf{v}\right] \tag{27}$$

Опишем теперь итерационный алгоритм решения задачи (27). 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 $\mathbf{A}$).

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

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

Преобразуем полученное условие и получаем

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

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

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

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

При фиксированном $\boldsymbol{\xi}$ задача (27) эквивалентна следующей задаче нахождения $\mathbf{A}$

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

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

Случай, когда множество допустимых матриц $\mathcal{M}$ есть множество диагональных матриц с положительными элементами на диагонали размера $n \times n$.

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

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

Отсюда получаем выражение для α_j

$$\alpha_j^{new} = \frac{1}{(\mathbf{A}'^{-1})^{\text{old}}_{jj} + [(\mathbf{A}'^{-1}\mathbf{v})^{\text{old}}_j]^2}. \tag{28}$$

Случай, когда $\mathcal{M}$ есть множество всех симметричных положительно определенных матриц размера $n \times n$.

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

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

$$\mathbf{A}^{new} = \left(\mathbf{I} + \mathbf{v}\mathbf{v}^\intercal\mathbf{A}'^{-1}_{\text{old}}\right)^{-1}_{\text{old}}. \tag{29}$$

Заметим, что положительная определенность матрицы $\mathbf{A}$, определенной в (29), следует из положительной определенности $\mathbf{A}'$ на предыдущей итерации и эквивалентной записи (29) в виде

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

EM-алгоритм для решения задачи (27). Заметим, что решить задачу (27) 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.

E-шаг.

На данном шаге вычисляем апостериорное распределение на вектор скрытых переменных $\mathbf{w}$ $p(\mathbf{w}|\mathbf{X}, \mathbf{y})$. Отметим, что правдоподобие $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$ здесь уже заменено на свою нижнюю оценку. Тогда $p(\mathbf{y}|\mathbf{X}, \mathbf{w})$ и $p(\mathbf{w}|\mathbf{A})$ оказываются сопряжены и имеем апостериорное нормальное распределение на вектор $\mathbf{w}$ (аналогично взятию интеграла в (25)) вида

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

$\mathbf{v} = 1/2 \sum_{i=1}^{n} y_i \mathbf{x}_i$ и

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

М-шаг.

На данном шаге решаем следующую задачу максимизации.

$$\mathbb{E}_{q(\mathbf{w})} \log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) \to \max_{\mathbf{A},\,\boldsymbol{\xi}},\ \text{где} \tag{30}$$

под $q(\mathbf{w})$ понимается апостериорное распределение на $\mathbf{w}$, полученное на E-шаге.

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

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

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

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 (30) эквивалентна следующей задаче нахождения $\mathbf{A}$

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

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

Случай, когда множество допустимых матриц $\mathcal{M}$ есть множество диагональных матриц с положительными элементами на диагонали размера $n \times n$.

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

$$\alpha_j^{new} = \frac{1}{(\mathbb{E}_q \mathbf{w}\mathbf{w}^\intercal)_{jj}^{\text{old}}} = \frac{1}{(\mathbf{w}_0\mathbf{w}_0^\intercal + \mathbf{A}'^{-1})_{\text{old}}}. \tag{31}$$

Случай, когда $\mathcal{M}$ есть множество всех симметричных положительно определенных матриц размера $n \times n$.

Условие оптимальности первого порядка имеет вид

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

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

$$\mathbf{A}^{new} = (\mathbf{w}_0\mathbf{w}_0^\intercal + \mathbf{A}'^{-1})_{\text{old}}^{-1}. \tag{32}$$

Модификация метода Белсли для задачи логистической регрессии. Воспользуемся найденной оценкой ковариационной матрицы $\mathbf{A}$ для отбора признаков. Для матрицы $\mathbf{A}$ выполним разложение Холецкого, получим

$$\mathbf{A} = \tilde{\mathbf{A}}^{\mathsf{T}}\tilde{\mathbf{A}},$$

где $\tilde{\mathbf{A}}$ есть верхнетреугольная матрица с неотрицательными элементами на диагонали.

Выполним сингулярное разложение матрицы $\tilde{\mathbf{A}}$

$$\tilde{\mathbf{A}} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}}.$$

Тогда

$$\mathbf{var}(\mathbf{w}) = \mathbf{A}^{-1} = (\tilde{\mathbf{A}}^{\mathsf{T}}\tilde{\mathbf{A}})^{-1} = (\mathbf{V}\mathbf{\Lambda}\mathbf{U}^{\mathsf{T}}\mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}})^{-1} = \mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^{\mathsf{T}}.$$

Для обнаружения мультиколлинеарности признаков построим таблицу, в которой каждому индексу обусловленности η_j соответствуют значения q_{ij} — долевые коэффициенты. Сумма долевых коэффициентов по индексу j равна единице.

$$\mathbf{var}(w_i) = \sum_{j=1}^{n} \frac{v_{ij}^2}{\lambda_j^2} = (q_{i1} + q_{i2} + \ldots + q_{in}) \sum_{j=1}^{n} \frac{v_{ij}^2}{\lambda_j^2},$$

где q_{ij} — отношение соответствующего слагаемого в разложении вектора $\mathbf{var}(w_i)$ ко всей сумме, а $\mathbf{V} = (v_{ij})$.

Таблица 1: Разложение $\mathbf{var}(\mathbf{w})$

Индекс обусловленности	$\mathbf{var}(w_1)$	$\mathbf{var}(w_2)$	$\ldots$	$\mathbf{var}(w_n)$
η_1	q_{11}	q_{21}	$\cdots$	q_{n1}
η_2	q_{12}	q_{22}	$\cdots$	q_{n2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
η_n	q_{1n}	q_{2n}	$\cdots$	q_{nn}

Чем больше значение долевого коэффициента q_{ij}, тем больший вклад вносит j-ый признак в дисперсию i-го регрессионного коэффициента.

Из табл. 1 определяется мультиколлинеарность: большие величины η_j означают, что, возможно, есть зависимость между признаками. Если один из признаков является линейной комбинацией остальных, то матрица $\tilde{\mathbf{A}}$ будет матрицей неполного ранга, а потому у матрицы $\mathbf{\Lambda}$ одним из собственных значений будет 0. В случае мультиколлинеарности признаков матрица $\mathbf{\Lambda}$ будет иметь близкие к нулю собственные значения, которым соответствуют большие коэффициенты обусловленности. На этом и будет основан метод отбора признаков, изложенный ниже.

Алгоритм отбора признаков, основанный на методе Белсли. Заметим, что чем больше параметров в модели, тем точнее она описывает имеющиеся данные, тем меньше будет ее применимость для произвольных данных, то есть наблюдается переобучение. Поэтому если при последовательном добавлении признаков учитывать только внутренний критерий качества модели на обучающей выборке, то чтобы

уменьшить переобучение, нужно требовать значительного роста качества модели после добавления признака (фактически происходит введение штрафа за сложность модели). В данной работе использовался другой подход.

Опишем два этапа алгоритма: Add и Del. На первом этапе (Add) последовательно добавляются признаки. На втором этапе (Del) происходит последовательное удаление признаков, согласно методу Белсли. Пусть перед началом работы алгоритма модель задается множеством индексов объектов $\mathcal{B} \subseteq \mathcal{S}$ и множеством индексов признаков $\mathcal{A} \subseteq \mathcal{J}$. Опишем как построить новую модель, то есть определить новое множество индексов признаков $\tilde{\mathcal{A}}$. Множество индексов объектов $\mathcal{B} \subseteq \mathcal{S}$, соответствующих модели, остается неизменным. Выделим из контрольной выборки объектов, задаваемой множеством индексов $\mathcal{T}$, некоторую подвыборку, задаваемую множеством индексов $\mathcal{T}_1 \subseteq \mathcal{T}$, то есть $\mathcal{T} = \mathcal{T}_1 \sqcup \mathcal{T}_2$. Выделенная выборка будет применяться для отбора признаков, а выборка, задаваемая индексным множеством $\mathcal{T}_2$ — для контроля качества модели.

Этап Add. Для каждого $j \in \mathcal{J} \setminus \mathcal{A}$ найдем $\hat{\mathbf{w}}_j$ согласно (4)

$$\hat{\mathbf{w}}_j = \arg\max_{\mathbf{w} \in \mathbb{R}^{|\mathcal{A}|+1}} L(\mathbf{w}|\mathbf{X}(\mathcal{B},\, \mathcal{A} \cup \{j\}),\, \mathbf{y}(\mathcal{B}))$$

и сосчитаем согласно (6)

$$l(j) = l(\hat{\mathbf{w}}_j|\mathbf{X}(\mathcal{T}_1,\, \mathcal{A} \cup \{j\},\, \mathbf{y}(\mathcal{T}_1)).$$

Найдем признак $j^* \in \mathcal{J} \setminus \mathcal{A}$, для которого l_j принимает минимальное значение. Обозначим значение той же функции для модели с множеством признаков $\mathcal{A}$ как l_0, то есть

$$l_0 = l(\hat{\mathbf{w}}_0|\mathbf{X}(\mathcal{T}_1,\, \mathcal{A},\, \mathbf{y}(\mathcal{T}_1)),$$

где

$$\hat{\mathbf{w}}_0 = \arg\max_{\mathbf{w} \in \mathbb{R}^{|\mathcal{A}|}} L(\mathbf{w}|\mathbf{X}(\mathcal{B},\, \mathcal{A}),\, \mathbf{y}(\mathcal{B})).$$

Если $l_{j^*} - l_0 \leq Z_1 < 0$, где $Z_1 \in \mathbb{R}$, то есть произошло значительное улучшение качества модели, признак j^* добавляется в модель, то есть $\mathcal{A} \to \mathcal{A} \cup \{j^*\}$. Этап повторяется до тех пор, пока происходит добавление признаков в модель.

Этап Del. После того, как добавить признаки в модель на этапе Add же не получается, переходим к этапу Del. Находим индексы обусловленности и долевые коэффициенты для текущего набора признаков $\mathcal{A}$ согласно методу Белсли, описание которого приведено выше. Далее находим количество достаточно больших индексов обусловленности. Достаточно большими будем считать индексы, квадрат которых превосходит максимальный индекс обусловленности η_t, где $t = |\mathcal{A}|$, количество признаков в текущем наборе $\mathcal{A}$. Количество таких индексов обозначим

$$i^* = \sum_{g=1}^{t} \left[\eta_g^2 > \eta_t\right]. \tag{33}$$

Затем ищем в матрице долевых коэффициентов $\mathbf{var}(\mathbf{w})$ столбец j^* с максимальной суммой по последним i^* долевым коэффициентам

$$j^* = \arg\max_{j \in \mathcal{A}} \sum_{g=t-i^*+1}^{t} q_g^j. \tag{34}$$

Как и для этапа Add считаем по обучающей выборке $\hat{\mathbf{w}}_0$ и $\hat{\mathbf{w}}_{j^*}$

$$\hat{\mathbf{w}}_0 = \arg\max_{\mathbf{w}\in\mathbb{R}^{|\mathcal{A}|}} L(\mathbf{w}|\mathbf{X}(\mathcal{B},\, \mathcal{A}),\, \mathbf{y}(\mathcal{B})),$$
$$\hat{\mathbf{w}}_{j^*} = \arg\max_{\mathbf{w}\in\mathbb{R}^{|\mathcal{A}|-1}} L(\mathbf{w}|\mathbf{X}(\mathcal{B},\, \mathcal{A}\setminus\{j^*\}),\, \mathbf{y}(\mathcal{B})),$$

считаем l_0 и l_{j^*}

$$l_0 = l(\hat{\mathbf{w}}_0|\mathbf{X}(\mathcal{T}_1,\, \mathcal{A},\, \mathbf{y}(\mathcal{T}_1)),$$
$$l_{j^*} = l(\hat{\mathbf{w}}_{j^*}|\mathbf{X}(\mathcal{T}_1,\, \mathcal{A}\setminus\{j^*\},\, \mathbf{y}(\mathcal{T}_1)). \tag{35}$$

Если $l_{j^*} - l_0 \le Z_2$, то есть не происходит значительного ухудшения качества модели, признак j^* исключается из модели, то есть $\mathcal{A} \to \mathcal{A}\setminus\{j^*\}$. Этап повторяется до тех пор, пока происходит удаление признаков из модели.

Поочередное повторение этапов Add и Del осуществляется до тех пор, пока происходит удаление или добавление признаков. Заметим, что условием остановки алгоритма является $Z_1 + Z_2 < 0$.

2.2 Предлагаемый метод отбора объектов и фильтрация выбросов

Отбор объектов наряду с отбором признаков очень важен для построения качественной и устойчивой модели. Под выбросом будем понимать объект, добавление которого в модель значимо влияет на ее параметры. Для того, чтобы определить влияние объекта $(\mathbf{x}_i,\, y_i)$ на модель определим его *специфичность* следующим образом

$$Sp(\mathbf{x}_i) = (\Delta_i\mathbf{w})^\intercal \mathbf{H} (\Delta_i\mathbf{w}), \tag{36}$$

$$\Delta_i\mathbf{w} = \hat{\mathbf{w}}_i - \hat{\mathbf{w}},$$

$\mathbf{H}^{-1}$ – ковариационная матрица оценок параметров модели $\hat{\mathbf{w}}$, а $\hat{\mathbf{w}}_i$ и $\hat{\mathbf{w}}$ оценки параметров модели, полученные согласно (4) по выборке без и с объектом $(\mathbf{x}_i,\, y_i)$ соответственно, то есть

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}\in\mathbb{R}^{|\mathcal{A}|}} L(\mathbf{w}|\mathbf{X}(\mathcal{S},\, \mathcal{A}),\, \mathbf{y}(\mathcal{S})), \tag{37}$$
$$\hat{\mathbf{w}}_i = \arg\max_{\mathbf{w}\in\mathbb{R}^{|\mathcal{A}|}} L(\mathbf{w}|\mathbf{X}(\mathcal{S}\setminus\{i\},\, \mathcal{A}),\, \mathbf{y}(\mathcal{S}\setminus\{i\})),$$

где $\mathcal{A}$ – некоторое ранее отобранное множество признаков, а $\mathcal{S}$, как и ранее, множество индексов объектов обучения.

Относительно параметров модели $\mathbf{w}$ предполагаем априорное равномерное псевдораспределение на всем пространстве. Тогда апостериорное распределение параметров модели $\hat{\mathbf{w}}$ локально нормально, то есть $\mathbf{w} \sim N(\hat{\mathbf{w}},\, \mathbf{H}^{-1})$, то есть в условиях предположения о том, что объект $(\mathbf{x}_i,\, y_i)$ не является выбросом, $\Delta_i\mathbf{w} \sim N(\mathbf{0},\, \mathbf{H}^{-1})$. Матрицу $\mathbf{H}$ считаем невырожденной (хотя, возможно, плохо обусловленной), а потому в условиях предположения о том, что рассматриваемый объект $(\mathbf{x}_i, y_i)$ не является выбросом, получаем

$$Sp(\mathbf{x}_i) = (\Delta_i\mathbf{w})^\intercal \mathbf{H} (\Delta_i\mathbf{w}) \sim \chi^2(|\mathcal{A}|),$$

где $|\mathcal{A}|$ – число признаков в модели. Задавая уровень значимости α, можно отбирать некоторую долю объектов, которые являются выбросами.

Отметим, что вместо априорного равномерного псевдораспределения на всем пространстве для $\mathbf{w}$ можно использовать, например, нормальное. В таком случае, как отмечено в разделе 2.1, апостериорное распределение параметров также можно приблизить нормальным, но с другой ковариационной матрицей (19).

Модификация алгоритма отбора объектов. В случае, когда матрица $\mathbf{H}$ является вырожденной число степеней свободы будет определяться $rg(\mathbf{H})$, а не $|\mathcal{A}|$. Однако определение точного числа степеней свободы затруднительно в силу вычислительных ошибок. То есть требуется определить некоторый порог $\lambda_0 \geq 0$, что собственное число λ матрицы $\mathbf{H}$ считается нулевым, если $\lambda \leq \lambda_0$. Кроме того, в этом случае затруднено использование метода IRLS, поскольку в нем требуется обращать матрицу $\mathbf{H}$. Для стабильности алгоритма IRLS можно заменить матрицу $\mathbf{H}$ на матрицу $\mathbf{H} + \tau\mathbf{I}$, где $\mathbf{I}$ – единичная матрица соответствующего размера. Это будет соотвествовать априорному предположению о нормальном распределении параметров с ковариационной матрицей $\frac{1}{\tau}\mathbf{I}$ и математическим ожиданием $\hat{\mathbf{w}}$, то есть $\mathbf{w} \sim N\left(\hat{\mathbf{w}},\, \tau\mathbf{I}\right)$. В таком случае апостериорное распределение вектора оценок весов модели будет нормальным с ковариационной матрицей

$$\mathbf{w} \sim N\left(\hat{\mathbf{w}},\, \left(\mathbf{H} + \tfrac{1}{\tau}\mathbf{I}\right)^{-1}\right). \tag{38}$$

Отметим, однако, что для того, чтобы исключить слагаемое, связанное с априорным нормальным распределением из градиента, в качестве априорного среднего использована неизвестная априори оценка максимума правдоподобия. Если же использовать априорное нормальное распредление с центром в нуле, то изменится значение $\hat{\mathbf{w}}$, однако выражение (38) по-прежнему будет справедливо. Выполняя соответствующую замену матрицы $\mathbf{H}$ на матрицу $\mathbf{H} + \dfrac{1}{\tau}\mathbf{I}$ в формуле (36), получаем обновленное определение *специфичности* объекта.

$$Sp(\mathbf{x}_i) = (\Delta_i\mathbf{w})^\intercal \left(\mathbf{H} + \tfrac{1}{\tau}\mathbf{I}\right)(\Delta_i\mathbf{w}). \tag{39}$$

При $\tau \to \infty$ (39) совпадает с (36).

При введении указанной регуляризации появляется неопределенность с выбором τ. Для того, чтобы избежать этой неопределенности, предлагается рассмотреть для каждого признака с номером j оценку дисперсии его параметра

$$D_j = \frac{\sum_{i\in\mathcal{S}} (\Delta_i w_j)^2}{|\mathcal{S}| - 1},$$

а специфичность определить как

$$\hat{Sp}(\mathbf{x}_i) = \sum_j \frac{(\Delta_i w_j)^2}{D_j}. \tag{40}$$

Если при этом специфичности, определенные по формулам (40) и (36), будут иметь подобную зависимость от номера объекта, то имеет смысл применять вторую, как вычислительно более простую и не требующую введения регуляризации на случай плохо обусловленной или вырожденной матрицы $\mathbf{H}$. Именно этот случай и реализуется на практике, как показал вычислительный эксперимент (см.рис. 2).

Отделение малочисленной шумовой компоненты от основной выборки. Если в выборке есть шумовая компонента, влияние которой не является определяющим, то можно ее выделить, с помощью введенной специфичности объектов. Для этого отсортируем значения специфичности объектов по убыванию и нарисуем график зависимости специфичности от номера объекта в отсортированном наборе. Найдем i^*

$$i^* = \arg\max_i \frac{Sp(\mathbf{x}_i)}{Sp(\mathbf{x}_{i+1})},$$

тогда объект $\mathbf{x}_i$ считаем шумовым объектом, если $Sp(\mathbf{x}_i) \geq Sp(\mathbf{x}_{i^*})$.

3 Мультимодельный подход к задаче классификации

3.1 Смесь логистических моделей

Зачастую данные бывают неоднородны, могут представлять совокупность нескольких разных, но схожих между собой наборов объектов. Эта неоднородность может быть вызвана разными причинами. Например, можно предположить, что на возврат кредита состоятельными гражданами величина их дохода влияет слабее (так как они вполне платежеспособны и имеют возможность вернуть кредит), чем более бедными. Эти и подобные замечания обуславливают неоднородность данных.

Для того, чтобы учесть неоднородность, можно было бы предложить разбиение исходного множества объектов на несколько подмножеств. В каждом из этих подмножеств можно построить свою модель логистической регрессии. При появлении очередного объекта нужно решить, к какой совокупности он относится и применять соответствующую этой совокупности модель. Таким образом, наблюдается разделение объектов между моделями. При таком подходе требуется определить правила отнесения объекта к той или иной модели, а также могут наблюдаться проблемы с классификацией объектов, «лежащих далеко от центров» всех совокупностей. Особенности таких (многоуровневых моделей) будут рассмотрены в следующем параграфе и в соответствующем разделе вычислительного эксперимента.

Здесь предлагается модель с мягким разделением между моделями или смесь моделей. При таком разделении для каждой модели есть вероятность $\pi_k \in [0,\, 1]$ того, что объект описывается этой моделью. Предположим, что число моделей $K \geq 1$, каждой из которых соответствует свой вектор весов признаков $\mathbf{w}_k$. Тогда

$$p(y_i|\mathbf{x}_i,\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}_i,\, \mathbf{w}_k)^{y_i}(1 - f(\mathbf{x}_i,\, \mathbf{w}_k))^{1-y_i}. \tag{41}$$

Введем вектор $\boldsymbol{\pi} = [\pi_1 \ldots \pi_K]^\intercal$. Тогда по аналогии с (5) получим для правдоподобия данных выражение

$$L(\mathbf{y}|\mathbf{X},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K,\, \boldsymbol{\pi}) = \prod_{i=1}^{m} \left(\sum_{k=1}^{K} \pi_k f(\mathbf{x}_i,\, \mathbf{w}_k)^{y_i}(1 - f(\mathbf{x}_i,\, \mathbf{w}_k))^{1-y_i} \right). \tag{42}$$

В (42) учтено, что не предполагается случайности в описании объектов $\mathbf{x}_i$, а также, что пока априорные распределения на параметры не вводились, то есть считаем параметры пока неизвестными, но неслучайными. Введем бинарную матрицу скрытых

переменных $\mathbf{Z} = \|z_{ik}\|$ размеров $m \times K$, при этом z_{ik} определяет принадлежность $\mathbf{x}_i$ модели с номером k. Тогда полная функция правдоподобия будет записана в виде

$$L(\mathbf{y},\, \mathbf{Z}|\mathbf{X},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K,\, ,\boldsymbol{\pi}) = \prod_{i=1}^{m}\prod_{k=1}^{K}\{\pi_k f(\mathbf{x}_i,\, \mathbf{w}_k)^{y_i}(1 - f(\mathbf{x}_i,\, \mathbf{w}_k))^{1-y_i}\}^{z_{ik}}. \qquad (43)$$

Заметим, что число моделей в смеси K до сих пор предполагалось фиксированным. Для того, чтобы число моделей было выбрано автоматически, выберем K, заведомо большим, чем число моделей в смеси, которая описывает данные. Введем далее априорное распределение Дирихле на вероятности моделей $\boldsymbol{\pi}$ для поощрения разреженности по компонентам $\boldsymbol{\pi}$ (в результате оптимизации большая часть компонент $\boldsymbol{\pi}$ станет равной нулю).

$$\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\pi}|\alpha), \text{ где } \alpha \in (0,\, 1), \text{ то есть} \qquad (44)$$

$$p(\boldsymbol{\pi}|\alpha) = \begin{cases} 0, \ \min_k \pi_k = 0, \\ \dfrac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha-1}, \text{ иначе.} \end{cases}$$

Тогда совместное правдоподобие имеет вид

$$L(\mathbf{y}, \mathbf{Z}, \boldsymbol{\pi}|\mathbf{X}, \mathbf{w}_1, \ldots, \mathbf{w}_K) = \prod_{i=1}^{m}\prod_{k=1}^{K}\{\pi_k f(\mathbf{x}_i,\, \mathbf{w}_k)^{y_i}(1 - f(\mathbf{x}_i,\, \mathbf{w}_k))^{1-y_i}\}^{z_{ik}} \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha-1}. \qquad (45)$$

Далее опишем применяемый EM-алгоритм для оценки параметров моделей $\mathbf{w}_1,\, \ldots,\, \mathbf{w}_K$ и их вероятностей $\pi_1, \ldots, \pi_K$. Выберем некоторые начальные приближения для $\mathbf{w}_1,\, \ldots,\, \mathbf{w}_K$. На E-шаге считаем апостериорные вероятности каждой из компонент смеси для каждого объекта $\mathbf{x}_i$ (они задают апостериорное распределение скрытых бинарных переменных z_{ik}) γ_{ik}

$$\gamma_{ik} = \mathbb{E}[z_{ik}] = p(k|\mathbf{x}_i,\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K,\, \boldsymbol{\pi}) = \frac{\pi_k f(\mathbf{x}_i,\, \mathbf{w}_k)^{y_i}(1 - f(\mathbf{x}_i,\, \mathbf{w}_k))^{1-y_i}}{\sum_{j=1}^{K} \pi_j f(\mathbf{x}_i,\, \mathbf{w}_j)^{y_i}(1 - f(\mathbf{x}_i,\, \mathbf{w}_j))^{1-y_i}}.$$

Запишем ожидаемое значение отрицательного логарифма полной функции правдоподобия

$$\tilde{l}(\mathbf{y},\, \boldsymbol{\pi}|\mathbf{X},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K) = \mathbb{E}_{\mathbf{Z}}[-\log L(\mathbf{y},\, \boldsymbol{\pi},\, \mathbf{Z}|\mathbf{X},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K)] =$$
$$= -\sum_{i=1}^{m}\sum_{k=1}^{K}\gamma_{ik}\{\log \pi_k + y_i \log(f(\mathbf{x}_i,\, \mathbf{w}_k)) + (1 - y_i)\log(1 - f(\mathbf{x}_i,\, \mathbf{w}_k))\} + \sum_{k=1}^{K}(\alpha - 1)\log \pi_k + \text{const}. \qquad (46)$$

На M-шаге происходит минимизация функции $\tilde{l}(\mathbf{y}, \boldsymbol{\pi}|\mathbf{X}, \mathbf{w}_1, \ldots, \mathbf{w}_K)$ по $\mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}$ при ограничении $\sum_{k=1}^{K} \pi_k = 1$. Решение задачи минимизации для $\boldsymbol{\pi}$ в явном виде дает

$$\pi_k = \begin{cases} 0, \ \sum_{i=1}^{m} \gamma_{ik} + \alpha - 1 < 0, \\ \dfrac{\sum_{i=1}^{m} \gamma_{ik} + \alpha - 1}{\sum_{l:\gamma_{il}+\alpha-1>0} (\sum_{i=1}^{m} \gamma_{il} + \alpha - 1)}, \text{ иначе.} \end{cases}$$

Отсюда получаем, что в процессе итераций EM-алгоритма часть компонент смеси будет исключена из рассмотрения, причем при $\alpha \to 0$ будут исключаться все новые компоненты смеси, а потому выбор гиперпараметра α определяет то, сколько компонент будет в смеси. Значение гиперпараметра α можно выбирать, например, кросс-валидацией.

Заметим, что при фиксированных $\{\gamma_{ik}\}$ функция $\tilde{l}(\mathbf{y},\, \boldsymbol{\pi}|\mathbf{X},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K)$ представима в виде

$$\tilde{l}(\mathbf{y},\, \boldsymbol{\pi}|\mathbf{X},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K) = -\sum_{k=1}^{K}\{\log \pi_k \left(\sum_{i=1}^{m} \gamma_{ik} + \alpha - 1\right)\} + \sum_{k=1}^{K} \tilde{l}_k(\mathbf{w}_k|\mathbf{X},\, \mathbf{y}). \quad (47)$$

Поэтому минимизация каждой из функций $\tilde{l}_k(\mathbf{w}_k|\mathbf{X},\, \mathbf{y})$ одного из векторов весов $\mathbf{w}_k$ производится независимо от остальных с помощью описанного выше метода IRLS. Но изменится выражение для градиента и гессиана. Опуская выкладки, соответствующие дифференцированию, запишем результат.

$$\frac{\partial \tilde{l}_k}{\partial \mathbf{w}_k} = \mathbf{X}^\intercal \boldsymbol{\Gamma}_k(\mathbf{f} - \mathbf{y}), \quad (48)$$

$$\mathbf{H}_k = \mathbf{X}^\intercal \mathbf{R}_k \mathbf{X}, \quad (49)$$

где $\boldsymbol{\Gamma}_k$ – диагональная матрица с элементами γ_{ik} на диагонали, а $\mathbf{R}_k$ – диагональная матрица с элементами $\gamma_{ik} f(\mathbf{x}_i^\intercal \mathbf{w}_k) f(-\mathbf{x}_i^\intercal \mathbf{w}_k)$ на диагонали.

Так как исходная задача минимизации разбивается на K независимых подзадач, возможно примение алгоритма совместного отбора объектов и признаков, приведенного в этой работе для каждой задачи в отдельности.

3.2 Многоуровневые модели

Определение. Многоуровневой регрессионной моделью называется набор регрессионных моделей f_k, $k = 1,\, \ldots,\, K$ такой, что при разбиении множества индексов объектов $\mathcal{I} = \sqcup_{k=1}^{K} \mathcal{I}_k$ для всех объектов из $\mathcal{I}_k$ используется модель f_k.

Опишем правило выбора модели на обучении. Запишем правдоподобие модели f_k

$$p(f_k|\mathbf{x}_i, y_i) = \frac{p(f_k, \mathbf{x}_i, y_i)}{p(\mathbf{x}_i, y_i)} = \frac{p(y_i|f_k, \mathbf{x}_i) p(f_k, \mathbf{x}_i)}{p(\mathbf{x}_i, y_i)}.$$

Рассмотрим две модели, без ограничения общности модели f_1 и f_2 и определим, какая из них предпочтительнее для объекта $(\mathbf{x}_i,\, y_i)$. Для этого запишем отношение правдоподобия моделей

$$\frac{p(f_1|\mathbf{x}_i, y_i)}{p(f_2|\mathbf{x}_i, y_i)} = \frac{p(y_i|f_1, \mathbf{x}_i)}{p(y_i|f_2, \mathbf{x}_i)} \frac{p(f_1)}{p(f_2)}.$$

Модель f_1 будет предпочтительнее, чем f_2, если

$$\frac{p(y_i|f_1, \mathbf{x}_i)}{p(y_i|f_2, \mathbf{x}_i)} > 1 \quad (50)$$

и наоборот. В случае K моделей имеем тогда следующее решающее правило отнесения к модели

$$k_i^* = \arg \max_{k \in \{1..K\}} p(y_i|f_k, \mathbf{x}_i).$$

Однако такой алгоритм выбора ведет к переобучению, а именно уже многоуровневая модель с числом моделей 2 идеально разделяет любую обучающую выборку [47]. Поэтому, и для объектов обучения, и для объектов контроля предлагается следующая процедура выбора модели.

Процедура выбора модели для объектов обучения и контроля. Предлагается использовать осторожный выбор модели

$$k_i^* = \arg \max_{k \in \{1..K\}} \min_{u \in \{0,1\}} p(u|f_k, \mathbf{x}_i),$$

который в случае логистической регрессии принимает вид

$$k_i^* = \arg \max_{k \in \{1..K\}} \{\min(f(\mathbf{x}_i^\mathsf{T}\mathbf{w}_k), f(-\mathbf{x}_i^\mathsf{T}\mathbf{w}_k))\}.$$

Преобразуем это выражение и получаем правило отнесения объектов контроля к модели

$$k_i^* = \arg \max_{k \in \{1..K\}} f(-|\mathbf{x}_i^\mathsf{T}\mathbf{w}_k|) = \arg \min_{k \in \{1..K\}} f(|\mathbf{x}_i^\mathsf{T}\mathbf{w}_k|).$$

С учетом монотонности сигмоидной функции связи получаем окончательное выражение для решающего правила отнесения объектов контроля к моделям

$$k_i^* = \arg \min_{k \in \{1..K\}} |\mathbf{x}_i^\mathsf{T}\mathbf{w}_k|. \tag{51}$$

Формула (51) фактически означает, что с точностью до $\|\mathbf{w}\|$ объект контроля относится к той модели, расстояние до разделяющей гиперплоскости которой меньше (см.рис. 1).

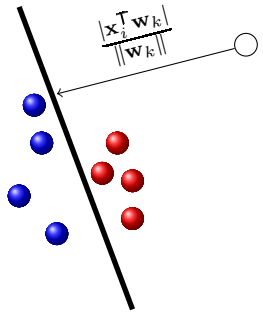


Рис. 1: Иллюстрация отнесения объекта контроля к модели.

Алгоритм отбора объектов для смесей логистических моделей и многоуровневых моделей. Рассмотрим сначала многоуровневые модели. В случае многоуровневых моделей множество индексов объектов $\mathcal{I}$ разбивается на непересекающиеся подмножества $\mathcal{I} = \sqcup_{k=1}^K \mathcal{I}_k$, что для объектов с индексами из $\mathcal{I}_k$ используется модель f_k. Для нового объекта $\mathbf{x}$ требуется определить, к какой модели он относится и проводить классификацию в рамках этой модели. Из-за жесткого разбиения множества объектов на подмножества для многоуровневых моделей также применим

метод фильтрации выбросов, описанный ранее для одной модели логистической регрессии внутри каждой из моделей по отдельности, то есть для объекта $\mathbf{x}_i$, который отнесен к модели с номером k имеем

$$\mathrm{Sp}(\mathbf{x}_i) = (\mathbf{w}_k^{'} - \mathbf{w}_k)^{\intercal}\mathbf{H}_k^{-1}(\mathbf{w}_k^{'} - \mathbf{w}_k),$$

где $\mathbf{w}_k$ – оценка параметров соответствующей модели с объектом $\mathbf{x}_i$, а $\mathbf{w}_k^{'}$ – без этого объекта, а $\mathbf{H}_k$ – гессиан логарифма правдоподобия для модели k.

Рассмотрим теперь случай смеси логистических моделей. Предлагается следующий алгоритм подсчета специфичности объектов.

1. Зафиксируем γ_{jk} с последней итерации EM-алгоритма.
2. Удалим объект $\mathbf{x}_i$ из выборки и пересчитаем $\pi_1, \ldots, \pi_K$
3. Переоптимизируем $\tilde{l}_k(\mathbf{w}_k|\mathbf{X}, \mathbf{y})$ по $\mathbf{w}_1, \ldots, \mathbf{w}_K$
4. Определим частную специфичность объекта $\mathbf{x}_i$ $\mathrm{Sp}_k(\mathbf{x}_i)$ для каждой модели как $\mathrm{Sp}_k(\mathbf{x}_i) = (\mathbf{w}_k^{'} - \mathbf{w}_k)^{\intercal}\mathbf{H}_k^{-1}(\mathbf{w}_k^{'} - \mathbf{w}_k)$
5. Определим общую специфичность объекта $\mathbf{x}_i$ как $\mathrm{Sp}(\mathbf{x}_i) = \sum_{k=1}^{K}\mathrm{Sp}_k(\mathbf{x}_i)$

Замечание:
Отметим, что из оптимальности уже найденных оценок на $\mathbf{w}_1, \ldots, \mathbf{w}_K$ после удаления объекта $\mathbf{x}_i$ из выборки имеем следующие возмущение в норме градиента оптимизируемого по $\mathbf{w}_k$ члена в правдоподобии

$$\nabla\tilde{l}_k(\mathbf{w}_k) = \sum_{j=1,\, j\neq i}^{m}\gamma_{jk}\mathbf{x}_j(f_j - y_j) = -\gamma_{ik}\mathbf{x}_i(f_i - y_i).$$

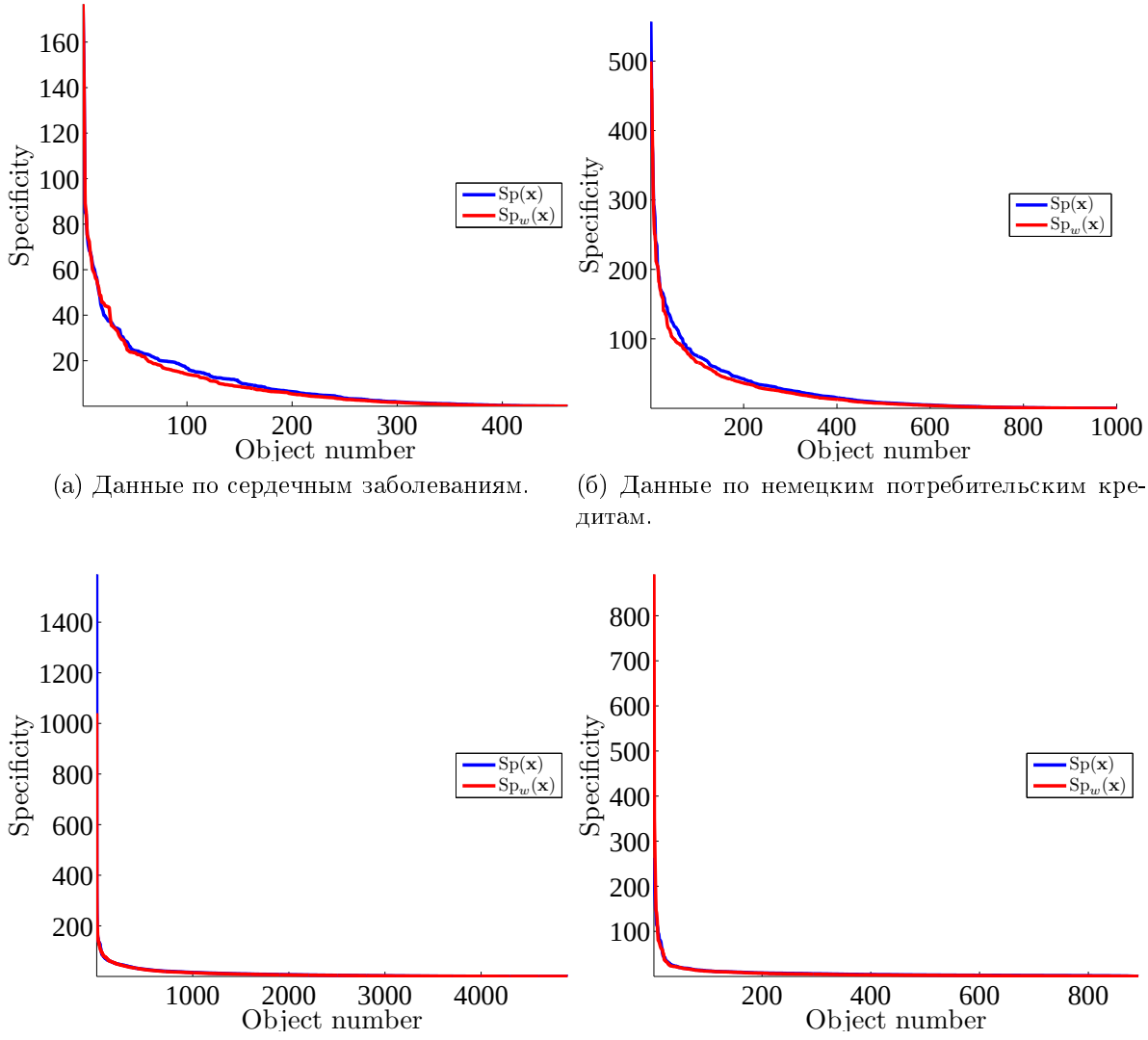
Отсюда заключаем, что объекты, плохо описанные моделью (с большой величиной невязки $f_i - y_i$) и имеющие высокую вероятность принадлежности модели k (большое γ_{ik}) дают больший вклад к отклонение градиента от нуля.

4 Вычислительный эксперимент

4.1 Тестирование предложенного алгоритма фильтрации выбросов

В вычислительном эксперименте анализировался предлагаемый метод отбора объектов, основанный на введенном определении специфичности. Также рассматривался метод отбора признаков, основанный на применении обоснованности модели вместе с методом Белсли. Для анализа метода отбора объектов использовались четыре набора данных из репозитория по машинному обучению UCI. Первый набор данных есть данные по немецким потребительским кредитам (1000 объектов, 24 признака, 2 класса) [6]. Второй набор данных есть данные по заболеваниям сердца в Южной Африке (462 объекта, 11 признаков, 2 класса) [7]. Третий набор данных есть данные по качеству белого вина, качество меняется от 0 до 10 [8]. Задача многоклассовой

классификации для данного набора данных была заменена на двухклассовую следующим образом: значения качества 0-5 были заменены на аттрибут класса 0, а значения качества 6-10 – на аттрибут класса 1. Полученный набор данных имеет 4898 объектов, 11 признаков и 2 класса. Четвертый набор данных есть данные по локализации белков в клетке [9]. Были выбраны два крупнейших класса из этого набора данных. Полученный набор данных имеет 892 объекта, 8 признаков и 2 класса. Все признаки во всех случаях числовые. Решается задача классификации на два класса с помощью логистической регрессии, оценка качества производится с помощью функционала качества AUC [25]. Покажем, что улучшение качества, обусловленное удалением выбросов является статистически значимым.



(а) Данные по сердечным заболеваниям.

(б) Данные по немецким потребительским кредитам.

(в) Данные по качеству белого вина.

(г) Данные по расположению белка в клетке.

Рис. 2: Сравнение двух определений специфичности $\mathrm{Sp}(\mathbf{x})$ и $\mathrm{Sp}_w(\mathbf{x})$.

Сначала сравним два определения специфичности, $\mathrm{Sp}(\mathbf{x}, y)$ (36) и $\mathrm{Sp}_w(\mathbf{x}, y)$ (40). Отсортируем объекты в выьорке по убыванию $\mathrm{Sp}(\mathbf{x}, y)$. Голубые линии на рис. 2 есть график нормализованной специфичности $m\mathrm{Sp}(\mathbf{x}, y)$ в зависимости от номера объекта в отсортированной выборке для четырех рассматриваемых наборов реальных дан-

ных. Отсортируем теперь объекты по убыванию $\mathrm{Sp}_w(\mathbf{x},\, y)$. Красные линии на рис. 2 есть график эмпирической специфичности $\mathrm{Sp}_w(\mathbf{x},\, y)$ в зависимости от номера объекта в отсортированной выборке для четырех рассматриваемых наборов реальных данных. Отметим, что для всех наборов данных рассматриваемые зависимости не имеют скачков. Кроме того, только малая доля объектов имеет высокое значение специфичности $\mathrm{Sp}(\mathbf{x},\, y)$ или эмпирической специфичности $\mathrm{Sp}_w(\mathbf{x},\, y)$.

Далее сосчитаем корреляции Пирсона и Кендалла между $\mathrm{Sp}(\mathbf{x})$ and $\mathrm{Sp}_w(\mathbf{x})$ для всех четырех рассматриваемых наборов данных (см. табл. 2). Для 3 из 4 наборов данных оба корреляционных коэффициента близки к 1, то есть эмпирическая специфичность $\mathrm{Sp}_w(\mathbf{x}, y)$ имеет сильную линейную и монотонную связь со специфичностью $\mathrm{Sp}(\mathbf{x},\, y)$. Отсюда порядок, порожденный эмпирической специфичностью $\mathrm{Sp}_w(\mathbf{x},\, y)$ на выборке почти тот же, что и порядок, порожденный $\mathrm{Sp}(\mathbf{x},\, y)$. Корреляция Пирсона между $\mathrm{Sp}(\mathbf{x})$ и $\mathrm{Sp}_w(\mathbf{x})$ для данных по локализации белка средняя. Однако порядок, порожденный эмпирической специфичностью $\mathrm{Sp}_w(\mathbf{x},\, y)$ на выборке, почти такой же, как порожденный специфичностью $\mathrm{Sp}(\mathbf{x},\, y)$, так как корреляция Кендалла близка к 1. Отсюда заключаем, что эмпирическая специфичность более предпочтительна для отбора объектов, так как ее подсчет не требует знания плохо обусловленной (или даже вырожденной) матрицы $\mathbf{H}$. Теперь оцениваем качество классификации

Таблица 2: Корреляции специфичностей $\mathrm{Sp}(\mathbf{x},\, y)$ (36) и $\mathrm{Sp}_w(\mathbf{x},\, y)$ (40).

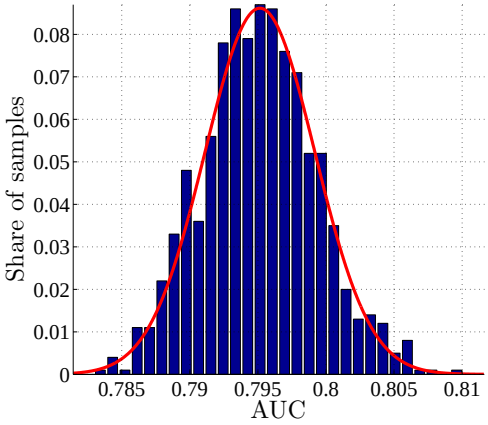
Данные \ Корреляции	Пирсон	Кендалл
SAHD	0.9736	0.9132
Кредиты	0.9794	0.9377
Вино	0.9528	0.9028
Белки	0.5230	0.8597

с помощью AUC [25]. Уберем из выборки небольшую долю объектов, обладающих наибольшей специфичностью $\mathrm{Sp}(\mathbf{x},\, y)$. Таблица 3 показывает увеличение AUC для всех четырех рассматриваемых наборов реальных данных.

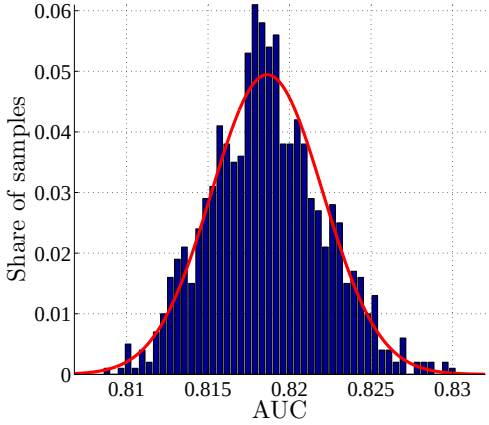
Таблица 3: Изменение AUC после применения процедуры отбора объектов.

Данные	AUC до отбора	AUC после отбора	# удаленных объектов
SAHD	0.7948	0.8275	15 из 462
Кредиты	0.8179	0.8779	50 из 1000
Вино	0.7992	0.8105	48 из 4898
Белки	0.7123	0.7332	18 из 892

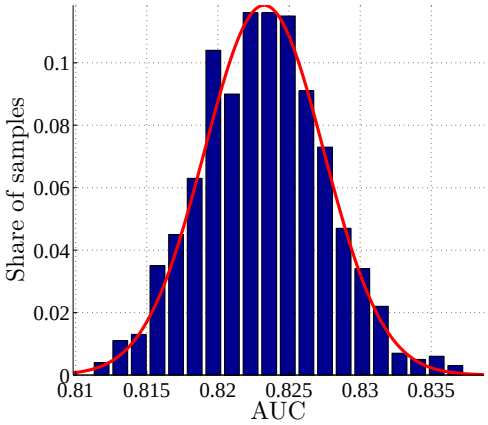
Покажем, что увеличение AUC после отбора объектов статистически значимо. Для доказательства этого рассмотрим в качестве гипотезы H_0 гипотезу о том, что увеличеие AUC вызвано лишь уменьшением размера выборки. Далее для всех рассматриваемых наборов данных делаем следующее. Случайно генерируем подвыборки исходной выборки размера 950 для данных по кредитам, 447 для данных по сердеч-



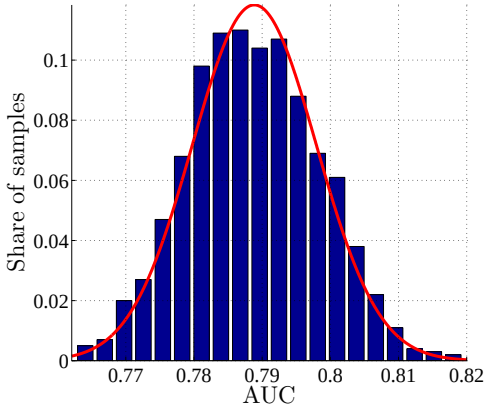
(а) SAHD данные. Вся выборка.



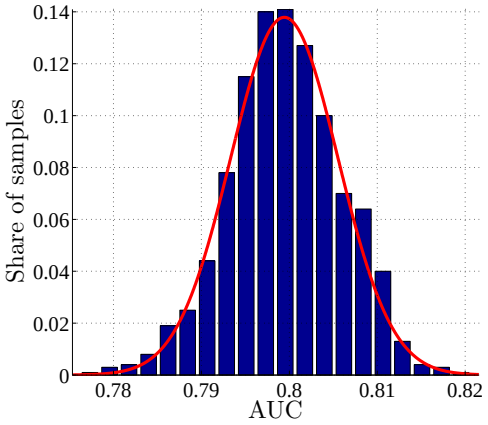
(б) Данные по кредитам. Вся выборка.



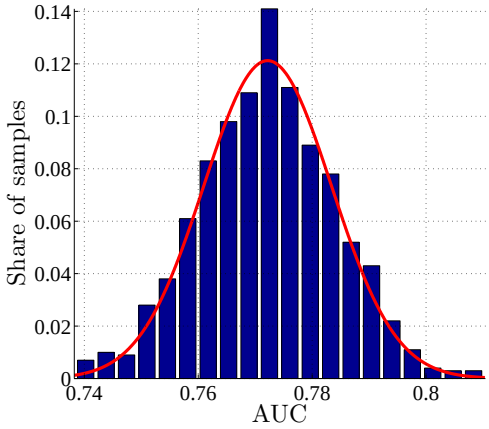
(в) Данные по кредитам. Обучение.



(г) Данные по кредитам. Тестовая выборка.

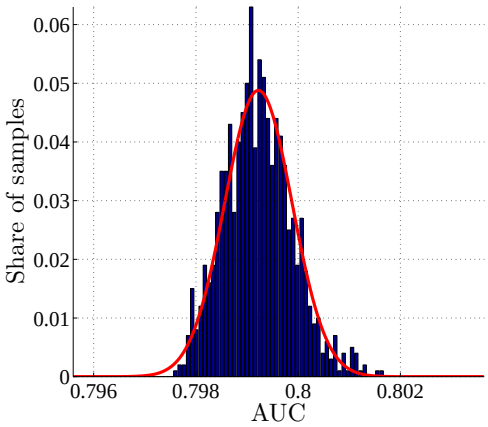


(д) SAHD данные. Обучение.

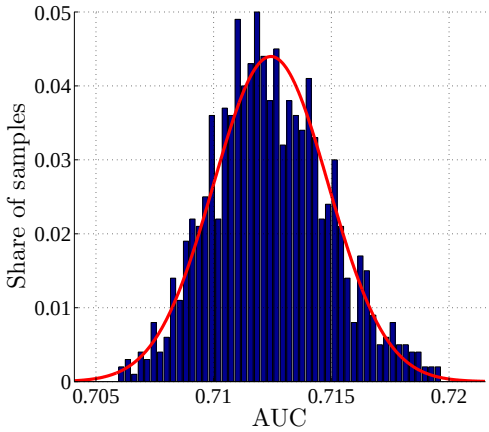


(е) SAHD данные. Тестовая выборка.

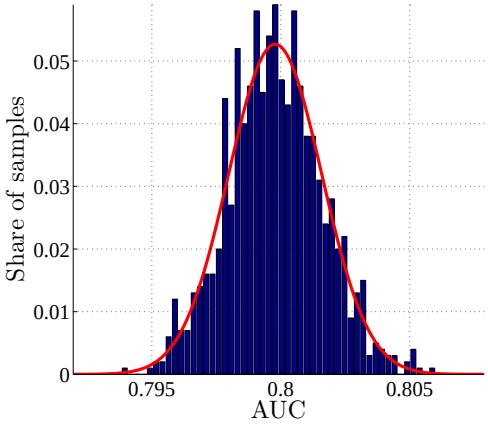
Рис. 3: Эмпирическое распределение AUC и его нормальная аппроксимация для данных по немецким потребительским кредитам и данным по сердечным заболеваниям в Южной Африке.



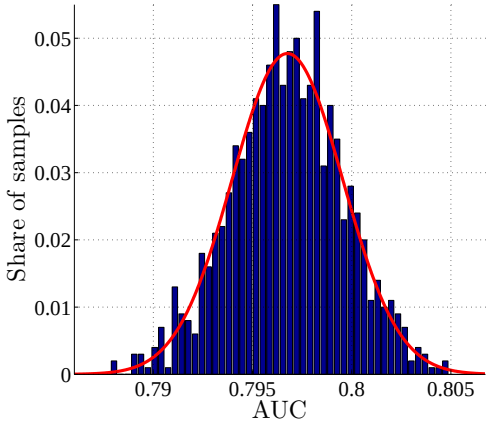
(а) Данные по вину. Вся выборка.



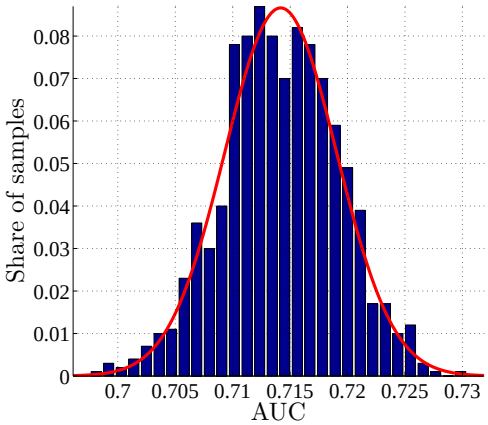
(б) Данные по белкам. Вся выборка.



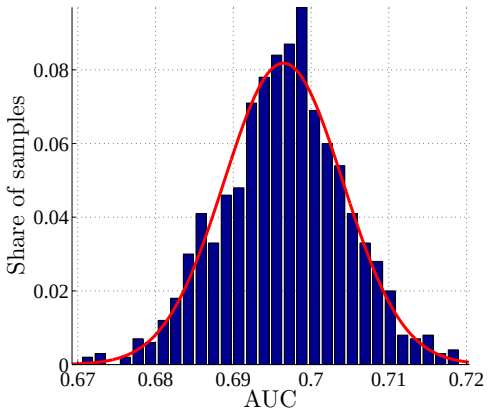
(в) Данные по вину. Обучение.



(г) Данные по вину. Тестовая выборка.



(д) Данные по белкам. Обучение.



(е) Данные по белкам. Тестовая выборка.

Рис. 4: Эмпирическое распределение AUC и его нормальная аппроксимация для данных по качеству вину и локализации белка в клетке.

ным заболеваниям, 4840 для данных по качеству вина и 874 для данных по локализации белка. Обозначим j – ую сгенерированную подвыборку D_j. Используя (2) получаем оценки максимума правдоподобия для параметров модели $\hat{\mathbf{w}}^j$. Сосчитаем соответствующее значение AUC^j. По полученным значениям AUC получаем эмпирическое распределение AUC. Обозначим достигаемый уровень значимости p для гипотезы H_0. Пусть p есть доля подвыборок, для которых AUC^j выше, чем достигаемый на подвыборке того же размера, получаемой после отбора объектов. Так как даже для 1000 сгенерированных подвыборок для всех четырех наборов данных нет ни одной, для которой AUC^j выше, чем достигаемый на подвыборке того же размера, получаемой после отбора объектов, такое определение дает $p = 0$.

Чтобы получить ненулевой достигаемый уровень значимости без генерации очень большого числа подвыборок используем нормальную аппроксимацию эмпирического распределения AUC. Для проверки нормальности используем тест Шапиро-Уилка. Рис. 3 a), b) и 4 a), b) показывают эмпирическое распределение AUC после 1000 генераций подвыборок для полной выборки и нормальные аппроксимации эмпирических распределений.

Тестируем гипотезу H_0 для обучающей и тестовой выборки следующим образом. Генерируем подвыборки исходной выборки 1000 раз. Каждую сгенерированную подвыборку случайно делим на обучающую и тестовую выборки 50 раз. Размер обучающей выборки следующий: 690 для данных по кредитам, 300 для данных по сердечным заболеваниям, 3000 для данных по качеству вина и 550 для данных по локализации белка. Для каждого разбиения подвыборки на обучение и тест параметры модели оцениваются с помощью (4). Получим значения AUC на обучающей и тестовой выборках для каждой подвыбоорки и для каждого из 50 ее разбиений. Усредним полученные 50 значений для каждой сгенерированной подвыборки. Полученные усредненные значения дают эмпирическое распределение AUC на обучающей и тестовой выборке. Рис. 3 c)-f) и 4 c)-f) показывают эмпирическое распределение AUC на обучающей и тестовой выборках вместе с нормальной аппроксимацией для каждого из четырех рассматриваемых наборов данных.

Свойства эмпирических распределений и их нормальных аппроксимаций приведены в табл. 4 и табл. 5. Верхняя часть таблиц содержит результаты при сэмплировании для полной выборке. Для всех рассматриваемых наборов данных достигаемый уровень значимости для гипотезы H_0 равен нулю с машинной точностью, а потому гипотеза H_0 отвергается. Отсюда фильтрация выбросов порождает статистически значимое улучшение качества в терминах AUC.

Средняя часть табл. 4 и табл. 5 содержит значения AUC на обучающей и тестовой выборках после отбора объектов. Нижняя часть табл. 4 и табл. 5 содержит результаты при сэмплировании с разделение на обучающую и тестовую выборки. Для всех рассматриваемых наборов данных (данные по кредитам, сердечным заболеваниям, качеству вина, локализации белка) и на обучающей, и тестовой выборках достигаемый уровень значимости для гипотезы H_0 не превышает 10^{-3}. Это соответсвует более, чем трем стандартным отклонениям от среднего. Поэтому гипотеза H_0 отклоняется. Отсюда отбор объектов порождает статистически значимое увеличение AUC.

Тестирование алгоритма на синтетических данных, содержащих кластеризованные и некластеризованные выбросы. В данном разделе предложенный

Таблица 4: Эмпирическое распределение AUC и его нормальная аппроксимация.

Свойство \ Данные	Кредиты	SAHD
1. Свойства при сэмплировании для полной выборки		
Достиг. p-value в тесте Шапиро-Уилка, p_{SW}	0.0317	0.2035
Оценка мат. ожид. AUC, $\hat{m}$	0.8186	0.7951
Оценка дисперсии AUC, $\hat{\sigma}^2$	$1.17 \cdot 10^{-5}$	$1.63 \cdot 10^{-5}$
Оценка станд. откл. AUC, $\hat{\sigma}$	0.00342	0.00404
Значение t-стат. для гип. H_0, M	17.33	7.39
Достиг. p-value для гип. H_0, p_0	0	0
2. Значения AUC на обучении и тесте после фильтрации выбросов		
AUC на обучении, $\text{AUC}_{\text{learn}}$	0.8819	0.8507
AUC на тесте,AUC_{test}	0.8308	0.8093
3. Свойства при сэмплировании с разделением на обучение и тест		
Достиг. p-value в тесте Шапиро-Уилка, p_{SW}	0.2655; 0.2364	0.2786; 0.7879
Оценка мат. ожид. AUC, $\hat{m}$	0.8233; 0.7889	0.7994; 0.7722
Оценка дисперсии AUC, $\hat{\sigma}^2$	$1.75 \cdot 10^{-5}$; $8.3 \cdot 10^{-5}$	$3.7 \cdot 10^{-5}$; $1.3 \cdot 10^{-4}$
Оценка станд. откл. AUC, $\hat{\sigma}$	0.0042; 0.0091	0.0061; 0.011
Значение t-стат. для гип. H_0, M	14.0; 6.8	5.15; 3.32
Достиг. p-value для гип. H_0, p_0	0; $5.3 \cdot 10^{-12}$	$1.3 \cdot 10^{-7}$; $4.6 \cdot 10^{-4}$

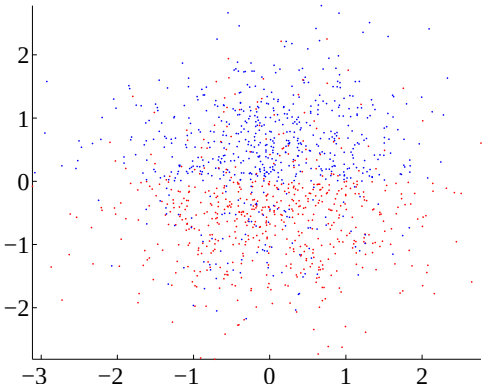
Таблица 5: Эмпирическое распределение AUC и его нормальная аппроксимация.

Свойство \ Данные	Вино	Белки
1. Свойства при сэмплировании для полной выборки		
Достиг. p-value в тесте Шапиро-Уилка, p_{SW}	$1.89 \cdot 10^{-5}$	0.0028
Оценка мат. ожид. AUC, $\hat{m}$	0.7992	0.7124
Оценка дисперсии AUC, $\hat{\sigma}^2$	$4.34 \cdot 10^{-7}$	$5.95 \cdot 10^{-6}$
Оценка станд. откл. AUC, $\hat{\sigma}$	$6.60 \cdot 10^{-4}$	0.0024
Значение t-стат. для гип. H_0, M	17.10	8.51
Достиг. p-value для гип. H_0, p_0	0	0
2. Значения AUC на обучении и тесте после фильтрации выбросов		
AUC на обучении, $\text{AUC}_{\text{learn}}$	0.8109	0.7346
AUC на тесте,AUC_{test}	0.8084	0.7225
3. Свойства при сэмплировании с разделением на обучение и тест		
Достиг. p-value в тесте Шапиро-Уилка, p_{SW}	0.3103; 0.6989	0.5326; 0.4288
Оценка мат. ожид. AUC, $\hat{m}$	0.7998; 0.7968	0.7142; 0.6965
Оценка дисперсии AUC, $\hat{\sigma}^2$	$3.26 \cdot 10^{-6}$; $7.8 \cdot 10^{-6}$	$2.4 \cdot 10^{-5}$; $5.8 \cdot 10^{-5}$
Оценка станд. откл. AUC, $\hat{\sigma}$	0.0018; 0.0028	0.0049; 0.0076
Значение t-стат. для гип. H_0, M	6.15; 4.15	4.18; 3.41
Достиг. p-value для гип. H_0, p_0	$3.9 \cdot 10^{-10}$; $1.7 \cdot 10^{-5}$	$1.4 \cdot 10^{-5}$; $3.2 \cdot 10^{-4}$

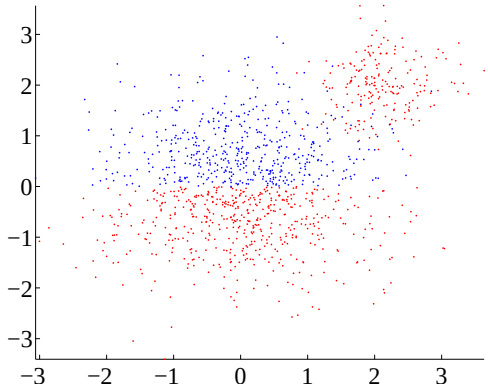
алгоритм фильтрации выбросов тестируется на синтетических данных, содержащих кластеризованные и некластеризованные выбросы. Используя значение AUC, равное 0.7, в качестве границы, определяющей приемлемое качество классификации [23], определим максимальную долю выбросов в данных, при которой AUC находится выше заданной границы.

Данные с некластеризованными выбросами генерируются следующим образом. $\mathbf{x} \in N(\mathbf{0}, \mathbf{I})$, где $\mathbf{I}$ есть 2×2 единичная матрица. $y_i = 1$ для $\mathbf{x}_i$, если $x_2 > 0$ и $y_i = 0$ иначе. Выбросы это объекты, сгенерированные из того же распределения, но с противоположным правилом выбора класса: $y_i = 0$ для $\mathbf{x}_i$, если $x_2 > 0$ и $y_i = 1$ иначе. Рис. 5a показывает сгенерированную выборку из 1000 обычных объектов и 200 выбросов.

Данные с кластеризованными выбросами генерируются следующим образом. Обычные объекты генерируются из $N(\mathbf{0}, \mathbf{I})$, где $\mathbf{I}$ есть 2×2 единичная матрица. Для таких объектов $y_i = 1$ для $\mathbf{x}_i$, если $x_2 > 0$ и $y_i = 0$ иначе. Выбросы генерируются из $N([2, 2]^\intercal, 0.5\mathbf{I})$, где $\mathbf{I}$ есть 2×2 единичная матрица. Все выбросы относятся к классу 0. Рис. 5b показывает сгенерированную выборку из 1000 обычных объектов и 200 выбросов.



(а) Синтетические данные. Некластеризованные выбросы.

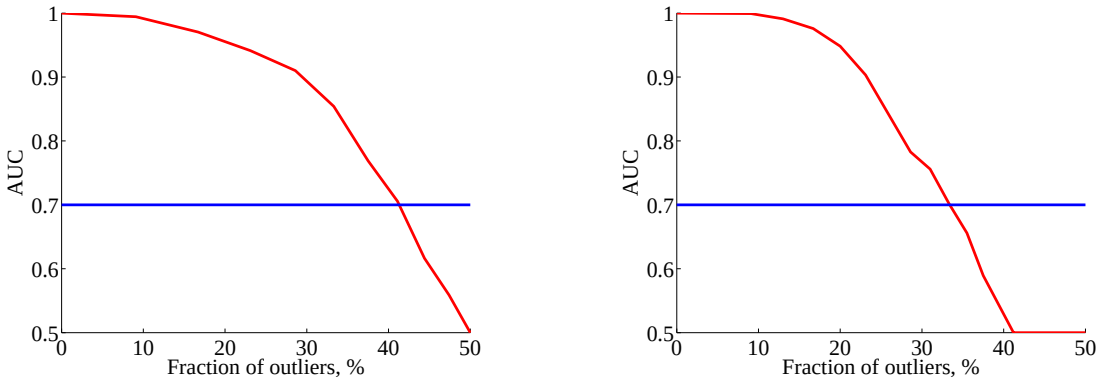


(б) Синтетические данные. Кластеризованные выбросы.

Рис. 5: Синтетические данные с кластеризованными и некластеризованными выбросами.

Рассмотрим зависимость AUC для отбора выбросов для выборок, имеющих от 0 до 50% выбросов. Рис. 6 показывает такую зависимость.

Для выборок с некластеризованными выбросами граница в 0.7 достигается на выборке, имеющей 41.1% выбросов. Для выборок с кластеризованными выбросами та же граница достигается при 33.3% доле выбросов. Эти результаты показывают, что для выборок среднего размера (1000 объектов) рассматриваемый метод применим даже для выборок, имеющих высокую долю выбросов. При этом для выборос с некластеризованными выбросами метод работает лучше. Отметим, что как для выборок с кластеризованными выбросами, так и для выборок с некластеризованными для всех рассмотренных долей выбросов в диапазоне от 0 до 50% корреляции Пирсона между специфичностью и эмпирической специфичностью выше, чем 0.8, а корреляция Кендалла выше, чем 0.7. Отсюда даже для выборок с большой долее выбросов применение эмпирической специфичности вместо специфичности оправдано и имеет смысл особенно в случае плохо обусловленной ковариационной матрицы.



(а) Некластеризованные выбросы. (б) Кластеризованные выбросы.

Рис. 6: Зависимость AUC для отбора выбросов от доли выбросов в выборке.

Сравнение предлагаемого метода фильтрации выбросов с другими. В данном разделе введенная функция специфичности объекта сравнивается с тремя другими функциями, широко используемыми для фильтрации выбросов: байесовыми, пирсоновыми и дисперсными остатками [54]. Однако можно показать, что хотя остатки Пирсона и дисперсные остатки определяются по-разному, они порождают один порядок на выьорке объектов, а потому дают одинаковые результаты при фильтрации выбросов.

Рассматриваемые методы сравниваются с помощью кросс-валидации следующим образом. Выборка разбивается на обучающую и тестовую. Выбросы исключаются из обучающей выборки с помощью каждого из четырех рассматриваемых методов. Обученная на полученной обучающей выборке модель используется для классификации тестовой выборки. Качество на тестовой выборке в терминах AUC сравнивается.

Для сравнения используются четыре набора данных из репозитория UCI и синтетические данные, имеющие кластеризованные и некластеризованные выбросы. Для каждой выборки деление на обучение и тест производится 100 раз. t-статистика Стьюдента используется для проверки значимости отличия в AUC для разных методов. Табл. 6 показывает результаты. Обозначим AUC_p, AUC_b и AUC_s значения AUC для тестовой выборки, полученные после фильтрации выбросов в обучающей выборке с помощью пирсоновых / дисперсных остатков, байесовых остатков и специфичности. t_P и t_B есть t-статистики Стьюдента для разностей AUC_s - AUC_p и AUC_s - AUC_b соответственно.

Табл. 6 показывает результаты для четырех рассматриваемых наборов реальных данных, а также синтетических данных, описанных ранее. Для каждого набора синтетических данных в таблице приведена доля выбросов. Звездочка означает, что фильтруется больше объектов, чем есть выбросов на самом деле. Если метод показывает значительно худше результаты в таком случае, это означает, что он начинает удалять действительно выжные объекты для построения модели. Для наборов реальных данных почти все отличия в качестве незначительны. Однако на синтетических данных метод фильтрации выбросов, основанный на специфичности работает в целом значимо лучше, чем остальные. При этом отличие в качестве особенно значимо для случая, когда удаляется больше объектов, чем есть выбросов. Одним из возможных объяснений является следующее. Как дисперсные / пирсоновы, так и байесовы

Таблица 6: Сравнение разных методов отбора объектов в терминах AUC на тестовой выборке.

Данные \ Метод	Дисп./пирсон.	Байес.	Специфичн.	t_p	t_b
SAHD	**0.7716**	0.7676	0.7661	-1.6395	-0.448
Кредиты	**0.7868**	0.7864	0.7802	-2.7093	-2.5345
Вино	**0.7977**	0.7974	0.7970	-0.8471	-0.4220
Белки	0.6845	**0.6951**	0.6944	5.8773	-0.3997
Синт. некласт., 9.1%	0.8997	**0.9021**	0.9002	0.2450	-1.1300
Синт. некласт.*, 9.1%	0.8945	0.8956	**0.8958**	0.8014	0.1583
Синт. некласт., 23.1%	0.7646	0.7653	**0.7665**	0.7945	0.5036
Синт. некласт.*, 23.1%	0.7671	0.7593	**0.7694**	0.9949	4.3273
Синт. некласт., 33.3%	0.6673	0.6679	**0.6680**	0.6450	0.1075
Синт. некласт.*, 33.3%	0.5372	0.6666	**0.6681**	64.5832	0.7482
Синт. класт., 9.1%	0.8885	0.9261	**0.9269**	20.9410	0.4443
Синт. класт.*, 9.1%	0.8740	0.9515	**0.9541**	66.9012	2.1318
Синт. класт., 16.7%	0.8393	**0.8471**	0.8456	2.5400	-0.6264
Синт. класт.*, 16.7%	0.8379	0.8305	**0.9060**	44.4005	49.1751
Синт. класт., 23.1%	0.8107	0.8171	**0.8174**	3.4906	0.1210
Синт. класт.*, 23.1%	0.8105	0.7923	**0.8113**	0.2828	6.5297
Синт. класт., 33.3%	**0.7860**	0.7856	0.7853	-0.4075	-0.1803
Синт. класт.*, 33.3%	0.7675	**0.7762**	0.7671	-0.1078	-2.4158

остатки определяют в качестве выбросов объекты, плохо описанные моделью. Такие процедуры хорошо работают для выборок с небольшим число выбросов [18]. Однако для выборок со значительным числом выбросов такие методы могут быть неэффективны. Метод, основанный на специфичности, более эффективен, так как он оценивает влияние объекта на модель и ее стабильность, а не то, насколько хорошо объект описан моделью (хотя в среднем, как уже указывалось именно объекты, плохо описанные моделью наибольшим образом влияют на оценку параметров).

4.2 Тестирование предложенного алгоритма отбора признаков

В данном разделе рассмотрим предлагаемый алгоритм отбора признаков на разных наборах синтетических данных, имеющих разные наборы признаков с разными значениями корреляции между ними, а также по-разному коррелированными с выходным вектором $\mathbf{y}$. Отметим исследование некоторых другим методов отбора признаков, которое было проведено на сходных синтетических данных [55]. Рассматриваем данные, попрожденные в соответствии с моделью логистической регрессии. Такие данные порождаем следующим образом.

Пусть $\mathbf{X}$ – матрица признаков, а $\mathbf{w}$ – истинный вектор параметров модели. Для каждого объекта определим вероятность $p_i \in [0,\, 1]$ принадлежности классу 1 в соответствии с (1). Для каждого объекта с признаковым описанием $\mathbf{x}_i$ и вероятностью принадлежности классу 1, равной p_i, сгенерируем реализацию целевой бернуллевской случайной величины $y_i \sim \mathrm{Be}(y_i|p_i)$. В качестве выборки рассматриваем пару $(\tilde{\mathbf{X}},\, \mathbf{y})$, где $\mathbf{y}$ есть вектор реализаций y_i, а $\tilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - \boldsymbol{\varepsilon}) + \boldsymbol{\Delta}\boldsymbol{\varepsilon}$, где $\boldsymbol{\Delta}$ – шумовая матрица с независимыми компонентами из распределения $U[-1.5,\, 1.5]$, а $\boldsymbol{\varepsilon} = \mathrm{diag}(\varepsilon_1,\, \ldots,\, \varepsilon_m)$, диагональная матрица, отращающая степень зашумленности каждого из признаков. Значения одного признака для всех объектов будем сэмплировать независимо из распределения $U[-1.5,\, 1.5]$. В случае коррелированности признаков таким образом будем сэмплировать ортогональную компоненту одного из признаков, а также всю матрицу $\boldsymbol{\Delta}$. Рассматриваем следующие случаи для $\mathbf{X}$, $\mathbf{w}$ и $\boldsymbol{\varepsilon}$.

1. Два активных независимых признака, $\mathbf{w} = (1,\, 1)^\intercal$;

2. Один активный независимый признак и один избыточный, независимый от активного, $\mathbf{w} = (1,\, 0)^\intercal$

3. Рассматриваем $n = 10$ совпадающих признаков, то есть $\mathbf{X}$ состоит из одинаковых столбцов, $\mathbf{w} = (1,\, 0,\, \ldots,\, 0)^\intercal$, $\varepsilon_1 = \ldots = \varepsilon_m$.

4. Тройка мультиколлинеарных признаков (третий равен сумме первых двух), $\mathbf{X}(1,\, 1,\, -1)^\intercal = \mathbf{0}$, $\varepsilon_1 = \varepsilon_2 = \varepsilon_3$, $\mathbf{w} = (1,\, 1)^\intercal$.

5. Две пары мультиколлинеарных признаков (первый и второй, третий и четвертый), $\mathbf{X}(1,\, -1,\, 0,\, 0)^\intercal = \mathbf{0}$, $\mathbf{X}(0,\, 0,\, 1,\, -1)^\intercal = \mathbf{0}$, $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon_4$, $\mathbf{w} = (1,\, 0,\, 1,\, 0)^\intercal$.

Будем рассматривать все приведенные случай последовательно. Размер выборки во всех примерах равен $m = 1000$.

Случай двух активных независимых признаков. В данном случае $\mathbf{w} = [1, 1]^\intercal$, а матрица $\mathbf{X}$ состоит из независимо сгенерированных покомпонентно из распределения $U[-1.5,\ 1.5]$ столбцов. Будем изменять интенсивность зашумления каждого из признаков ε_1 и ε_2. Рассматриваем предлагаемый метод оценки ковариационной матрицы как в множестве диагональных матриц с положительными элементами на диагонали, так и в множестве всех положительно определенных симметричных матриц соотвествующего размера.

Рассмотрим как одинаковую интенсивность зашумления $\varepsilon_1 = \varepsilon_2 = \varepsilon$, так и неодинаковую интенсивность $\varepsilon_1 \neq \varepsilon_2$. В случае одинаковой интенсивности зашумления изменяем ε от 0 до 1, выполняем отбор признаков с помощью предложенного метода для диагональных матриц ковариации и матриц ковариации общего вида. Для полученных оценок $\mathbf{w}_{MP}$ в каждом из способов вычисляем следующие меры качества

$$Q_1 = \frac{1}{m}\|\hat{\mathbf{p}} - \mathbf{p}\|_1,\ Q_2 = \sqrt{\frac{1}{m}\|\hat{\mathbf{p}} - \mathbf{p}\|_2^2},\ \text{где}$$

$\hat{\mathbf{p}}$ есть вычисленная с помощью (1) при найденном $\mathbf{w}_{MP}$ вероятность принадлежности классу 1, а $\mathbf{p}$ есть истинная такая верояность, найденная с помощью (1) при $\mathbf{w} = (1,\ 1)^\intercal$ по незашумленным признакам.

Одинаковая интенсивность зашумления. В случае одинаковой интенсивности зашумления рассматриваем следующие значения $\varepsilon \in \{0,\ 0.1,\ 0.3,\ 0.5,\ 0.8,\ 0.9,\ 0.95\}$. В табл. 7 приведены результаты при разных значениях ε.

Таблица 7: Результаты отбора признаков с помощью максимизации обоснованности модели

ε	$\mathbf{A}^{-1}$	Q_1^{diag}	Q_1^{full}	Q_2^{diag}	Q_2^{full}	$\mathbf{w}_{\text{MP}}^{\text{diag}}$	$\mathbf{w}_{\text{MP}}^{\text{full}}$
0	$\begin{pmatrix} 0.8385 & 0.7981 \\ 0.7981 & 0.7643 \end{pmatrix}$	0.0188	0.0180	0.0207	0.0197	$(0.905,\ 0.866)^\intercal$	$(0.910,\ 0.872)^\intercal$
0.1	$\begin{pmatrix} 1.0253 & 0.9725 \\ 0.9725 & 0.9227 \end{pmatrix}$	0.0271	0.0265	0.0327	0.0321	$(1.004,\ 0.952)^\intercal$	$(1.010,\ 0.958)^\intercal$
0.3	$\begin{pmatrix} 1.2632 & 1.1806 \\ 1.1806 & 1.1037 \end{pmatrix}$	0.0776	0.0774	0.0966	0.0964	$(1.114,\ 1.040)^\intercal$	$(1.120,\ 1.047)^\intercal$
0.5	$\begin{pmatrix} 0.7791 & 0.7437 \\ 0.7437 & 0.7102 \end{pmatrix}$	0.1435	0.1433	0.1759	0.1758	$(0.8700,\ 0.8293)^\intercal$	$(0.8782,\ 0.8385)^\intercal$
0.8	$\begin{pmatrix} 0.0428 & 0.0477 \\ 0.0477 & 0.0477 \end{pmatrix}$	0.2057	0.2054	0.2410	0.2410	$(0.178,\ 0.205)^\intercal$	$(0.198,\ 0.222)^\intercal$
0.9	$\begin{pmatrix} 0.0036 & 0.0047 \\ 0.0047 & 0.0064 \end{pmatrix}$	0.2115	0.2110	0.2465	0.2460	$(10^{-9},\ 0.036)^\intercal$	$(0.049,\ 0.062)^\intercal$
0.95	$10^{-5}\begin{pmatrix} 7.1 & 2.6 \\ 2.6 & 11.6 \end{pmatrix}$	0.2119	0.2119	0.2467	0.2467	$(0.,\ 0.)^\intercal$	$(5 \cdot 10^{-4},\ 0.010)^\intercal$

Из табл. 7 заключаем, что, когда $\mathbf{w}_{\text{MP}}^{\text{full}}$ существенно отличен от нуля, то в формуле (**??**) слагаемое $\mathbf{w}_{MP}\mathbf{w}_{MP}^\intercal$ является определяющим, а потому матрица $\mathbf{A}^{-1}$ близка к соответствующей вырожденной матрице ранга 1. Подобное поведение наблюдается и для других конфигураций данных. Близость матрицы $\mathbf{A}^{-1}$ к матрице ранга 1

затрудняет ее применение в методе Белсли, поскольку в таком случае избыточными могут оказаться все признаки, кроме одного. Для того, чтобы преодолеть эту проблему предлагается ввести априорное распределение на матрицу $\mathbf{A}$, которое будет поощрять матрицы большого ранга. Конкретный вид такого распределения и соответствующие экспериментальный результаты есть предмет дальнейшего исследования.

Отметим, что для полной матрицы ковариаций качество предсказаний в терминах Q_1 и Q_2 несколько выше. Однако различие крайне мало и не может считаться значимым. Приведем далее зависимости $\log \alpha_j$, $j = 1, 2$ в случае диагональной матрицы $\mathbf{A} = \mathrm{diag}(\alpha_1, \alpha_2)$ от номера итерации (см. рис. 7).

Из рис. 7 заключаем, что среднее число итераций до сходимости менее 10, а большее число итераций требуется только в пограничном случае ($\varepsilon = 0.9$). Отметим также, что при $\varepsilon = 0.95$ оба признака признаются неинформативными, а при $\varepsilon = 0.9$ – один из них. Перейдем далее к случай разной интенсивности зашумления для двух признаков.

Разная интенсивность зашумления для признаков $\varepsilon_1 \neq \varepsilon_2$. Для разной интенсивности зашумления для признаков рассмотрим три пары значений интенсивностей: $(\varepsilon_1, \varepsilon_2) \in \{(0., 0.99), (0.1, 0.9), (0.3, 0.8)\}$.

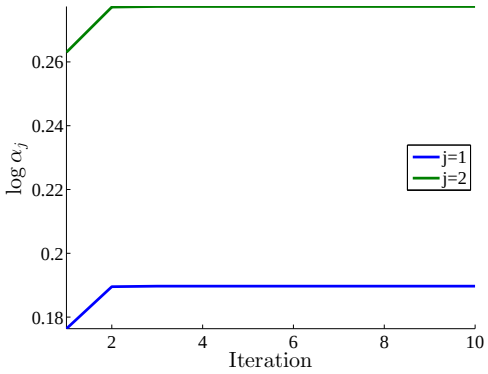
Таблица 8: Результаты отбора признаков с помощью максимизации обоснованности модели

ε_1	ε_2	$\mathbf{A}^{-1}$	Q_1^{diag}	Q_1^{full}	Q_2^{diag}	Q_2^{full}	$\mathbf{w}_{\mathrm{MP}}^{\mathrm{diag}}$	$\mathbf{w}_{\mathrm{MP}}^{\mathrm{full}}$
0	0.99	$\begin{pmatrix} 0.6674 & 0.0395 \\ 0.0395 & 0.0024 \end{pmatrix}$	0.1509	0.1509	0.1735	0.1737	$(0.811, 0.)^\intercal$	$(0.813, 0.048)^\intercal$
0.1	0.9	$\begin{pmatrix} 0.8357 & 0.1353 \\ 0.1353 & 0.0222 \end{pmatrix}$	0.1499	0.1495	0.1727	0.1728	$(0.907, 0.095)^\intercal$	$(0.910, 0.148)^\intercal$
0.3	0.8	$\begin{pmatrix} 1.0934 & 0.3033 \\ 0.3033 & 0.0843 \end{pmatrix}$	0.1519	0.1515	0.1795	0.1795	$(1.037, 0.252)^\intercal$	$(1.040, 0.289)^\intercal$

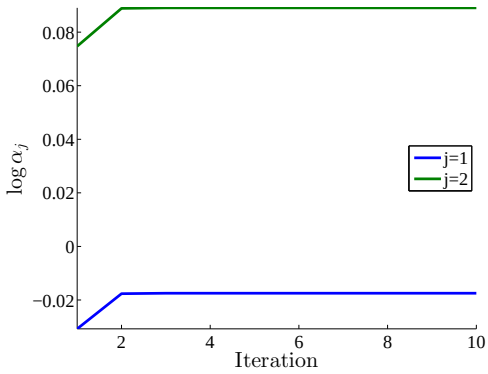
Из табл. 8 заключаем, что, уменьшения веса или полное исключение признака происходит для более зашумленного признака, причем тем сильнее, чем больше интенсивность шума ε_2. Приведем далее зависимости $\log \alpha_j$, $j = 1, 2$ в случае диагональной матрицы $\mathbf{A} = \mathrm{diag}(\alpha_1, \alpha_2)$ от номера итерации (см. рис. 8).

Из рис. 8 заключаем, что сходимость итеративного метода происходит за несколько первых итераций и примерно за 20 итераций для случая исключения одного из признаков. Также отметим, что оценка обратной априорной дисперсии выше для признака с большой интенсивностью зашумленности.

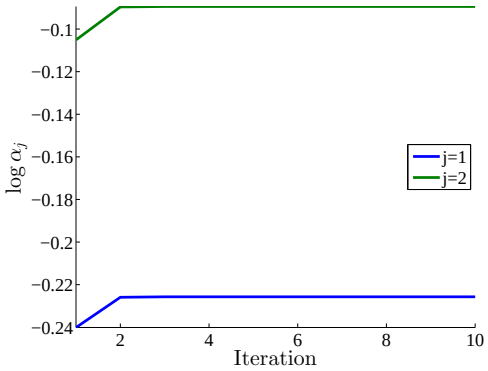
Случай, когда есть один информативный признак и один избыточный, независимый от активного. Рассмотрим теперь случай, когда матрица $\mathbf{X}$ имеет $n = 2$ столбца, первый признак является информативным, а второй неинформативным, независимым от первого. Генерацию выборки производим с $\mathbf{w} = [1, 0]^\intercal$. Интенсивности зашумленности признаков считаем одинаковыми для двух признаков, то есть $\varepsilon_1 = \varepsilon_2 = \varepsilon$. Приведем для разных ε сравнения полученных с помощью каждого из методов результатов (см. табл. 9).



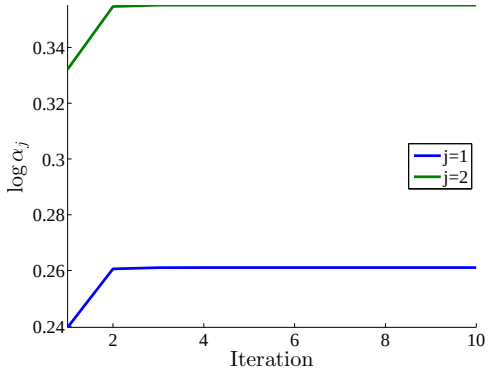
(а) $\varepsilon = 0.$



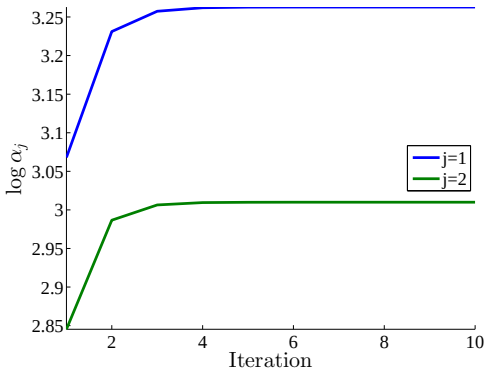
(б) $\varepsilon = 0.1$



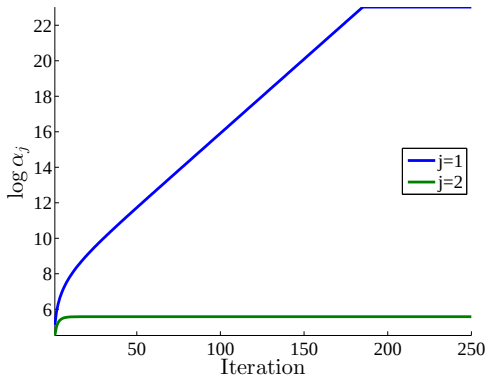
(в) $\varepsilon = 0.3$



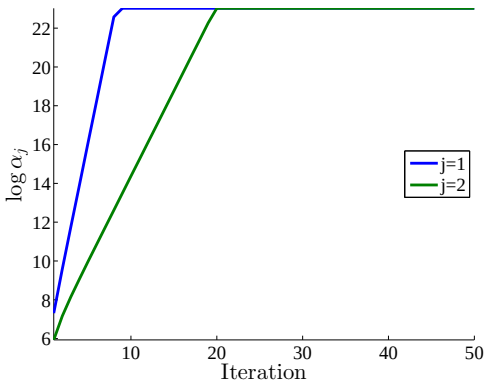
(г) $\varepsilon = 0.5$



(д) $\varepsilon = 0.8$

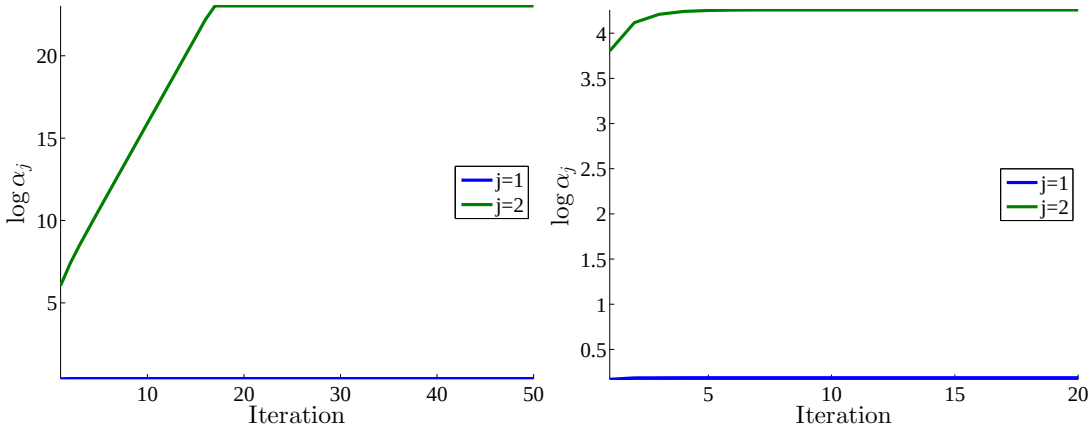


(е) $\varepsilon = 0.9$

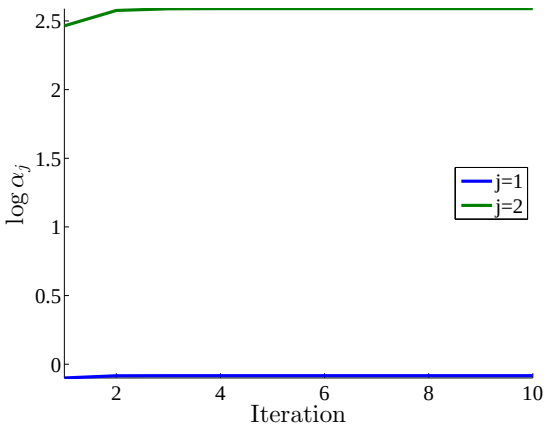


(ж) $\varepsilon = 0.95$

Рис. 7: Зависимость оценок обратных априорных дисперсий α_1, α_2 от номера итерации в зависимости от интенсивности зашумления ε



(а) $(\varepsilon_1,\, \varepsilon_2) = (0.,\, 0.99)$ (б) $(\varepsilon_1,\, \varepsilon_2) = (0.1,\, 0.9)$



(в) $(\varepsilon_1,\, \varepsilon_2) = (0.3,\, 0.8)$

Рис. 8: Зависимость оценок обратных априорных дисперсий α_1, α_2 от номера итерации в зависимости от интенсивностей зашумления ε_1, ε_2

Таблица 9: Результаты отбора признаков с помощью максимизации обоснованности модели

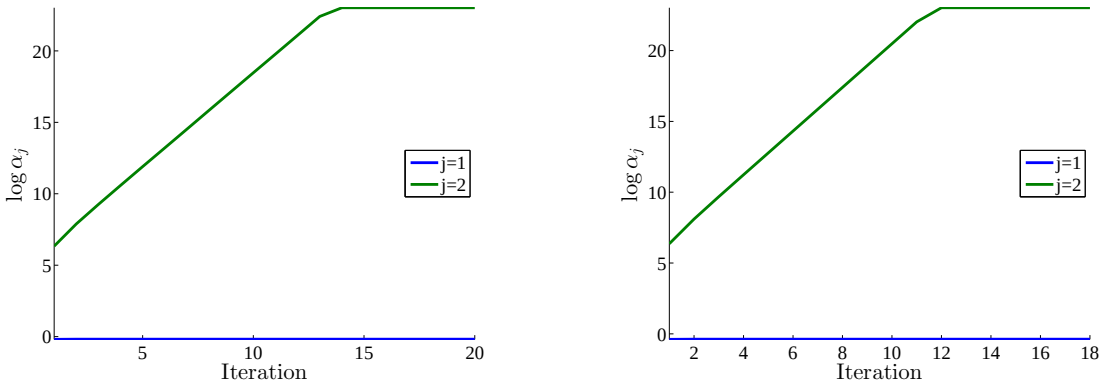
ε	0	0.1	0.3	0.5	0.6	0.7	0.75	0.8	0.9	0.95	0.99
w_1^{diag}	1.091	1.204	1.340	1.092	0.824	0.547	0.421	0.307	0.112	0.017	0
w_1^{full}	1.092	1.205	1.341	1.092	0.823	0.546	0.420	0.308	0.125	0.058	0.022
w_2^{diag}	0	0	0	0	0	0	0.008	0.024	0.041	0.044	0.046
w_2^{full}	-0.042	-0.041	-0.025	0.027	0.051	0.086	0.093	0.095	0.084	0.069	0.054
Q_1^{diag}	0.013	0.021	0.064	0.113	0.134	0.152	0.159	0.164	0.171	0.172	0.172
Q_1^{full}	0.014	0.022	0.064	0.113	0.134	0.152	0.159	0.165	0.171	0.172	0.172
Q_2^{diag}	0.014	0.025	0.076	0.138	0.163	0.180	0.186	0.190	0.194	0.195	0.195
Q_2^{full}	0.016	0.026	0.076	0.139	0.163	0.180	0.186	0.190	0.194	0.195	0.195
Тест. Q_1^{diag}	0.013	0.021	0.064	0.113	0.135	0.153	0.160	0.166	0.172	0.174	0.174
Тест. Q_1^{full}	0.014	0.022	0.064	0.114	0.135	0.153	0.160	0.166	0.172	0.173	0.174
Тест. Q_2^{diag}	0.014	0.025	0.076	0.140	0.165	0.181	0.187	0.190	0.195	0.195	0.196
Тест. Q_2^{full}	0.016	0.026	0.076	0.140	0.165	0.181	0.187	0.191	0.195	0.195	0.196

Результаты из табл. 9 показывают, что при интенсивности зашумления до $\varepsilon = 0.7$ включительно информативный признак признается информативным, а неинформативный фильтруется как избыточный. При дальнейшем увеличении интенсивности зашумления фильтрации избыточного признака не происходит,а при $\varepsilon = 0.99$ метод оценки ковариационной матрицы в множестве диагональных удаляет информативнй признак как избыточный. Такое поведение метода отбора признаков вызвано тем, что корреляция зашумленного информативного признака с целевым вектором $\mathbf{y}$ равна 0.0196, а для зашумленного информативного – 0.0446. Приведем на рис. 9– 11 зависимости оценок обратных априорных дисперсий α_j параметров w_j от номера итерации для всех рассматривавшихся значений интенсивности зашумления ε.

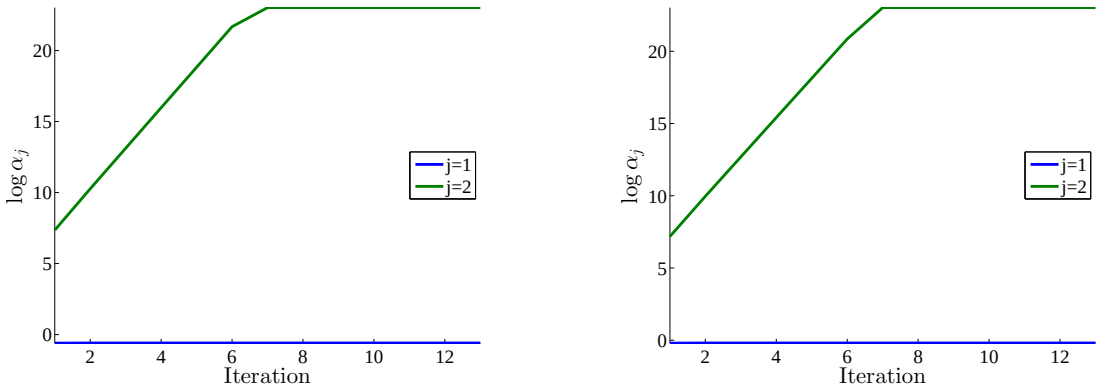
В рассматриваемом случае фильтрация неинформативного признака происходила вплоть до интенсивности зашумления $\varepsilon = 0.7$.

Случай нескольких повторяющихся признаков. Рассмотрим теперь случай, когда матрица $\mathbf{X}$ имеет $n = 10$ одинаковых столбцов, а интенсивность зашумления одинакова для всех признаков, то есть $\varepsilon_1 = \ldots = \varepsilon_{10} = \varepsilon$. В качестве интенсивности зашумления рассматриваем $\varepsilon \in \{0, 0.1, 0.3, 0.5, 0.8, 0.9, 0.95, 0.99\}$. Генерацию выборки производим с $\mathbf{w} = [1,\ 0,\ \ldots,\ 0]^\intercal$.

Будем называть признак с номером j активным, если модуль соответствующего параметра $|w_j| \geq 0.01$. Приведем для разных ε сравнения полученных с помощью каждого из методов результатов (см. табл. 10).

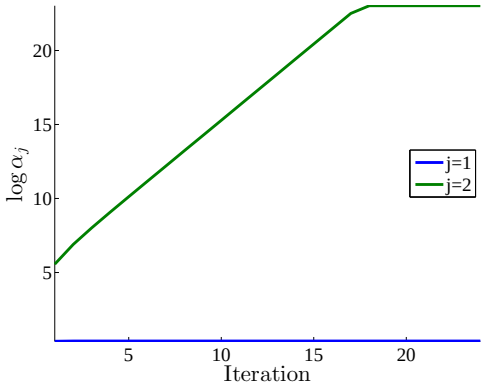


(а) $\varepsilon = 0.$ (б) $\varepsilon = 0.1$

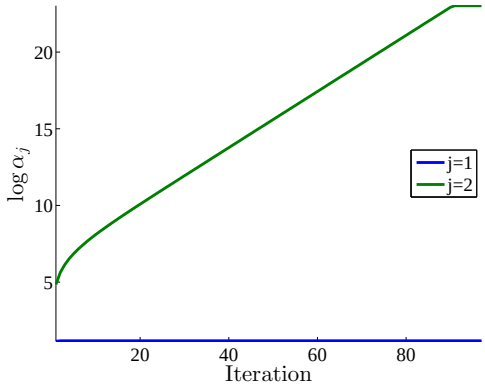


(в) $\varepsilon = 0.3$ (г) $\varepsilon = 0.5$

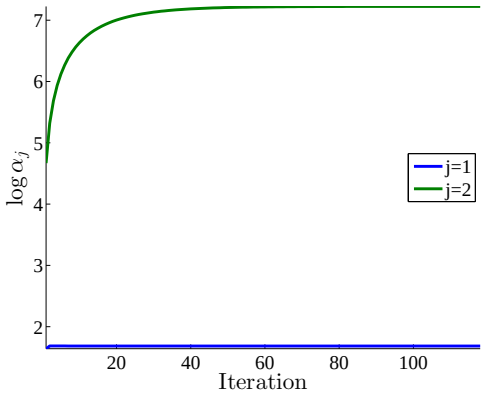
Рис. 9: Зависимость оценок обратных априорных дисперсий α_1, α_2 от номера итерации в зависимости от интенсивности зашумления ε



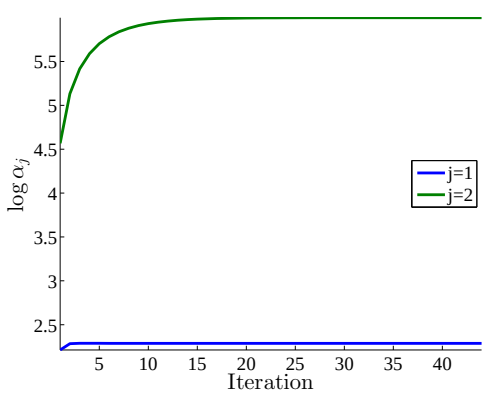
(а) $\varepsilon = 0.6$



(б) $\varepsilon = 0.7$

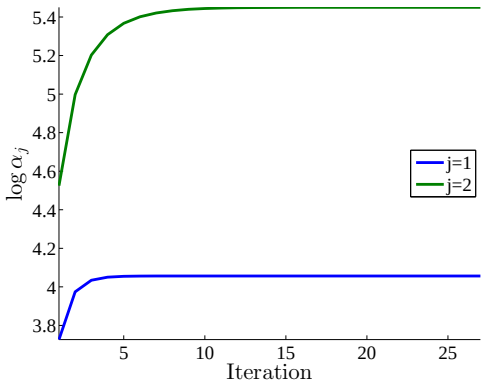


(в) $\varepsilon = 0.75$

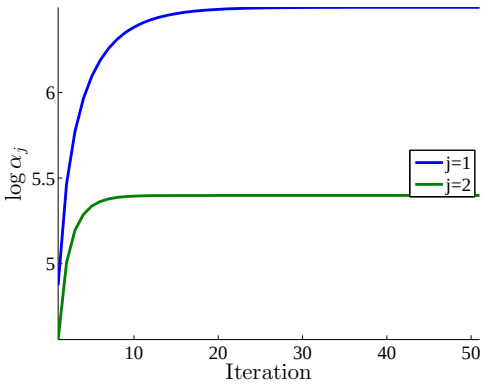


(г) $\varepsilon = 0.8$

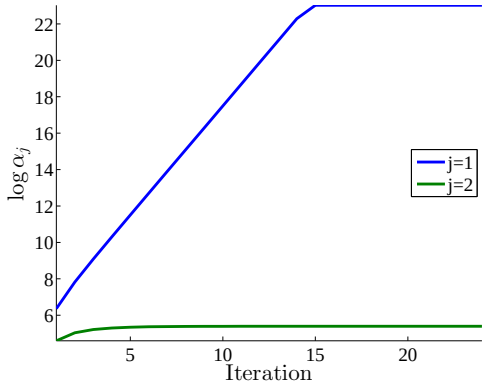
Рис. 10: Зависимость оценок обратных априорных дисперсий α_1, α_2 от номера итерации в зависимости от интенсивности зашумления ε



(а) $\varepsilon = 0.9$



(б) $\varepsilon = 0.95$



(в) $\varepsilon = 0.99$

Рис. 11: Зависимость оценок обратных априорных дисперсий α_1, α_2 от номера итерации в зависимости от интенсивности зашумления ε

Таблица 10: Результаты отбора признаков с помощью максимизации обоснованности модели

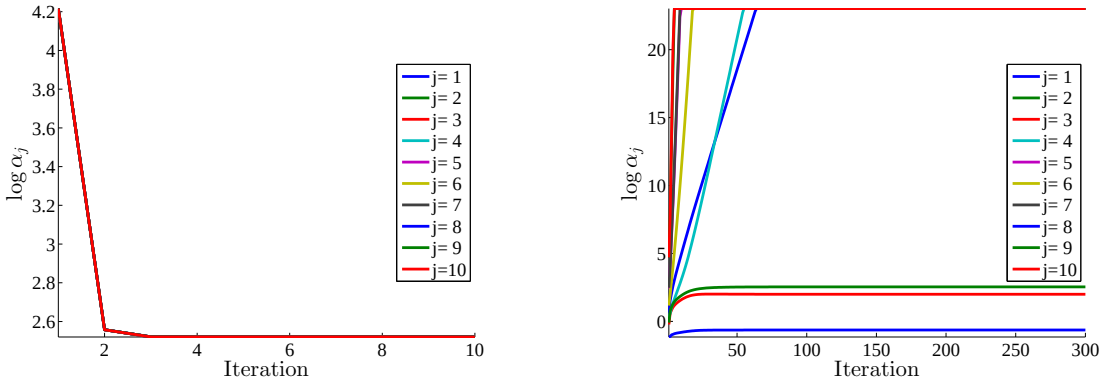
ε	0	0.1	0.3	0.5	0.8	0.9	0.95	0.99
# активн. диаг.	10	3	3	7	7	6	5	4
# активн. полн.	10	10	10	10	10	10	10	9
Q_1^{diag}	0.0157	0.0250	0.0431	0.0614	0.1329	0.1613	0.1677	0.1696
Q_1^{full}	0.0157	0.0453	0.0475	0.0621	0.1301	0.1588	0.1676	0.1706
Q_2^{diag}	0.0176	0.0305	0.0535	0.0782	0.1613	0.1894	0.1954	0.1974
Q_2^{full}	0.0176	0.0548	0.0606	0.0796	0.1593	0.1885	0.1964	0.1999
Тест. Q_1^{diag}	0.0160	0.0271	0.0453	0.0674	0.1434	0.1698	0.1750	0.1754
Тест. Q_1^{full}	0.0160	0.0453	0.0517	0.0696	0.1413	0.1681	0.1748	0.1761
Тест. Q_2^{diag}	0.0177	0.0327	0.0555	0.0835	0.1726	0.1974	0.2019	0.2023
Тест. Q_2^{full}	0.0177	0.0565	0.0642	0.0865	0.1718	0.1985	0.2043	0.2050
$\sum_{j=1}^n w_j^{\text{diag}}$	0.953	0.986	1.198	1.470	1.208	0.449	0.131	-0.0943
$\sum_{j=1}^n w_j^{\text{full}}$	0.953	0.974	1.209	1.517	1.389	0.666	0.225	-0.106
$\sum_{j=1}^n \lvert w_j^{\text{diag}}\rvert$	0.953	1.663	1.198	1.470	1.208	0.523	0.375	0.414
$\sum_{j=1}^n \lvert w_j^{\text{full}}\rvert$	0.953	6.920	2.685	1.955	1.509	1.043	0.810	0.701

Результаты из табл. 10 показывают, что при малом зашумлении ($\varepsilon = 0,\ 0.1$) метод с нахождением ковариационной матрицы в множестве диагональных работает лучше, в том числе на тестовой выборке, чем метод с нахождением ковариационной матрицы общего вида. Однако при увеличении коэффициента зашумления два метода дают сходное качество результатов. Отметим также, что по построению признаков оправданным является их использование с равными весами, поскольку в силу независимости шума это позволит сократить дисперсию шума в результрующей линейной комбинации наилучшим образом. Такая конфигурация весов наблюдается только при малом зашумлении, только при $\varepsilon = 0$ из рассмотренных значений интенсивности зашумления. Кроме того, что метод с оценкой ковариационной матрицы в множестве всех ковариационных матриц соответствующего размера обладает большей абсолютной суммой отрицательных весов, что выражается в большом отношении $(\sum_{j=1}^n |w_j| - \sum_{j=1}^n w_j)/\sum_{j=1}^n |w_j|$. Использование зашумленных признаков с отрицательными весами является нежелательным, поскольку таковое означает попытку компенсации шумов за счет вычитания признаков, однако это невозможно в силу независимости шумов. Отметим также, что при большой зашумленности ($\varepsilon = 0.99$) сумма весов признаков становится отрицательной, то есть при такой зашумленности извлечение зависимости из данных становится невозможным при данном числе объектов в рамках рассматриваемого способа.

На рис. 12 – 13 приведем зависимости оценок обратных априорных дисперсий α_j параметров w_j от номера итерации для всех рассматривавшихся значений интенсивности зашумления ε.

Рис. 12 и 13 показывают, что некоторая доля признаков считаются избыточными, а оценки обратных дисперсий остальных сходятся за малое число итераций.

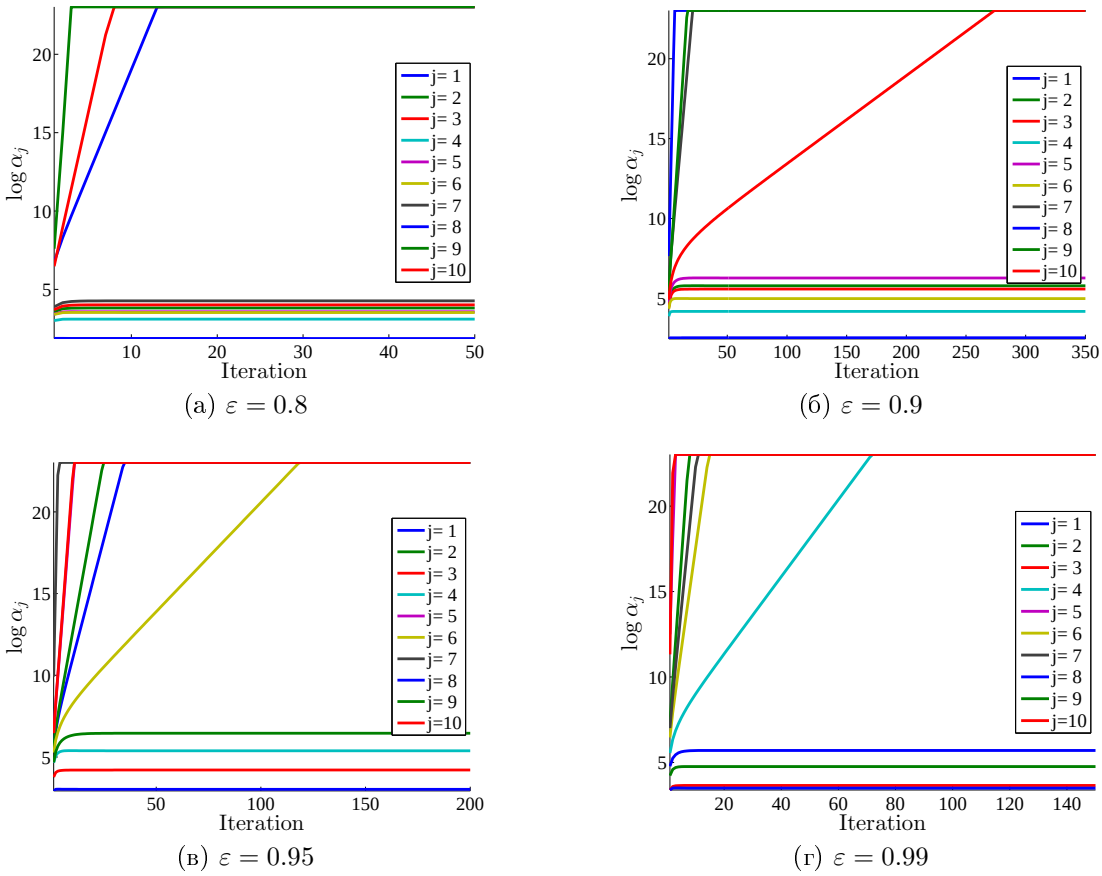
Случай тройки мультиколлинеарных признаков. Рассмотрим теперь случай, когда матрица $\mathbf{X}$ имеет $n = 3$ столбца, причем третий признак равен полусумме пер-



(а) $\varepsilon = 0.$ (б) $\varepsilon = 0.1$

(в) $\varepsilon = 0.3$ (г) $\varepsilon = 0.5$

Рис. 12: Зависимость оценок обратных априорных дисперсий $\alpha_1, \alpha_2, \ldots, \alpha_{10}$ от номера итерации в зависимости от интенсивности зашумления ε



(а) $\varepsilon = 0.8$ (б) $\varepsilon = 0.9$

(в) $\varepsilon = 0.95$ (г) $\varepsilon = 0.99$

Рис. 13: Зависимость оценок обратных априорных дисперсий $\alpha_1, \alpha_2, \ldots, \alpha_{10}$ от номера итерации в зависимости от интенсивности зашумления ε

вых двух, то есть $\mathbf{X}[1,\ 1,\ -2]^\intercal = \mathbf{0}$. Генерацию выборки производим с $\mathbf{w} = [1,\ 1,\ 0]^\intercal$. Интенсивности зашумленности признаков считаем одинаковыми для всех трех признаков, то есть $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon$. Отметим, что вектор весов признаков, минимизирующий дисперсию шума есть $w_1 = w_2 = w_3 = 2/3$. Приведем для разных ε сравнения полученных с помощью каждого из методов результатов (см. табл. 11).

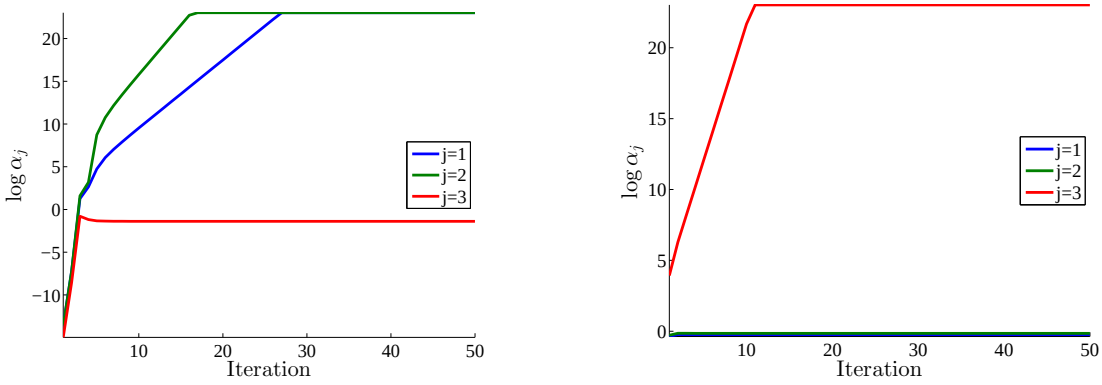
Таблица 11: Результаты отбора признаков с помощью максимизации обоснованности модели

ε	0	0.1	0.3	0.5	0.8	0.9	0.95	0.99
w_1^{diag}	0	1.122	0.975	0.777	0.203	0.009	0	0
w_1^{full}	881.6	1.197	0.968	0.787	0.232	0.084	0.028	-0.008
w_2^{diag}	0	1.073	0.995	0.889	0.417	0.276	0.211	0.165
w_2^{full}	881.6	1.148	0.988	0.896	0.421	0.273	0.210	0.166
w_3^{diag}	1.997	0	0.525	0.607	0.169	0	0	0
w_3^{full}	-1761.2	-0.140	0.567	0.625	0.204	0.070	0.018	-0.015
Q_1^{diag}	$2.5 \cdot 10^{-4}$	0.0216	0.0640	0.1225	0.1988	0.2091	0.2107	0.2117
Q_1^{full}	0.008	0.0230	0.0635	0.1223	0.1982	0.2081	0.2107	0.2117
Q_2^{diag}	$2.8 \cdot 10^{-4}$	0.0276	0.0817	0.1535	0.2353	0.2461	0.2473	0.2479
Q_2^{full}	0.0101	0.0293	0.0812	0.1549	0.2350	0.2448	0.2471	0.2480
Тест. Q_1^{diag}	$2.5 \cdot 10^{-4}$	0.0211	0.0643	0.1243	0.2016	0.2122	0.2126	0.2125
Тест. Q_1^{full}	0.0075	0.0225	0.0643	0.1241	0.2006	0.2101	0.2121	0.2127
Тест. Q_2^{diag}	$2.8 \cdot 10^{-4}$	0.0272	0.0821	0.1535	0.2350	0.2461	0.2464	0.2461
Тест. Q_2^{full}	0.010	0.0290	0.0816	0.1548	0.2345	0.2437	0.2457	0.2464

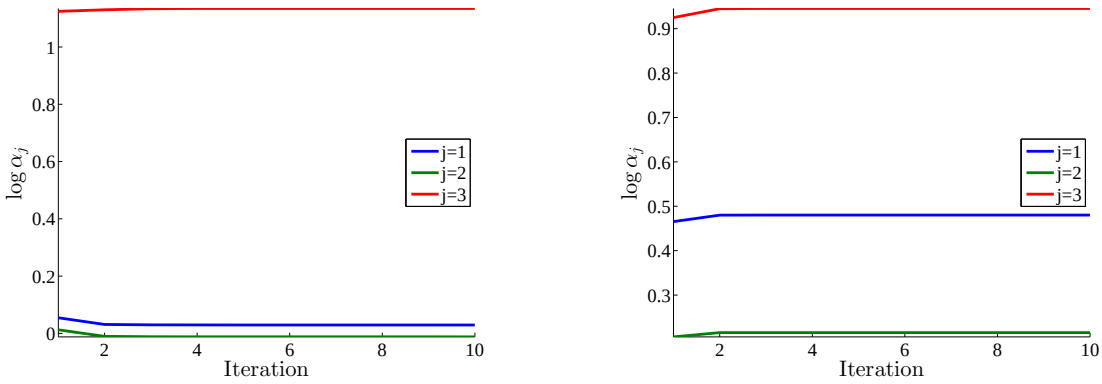
Результаты из табл. 11 показывают, что при отсутствии зашумления ($\varepsilon = 0$) метод с нахождением ковариационной матрицы в множестве диагональных работает лучше, в том числе на тестовой выборке, чем метод с нахождением ковариационной матрицы общего вида. Однако при увеличении коэффициента зашумления два метода дают сходное качество результатов. При отсутствии зашумления метод с оценкой полной ковариационной матрицы дает оценку вектора параметров с большой нормой, поскольку оценка полной матрицы $\mathbf{A}$ дает вырожденный результат. Отметим, однако, что диагональные элементы, соответствующие первым двум признакам ($\alpha_{11} = 7.5 \cdot 10^3$, $\alpha_{22} = 8.6 \cdot 10^3$) значительно больше, чем диагональный элемент, соответствующий третьему признаку ($\alpha_{33} = 12.3$), то есть при отборе с помощью предложенной модификации метода Белсли первые два признака исключаются как избыточные.

На рис. 14 – 15 приведем зависимости оценок обратных априорных дисперсий α_j параметров w_j от номера итерации для всех рассматривавшихся значений интенсивности зашумления ε.

Рис. 14 и 15 показывают, что при большой интенсивности зашумления два из трех признаков считаются избыточными, а оставшийся также имеет большое значение обратной априорной дисперсии, то есть близок к избыточности. Отметим, что даже при интенсивности зашумления, близкой к 1, то есть $\varepsilon = 0.99$, зашумленные признаки содержат некоторую информацию об исходной незашумленной матрице признаков. Рассмотрим теперь случай, когда признаки сэмплируются независимо от

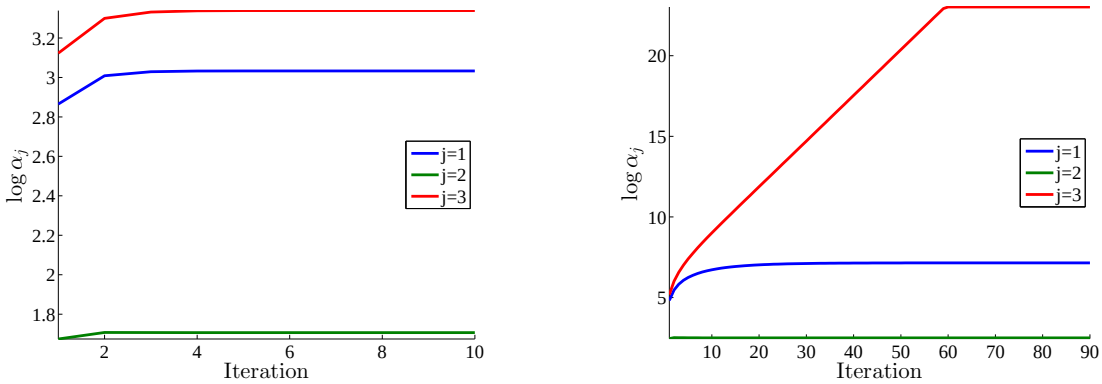


(а) $\varepsilon = 0.$ (б) $\varepsilon = 0.1$



(в) $\varepsilon = 0.3$ (г) $\varepsilon = 0.5$

Рис. 14: Зависимость оценок обратных априорных дисперсий α_1, α_2, α_3 от номера итерации в зависимости от интенсивности зашумления ε



(а) $\varepsilon = 0.8$ (б) $\varepsilon = 0.9$

(в) $\varepsilon = 0.95$ (г) $\varepsilon = 0.99$

Рис. 15: Зависимость оценок обратных априорных дисперсий α_1, α_2, α_3 от номера итерации в зависимости от интенсивности зашумления ε

целевой переменной $\mathbf{y}$.

Случай, когда все признаки являются избыточными. Рассмотрим случай, когда все $n = 10$ признаков являются избыточными, то есть сэмплировались независимо от целевой переменной $\mathbf{y}$, которая определялась в соответствии с (1), но по матрице $\mathbf{X}$, столбцы которой сэмплировались независимо от указанных признаков. Так как число объектов m конечно (в рассматривавшихся выборках $m = 1000$), при увеличении числа признаков растет вероятность того, что между $\mathbf{y}$ и одним из признаков возникнет значимо отличная от нуля корреляция, хотя признаки и сэмплируются независимо. По этой причине в дополнение к уже описанной выборке рассматривается выборка с тем же числом признаков, в которой признаки ортогонализованы по отношению к $\mathbf{y}$.

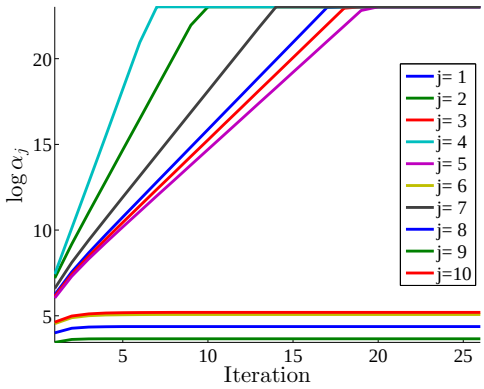
Результатом, который ожидается от алгоритма отбора признаков является признание всех признаков избыточными, что соответствует случаю $\sum_{j=1}^{10} |w_j| = 0$. Результаты отбора признаков с помощью оценки ковариационной матрицы в множестве диагональных ковариационных матриц и в множестве всех ковариационных матриц соответствующего размера приведены в табл. 12. $n_{\text{act}}^{\text{diag}}$ и $n_{\text{act}}^{\text{full}}$ есть количество активных признаков, то есть признаков с $|w_j| > 0.01$, для оценки ковариационной матрицы в рамках множества диагональных и множества всех ковариационных матрицы соответственно.

Таблица 12: Результаты отбора признаков с помощью максимизации обоснованности модели в случае, когда все признаки являются избыточными

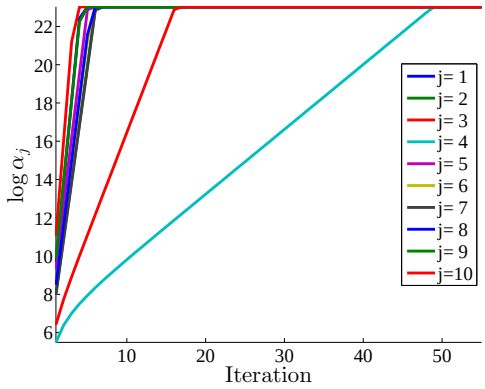
	Неортогонализованные признаки	Ортогонализованные признаки
$\sum_{j=1}^{n} \lvert w_j^{\text{diag}} \rvert$	0.351	0.
$\sum_{j=1}^{n} \lvert w_j^{\text{full}} \rvert$	0.692	$4.7 \cdot 10^{-4}$
$n_{\text{act}}^{\text{diag}}$	4	0
$n_{\text{act}}^{\text{full}}$	10	0

Отметим, что результаты (12) показывают, что если признаки ортогональны целевой переменной, оба метода признают их неинформативными. В случае отсутствия ортогонализации в силу конечности числа объектов и возникновения отличных от нуля корреляций с вектором целевых переменных $\mathbf{y}$, избыточные признаки получают ненулевой вес, хотя и малый. Отметим, что для учета данной проблемы, возможно применения методов, аналогичных методам контроля за значимостью при множественной проверке гипотез [56]. Зависимость оценок априорных обратных дисперсий от номера итерации приведена на рис. 16.

Случай двух пар активных признаков. Рассмотрим случая, когда имеется $n = 4$ признака, которые разбиваются на две пары совпадающих, а признаки из разных пар независимы, $\mathbf{X}[1, -1, 0, 0]^\intercal = \mathbf{0}$, $\mathbf{X}[0, 0, 1, -1]^\intercal = \mathbf{0}$. Генерация данных производится с $\mathbf{w} = [1, 0, 1, 0]^\intercal$. Будем рассматривать случай совпадающих интенсивностей зашумления, то есть $\varepsilon_1 = \varepsilon_2 = \varepsilon_3 = \varepsilon_4 = \varepsilon$. Результаты отбора признаков с помощью рассматриваемых двух методов приведены в табл. 13.



(а) Неортогонализованные признаки.



(б) Ортогонализованные признаки.

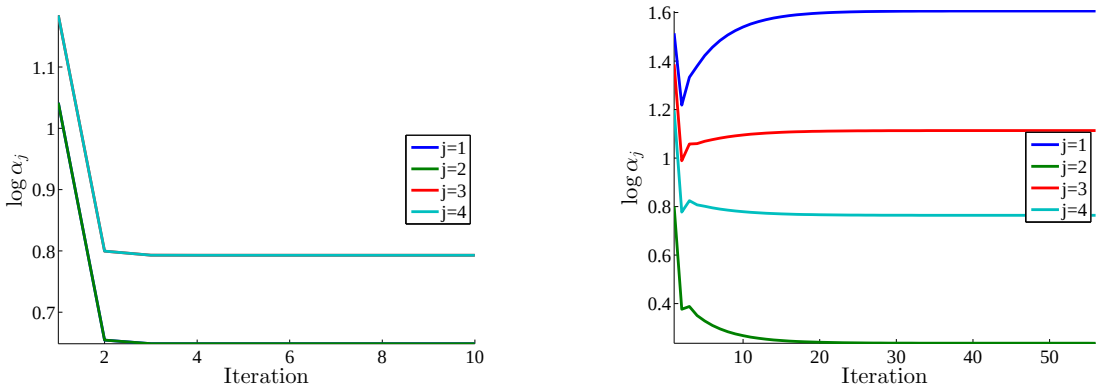
Рис. 16: Зависимость оценок обратных априорных дисперсий $\alpha_1, \alpha_2, \ldots, \alpha_{10}$ от номера итерации для ортогонализованных и неортогонализованных признаков

Таблица 13: Результаты отбора признаков с помощью максимизации обоснованности модели

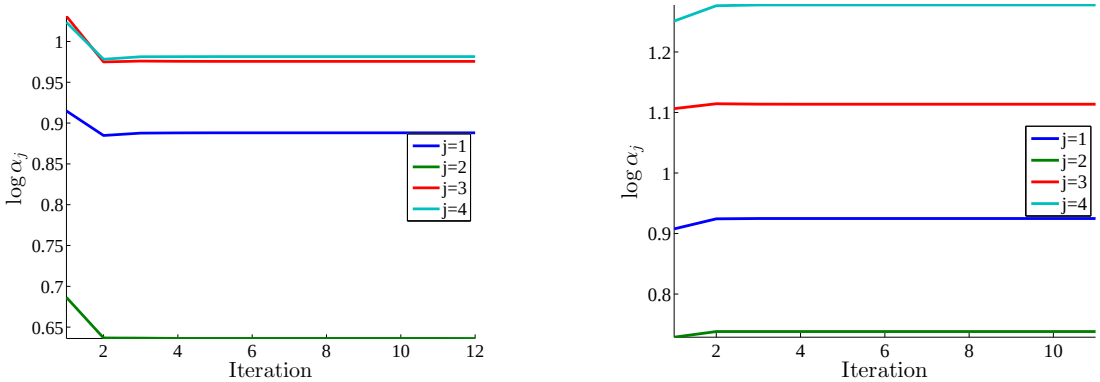
ε	0	0.1	0.3	0.5	0.8	0.9	0.95	0.99
w_1^{diag}	0.509	0.301	0.613	0.617	0.217	0.073	0	0
w_1^{full}	-0.744	0.466	0.629	0.631	0.249	0.109	0.051	0.019
w_2^{diag}	0.509	0.822	0.703	0.680	0.279	0.132	0.061	0
w_2^{full}	1.768	0.667	0.705	0.690	0.296	0.147	0.082	0.043
w_3^{diag}	0.473	0.451	0.582	0.559	0.182	0.027	0	0
w_3^{full}	0.520	0.497	0.594	0.571	0.210	0.084	0.034	0.007
w_4^{diag}	0.473	0.581	0.582	0.512	0.042	0	0	-0.034
w_4^{full}	0.430	0.547	0.596	0.531	0.112	-0.009	-0.044	-0.057
Q_1^{diag}	0.0081	0.0176	0.0560	0.1136	0.1949	0.2057	0.2083	0.2087
Q_1^{full}	0.0077	0.0166	0.0557	0.1131	0.1942	0.2049	0.2073	0.2080
Q_2^{diag}	0.0101	0.0231	0.0717	0.1406	0.2298	0.2410	0.2437	0.2440
Q_2^{full}	0.0098	0.0218	0.0715	0.1405	0.2291	0.2403	0.2427	0.2434
Тест. Q_1^{diag}	0.0083	0.0179	0.0563	0.1134	0.1941	0.2027	0.2035	0.2032
Тест. Q_1^{full}	0.0078	0.0169	0.0559	0.1128	0.1924	0.2025	0.2038	0.2038
Тест. Q_2^{diag}	0.0104	0.0231	0.0712	0.1406	0.2286	0.2374	0.2388	0.2388
Тест. Q_2^{full}	0.0100	0.0219	0.0709	0.1403	0.2273	0.2371	0.2388	0.2390

Отметим, что в силу одинаковой интенсивности зашумления весовой вектор $(0.5, 0.5, 0.5, 0.5)^\intercal$ минимизирует дисперсию шума в полученной линейной комбинации при условии, что $w_1 + w_2 = 1$, $w_3 + w_4 = 1$. При малом зашумлении ($\varepsilon = 0.1$) наблюдается близость оценки максимума апостериорной вероятности к указанным значениям, а при отсутствии шума метод с оценкой ковариационной матрицы в множестве всех ковариационных матриц соответствующего размера имеет $w_1 < 0$, $w_2 > 1$. Отметим, что на случай отсутствия шума указанное замечание не распространяется, поскольку дисперсия шума в линейной комбинации равна нулю при любой комбинации весов. Результаты из табл. 13 также показывают, что при увеличении коэффициента зашумления два метода дают сходное качество результатов.

На рис. 17 – 18 приведем зависимости оценок обратных априорных дисперсий α_j параметров w_j от номера итерации для всех рассматривавшихся значений интенсивности зашумления ε.



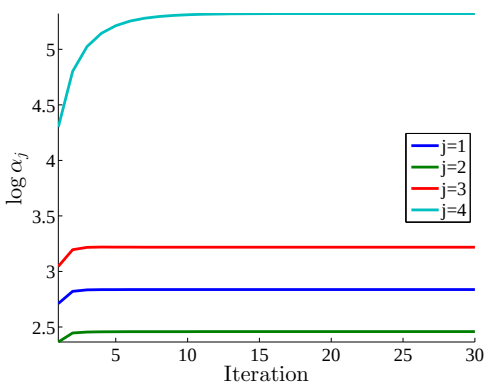
(а) $\varepsilon = 0.$ (б) $\varepsilon = 0.1$



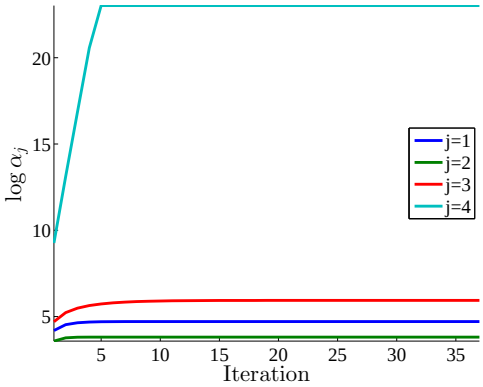
(в) $\varepsilon = 0.3$ (г) $\varepsilon = 0.5$

Рис. 17: Зависимость оценок обратных априорных дисперсий α_1, α_2, α_3, α_4 от номера итерации в зависимости от интенсивности зашумления ε

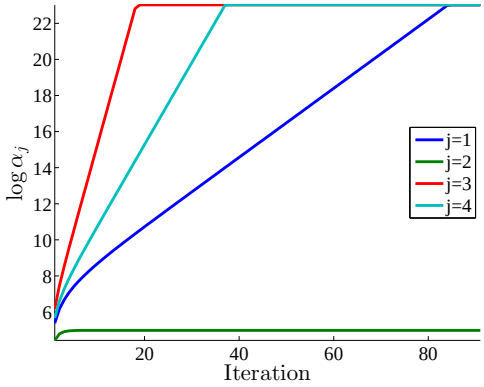
Рис. 17 и 18 показывают, что при большой интенсивности зашумления три из четырех признаков считаются избыточными, а оставшийся также имеет большое значение обратной априорной дисперсии, то есть близок к избыточности.



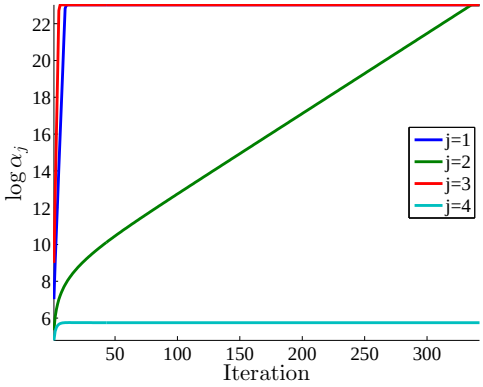
(а) $\varepsilon = 0.8$



(б) $\varepsilon = 0.9$



(в) $\varepsilon = 0.95$



(г) $\varepsilon = 0.99$

Рис. 18: Зависимость оценок обратных априорных дисперсий α_1, α_2, α_3, α_4 от номера итерации в зависимости от интенсивности зашумления ε

Заключение

В работе рассматривается задача классификации на два класса. Предложены решения для трех проблем, возникающих при решении задачи классификации: удаление объектов выбросов из выборки, отбор информативных признаков, учет неоднородности данных. Предложенный алгоритм отбора объектов и фильтрации выбросов основан на предложенной функции специфичности. Показано, что улучшение качества как на исходной, так и на тестовой выборке, связанное с отбором объектов, является статистически значимым. Проведено сравнение предлагаемого метода с другими на четырех реальных выборках данных, а также на синтетических данных, имеющих кластеризованные и некластеризованные выбросы. Предложенный алгоритм работает лучше в терминах функционала качества AUC на данных, имеющих кластеризованные и некластеризованные выбросы, чем другие методы. Предложенный алгоритм отбора признаков, основанный на оценке ковариационной матрицы параметров, был протестирован на синтетических данных, моделирующих разные случаи мультиколлинеарности между признаками. Полученные результаты позволяют сделать вывод о применимости предложенного метода для отбора признаков. Стоит, однако, отметить, что оценка ковариационной матрицы в множестве произвольных ковариационных матриц, требует оценки большого числа параметров по данным, а результат близок к матрице ранга 1. Поэтому в дальнейшем предлагается ввести ограничения на возможные матрицы ковариации, например, путем введения соответствующего априорного распределения на ковариационную матрицу.

Список литературы

[1] *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.

[2] Информация о среднерыночных ставках по разным типам кредитов, данные ЦБ РФ. http://www.cbr.ru/analytics/?PrtId=inf. Дата обращения: 24.05.2015.

[3] Описание российского сервиса равноправного кредитования ФинГуру. http://fingooroo.ru/AboutProject.action. Дата обращения: 24.05.2015.

[4] Устное сообщение В. В. Стрижова. 14.06.2013.

[5] Данные по объему кредитования физических лиц, 2015. http://www.cbr.ru/statistics/UDStat.aspx?Month=04& Year=2015& TblID=302-02M. Дата обращения: 24.05.2015.

[6] Данные по немецким потребительским кредитам. URL: http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/, 2000. Дата обращения: 04.05.2014.

[7] Данные по сердечным заболеваниям в Южной Африке. URL: http://svn.code.sf.net/p/mlalgorithms/code/Aduenko2013BsThesis/data/SAHD.csv. Дата обращения: 04.05.2014.

[8] Данные по качеству вина. URL: http://archive.ics.uci.edu/ml/datasets/Wine/, 1991. Дата обращения: 04.05.2014.

[9] Данные по расположения белка в клетке. URL: http://archive.ics.uci.edu/ml/datasets/Yeast/, 1996. Дата обращения: 04.05.2014.

[10] *Siddiqi N.* Credit risk scorecards: developing and implementing intelligent credit scoring // Wiley, 2006.

[11] *A. Gelman, J. B. Carlin, H. S. Stern and D. B. Rubin* Bayesian Data Analysis. Chapman and Hall, 2003.

[12] *Hahn E. D., and Soyer R.* Probit and logit models: Differences in the multivariate realm. Submitted to The Journal of the Royal Statistical Society, Series B, 2005.

[13] *Hardin J. W., and Hilbe J. M.*. Generalized linear models and extensions. Stata Press, 2007.

[14] *Bishop C. M.* Pattern recognition and machine learning. // Springer, 2006.

[15] *Bishop C. M., Nasrabadi N. M.* Pattern recognition and machine learning. // Journal of electronic imaging, 2007. Vol. 16. No. 4.

[16] Wisnowski, James W., Douglas C. Montgomery, and James R. Simpson. "A comparative analysis of multiple outlier detection procedures in the linear regression model."Computational statistics & data analysis 36.3 (2001): 351-382

[17] Sebert, David M., Douglas C. Montgomery, and Dwayne A. Rollier. "A clustering algorithm for identifying multiple outliers in linear regression."Computational statistics & data analysis 27.4 (1998): 461-484.

[18] Kosinski, Andrzej S. "A procedure for the detection of multivariate outliers."Computational statistics & data analysis 29.2 (1998): 145-161.

[19] Hardin, Johanna, and David M. Rocke. "Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator."Computational Statistics & Data Analysis 44.4 (2004): 625-638.

[20] Filzmoser, Peter, Ricardo Maronna, and Mark Werner. "Outlier identification in high dimensions."Computational Statistics & Data Analysis 52.3 (2008): 1694-1711.

[21] *Motrenko A., Strijov V., Weber G.-W.* Bayesian sample size estimation for logistic regression // Journal of Computational and Applied Mathematics, 2014, 255 — 743-752.

[22] *Croux C., Haesbroeck G.* Implementing the Bianco and Yohai estimator for Logistic Regression // Computational Statistics and Data Analysis, 2003. Vol. 44. Pp. 273–295.

[23] *Hosmer D. W., Lemeshow S.* Applied logistic regression // A Wiley-Interscience Publication, 2000.

[24] *Hastie T., Tibshirani R., Friedman J. H.* The Elements of Statistical Learning // Springer, 2001.

[25] *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.

[26] *Malkovich J. F., Afifi A. A.* On tests for multivariate normality // Journal of the American Statistical Association, 1973. Vol. 68. No. 341. Pp. 176–179.

[27] *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.

[28] *Leardi R., Boggia R., Terrile M.* Genetic algorithms as a strategy for feature selection. // Journal of chemometrics, 1992. Vol. 6. No. 5. Pp. 267–281.

[29] *Huang C. L., Wang C. J.* A GA-based feature selection and parameters optimization for support vector machines // Expert Systems with applications, 2006. Vol. 31. No. 2. Pp. 231–240.

[30] *Weston J. et al.* Feature selection for SVMs // Advances in neural information processing systems, 2001. Pp. 668-674.

[31] *Chapelle O. et al.* Choosing multiple parameters for support vector machines // Machine learning, 2002. Vol. 46. No. 1. Pp. 131–159.

[32] *Neumann J., Schnörr C., Steidl G.* Combined SVM-based feature selection and classification // Machine Learning, 2005. Vol. 61. No. 1. Pp. 129–150.

[33] *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.

[34] *Bissantz N. et al.* Convergence rates of general regularization methods for statistical inverse problems and applications // SIAM Journal on Numerical Analysis, 2007. Vol. 45. No. 6. Pp. 2610-2636.

[35] *Lee S. I. et al.* Efficient l1 regularized logistic regression // Proceedings of the National Conference on Artificial Intelligence, 2006. Vol. 21. No. 1. P. 401.

[36] *Нестеров Ю. Е.* Методы выпуклой оптимизации. М.: МЦНМО, 2010.

[37] *Liu Y., Wu Y.* Variable selection via a combination of the L0 and L1 penalties // Journal of Computational and Graphical Statistics, 2007. Vol. 16. No. 4.

[38] *Zare A., Gader P.* Sparsity promoting iterated constrained endmember detection in hyperspectral imagery // IEEE Geoscience and Remote Sensing Letters, 2007. Vol. 4. No. 3. P. 446.

[39] *Krishnapuram B. et al.* A Bayesian approach to joint feature selection and classifier design // Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. Vol. 26. No. 9. Pp. 1105-1111.

[40] *Krishnapuram B. et al.* Sparse multinomial logistic regression: Fast algorithms and generalization bounds // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005. Vol. 27. No. 6. Pp. 957–968.

[41] *Lee Y., Nelder J. A., Pawitan Y.* Generalized linear models with random effects: unified analysis via H-likelihood // Chapman&Hall/CRC, 2006. Vol. 106.

[42] *Леонтьева Л. Н.* Последовательный выбор признаков при восстановлении регрессии // Машинное обучение и анализ данных, 2012. Т. 1. № 3. С. 335–346.

[43] *Kohavi R. et al.* A study of cross-validation and bootstrap for accuracy estimation and model selection // IJCAI, 1995. Vol. 14. No. 2. Pp. 1137-1145.

[44] *MacKay D. J. C.* Bayesian methods for adaptive models // California Institute of Technology, 1992.

[45] *Kwok J. T. Y.* The evidence framework applied to support vector machines // Neural Networks, IEEE Transactions on, 2000. Vol. 11. No. 5. Pp. 1162-1173.

[46] *Mackay D. J, C.* The evidence framework applied to classification networks // Neural computation, 1992. Vol. 4. No. 5. Pp. 720-736.

[47] *Адуенко А. А.* Совместный выбор объектов и признаков при построении моделей в задачах банковского скоринга, 2013. URL: http://www.machinelearning.ru/wiki/images/b/b3/Thesis.pdf. Дата обращения: 24.05.2015.

[48] *Elfelly N. et al.* Multimodel control design using unsupervised classifiers // Studies in Informatics and Control, 2012. Vol. 21. No. 1. P. 102.

[49] *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.

[50] *Gelman A., Hill J.* Data analysis using regression and multilevel/hierarchical models // Cambridge University Press, 2006.

[51] *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.

[52] *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.

[53] *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.

[54] *Albert J., Chib S.* Bayesian residual analysis for binary response regression models // Biometrika, 1995. Vol. 82. No. 4. Pp. 747–769.

[55] *Katrutsa A. M., Strijov V. V.* Stresstest procedures for feature selection algorithms. // Journal of Chemometrics, 2015.

[56] *Benjamini Y., Hochberg Y.* Controlling the false discovery rate: a practical and powerful approach to multiple testing // Journal of the Royal Statistical Society. Series B (Methodological), 1995. Pp. 289-300.