Глава 6

Байесовский подход к теории вероятностей. Примеры байесовских рассуждений

В главе представлен байесовский подход к теории вероятностей, при котором вероятность интерпретируется как мера незнания, а не как объективная случайность. Приведены основные правила работы с условными вероятностями. Демонстрируются различия между частотным и байесовским подходами. Показано, что байесовский подход к теории вероятностей можно рассматривать как обобщение классической булевой логики для проведения логических рассуждений в условиях неопределенностей. В конце главы приведен пример байесовского вывода для ситуации, в которой классическая логика оказывается бессильна.

6.1 Ликбез: Формула Байеса

6.1.1 Sum- и Product- rule

Условная вероятность

- Пусть X и Y — случайные величины с плотностями $p(x)$ и $p(y)$ соответственно
- В общем случае их совместная плотность $p(x, y) \neq p(x)p(y)$. Если это равенство выполняется, величины называют **независимыми**
- Условной плотностью называется величина
$$p(x|y) = \frac{p(x, y)}{p(y)}$$
- Смысл: как факт $Y = y$ влияет на распределение X.
Заметим, что $\int p(x|y)dx \equiv 1$, но $\int p(x|y)dy$ не обязан равняться единице, т.к. относительно y это не плотность, а **функция правдоподобия**
- Очевидная система тождеств $p(x|y)p(y) = p(x, y) = p(y|x)p(x)$ позволяет легко переходить от $p(x|y)$ к $p(y|x)$
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Sum-rule

- Все операции над вероятностями базируются на применении всего двух правил
- Sum rule: Пусть $A_1, \ldots, A_k$ взаимоисключающие события, одно из которых **всегда происходит**. Тогда
$$P(A_i \cup A_j) = P(A_i) + P(A_j) \qquad \sum_{i=1}^{k} P(A_i) = 1$$
- Очевидное следствие (формула полной вероятности): $\forall B$ верно $\sum_{i=1}^{k} P(A_i|B) = 1$, откуда
$$\sum_{i=1}^{k} \frac{P(B|A_i)P(A_i)}{P(B)} = 1 \quad P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$
- В интегральной форме
$$p(b) = \int p(b, a)da = \int p(b|a)p(a)da$$

Product-rule

- Правило произведения (product rule) гласит, что любую совместную плотность всегда можно разбить на множители
$$p(a, b) = p(a|b)p(b) \quad P(A, B) = P(A|B)P(B)$$
- Аналогично для многомерных совместных распределений
$$p(a_1, \ldots, a_n) =$$
$$p(a_1|a_2, \ldots, a_n)p(a_2|a_3, \ldots, a_n) \ldots p(a_{n-1}|a_n)p(a_n)$$
- Можно показать (Jaynes, 1995), что Sum- и Product- rule являются единственными возможными операциями, позволяющими рассматривать вероятности как промежуточную ступень между истиной и ложью

6.1.2 Формула Байеса

Априорные и апостериорные суждения

- Предположим, мы пытаемся изучить некоторое явление
- У нас имеются некоторые знания, полученные до (лат. a priori) наблюдений/эксперимента. Это может быть опыт прошлых наблюдений, какие-то модельные гипотезы, ожидания
- В процессе наблюдений эти знания подвергаются постепенному уточнению. После (лат. a posteriori) наблюдений/эксперимента у нас формируются новые знания о явлении
- Будем считать, что мы пытаемся оценить неизвестное значение величины θ посредством наблюдений некоторых ее косвенных характеристик $x|\theta$

Формула Байеса

- Знаменитая формула Байеса (1763 г.) устанавливает правила, по которым происходит преобразование знаний в процессе наблюдений
- Обозначим априорные знания о величине θ за $p(\theta)$
- В процессе наблюдений мы получаем серию значений $\boldsymbol{x} = (x_1, \ldots, x_n)$. При разных θ наблюдение выборки $\boldsymbol{x}$ более или менее вероятно и определяется значением правдоподобия $p(\boldsymbol{x}|\theta)$
- За счет наблюдений наши представления о значении θ меняются согласно формуле Байеса

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int p(\boldsymbol{x}|\theta)p(\theta)d\theta}$$

- Заметим, что знаменатель не зависит от θ и нужен исключительно для нормировки апостериорной плотности

6.2 Два подхода к теории вероятностей

6.2.1 Частотный подход

Различия в подходах к теории вероятностей

- В современной теории вероятностей существуют два подхода к тому, что называть случайностью
- В частотном подходе предполагается, что случайность есть **объективная неопределенность**
  В жизни «объективные» неопределенности практически не встречаются. Чуть ли не единственным примером может служить радиоактивный распад (во всяком случае, по современным представлениям)
- В байесовском подходе предполагается, что случайность есть **мера нашего незнания**
  Почти любой случайный процесс можно так интерпретировать. Например, случайность при бросании кости связана с незнанием динамических характеристик кубика, сукна, руки кидающего, сопротивления воздуха и т.п.

Следствие частотного подхода

- При интерпретации случайности как «объективной» неопределенности **единственным** возможным средством анализа является проведение серии испытаний
- При этом вероятность события интерпретируется как предел частоты наступления этого события в n испытаниях при $n \to \infty$
- Исторически частотный подход возник из весьма важной практической задачи: анализа азартных игр — области, в которой понятие серии испытаний имеет простой и ясный смысл

Особенности частотного подхода

- Величины четко делятся на случайные и детерминированные
- Теоретические результаты работают на практике при больших выборках, т.е. при $n \gg 1$
- В качестве оценок неизвестных параметров выступают точечные, реже интервальные оценки
- Основным методом статистического оценивания является метод максимального правдоподобия (Фишер, 1930ые гг.)

6.2.2 Байесовский подход

Альтернативный подход

- Далеко не всегда при оценке вероятности события удается провести серию испытаний.
- Пример: оцените вероятность того, что человеческая цивилизация может быть уничтожена метеоритной атакой
- Очевидно, что частотным методом задачу решить невозможно (точнее вероятность этого события строго равна нулю, ведь подобного еще не встречалось). В то же время интерпретация вероятности как меры нашего незнания позволяет получить отличный от нуля осмысленный ответ
- Идея байесовского подхода заключается в переходе от априорных знаний (или точнее незнаний) к апостериорным с учетом наблюдаемых явлений

Особенности байесовского подхода

- Все величины и параметры считаются случайными
  Точное значение параметров распределения нам неизвестно, значит они случайны с точки зрения нашего незнания
- Байесовские методы работают даже при объеме выборки 0! В этом случае апостериорное распределение равно априорному
- В качестве оценок неизвестных параметров выступают апостериорные распределения, т.е. решить задачу оценивания некоторой величины, значит найти ее апостериорное распределение
- Основным инструментом является формула Байеса, а также sum- и product- rule

Недостатки байесовского подхода

- Начиная с 1930 гг. байесовские методы подвергались резкой критике и практически не использовались по следующим причинам
  - В байесовских методах предполагается, что априорное распределение известно до начала наблюдений и не предлагается конструктивных способов его выбора
  - Принятие решения при использовании байесовских методов в нетривиальных случаях требует колоссальных вычислительных затрат, связанных с численным интегрированием в многомерных пространствах
  - Фишером была показана оптимальность метода максимального правдоподобия, а следовательно — бессмысленность попыток придумать что-то лучшее
- В настоящее время (с начала 1990 гг.) наблюдается возрождение байесовских методов, которые оказались в состоянии решить многие серьезные проблемы статистики и машинного обучения

Точечные оценки при использовании метода Байеса

- Математическое ожидание по апостериорному распределению. Весьма трудоемкая процедура

$$\hat{\theta}_B = \int \theta p(\theta|\boldsymbol{x}) d\theta$$

- Максимум апостериорной плотности. Удобен в вычислительном плане

$$\hat{\theta}_{MP} = \arg\max P(\theta|\boldsymbol{x}) = \arg\max P(\boldsymbol{x}|\theta)P(\theta) = \arg\max \left(\log P(\boldsymbol{x}|\theta) + \log P(\theta)\right)$$

- Это фактически регуляризация метода максимального правдоподобия!

6.3 Байесовские рассуждения

6.3.1 Связь между байесовским подходом и булевой логикой

Попытки обобщения булевой логики

- Классическая булева логика плохо применима к жизненным ситуациям, которые далеко не всегда выразимы в терминах «истина» и «ложь»
- Неоднократно предпринимались попытки обобщить булеву логику, сохраняя при этом действие основных логических законов (Modus Ponens, Modus Tolens, правило де Моргана, закон двойного отрицания и пр.)
- Наиболее известные примеры:
  - Многозначная логика, расширившая множество логических переменных до $\{0, 1, \ldots, k-1\}$
  - Нечеткая логика, оперирующая континуумом значений между 0 и 1, характеризующими разную степень истинности

Недостатки нечеткой логики

- Несмотря на кажущуюся привлекательность нечеткая логика обладает рядом существенных недостатков
- Отсутствует строгое математическое обоснование ряду методов, использующихся в нечетких рассуждениях
- Существует множество эвристических правил, определяющих как именно нужно строить нечеткий вывод. Все они приводят к различным результатам
- Непонятна связь нечеткой логики с теорией вероятности

Логическая интерпретация байесовского подхода

- Байесовский вывод можно рассматривать как обобщение классической булевой логики. Только вместо понятий «истина» и «ложь» вводится «истина с вероятностью p».
- Обобщение классического правила Modus Ponens

$$\frac{A, A \Rightarrow B}{A\&B}$$

$$\frac{p(A), p(B|A)}{p(A\&B)}$$

- Теперь рассмотрим такую ситуацию

$$\frac{A \Rightarrow B, B}{A = ?}$$

$$\frac{p(B|A), p(B), p(A)}{p(A|B)}$$

Формула Байеса позволяет рассчитать изменение степени истинности A с учетом информации о B

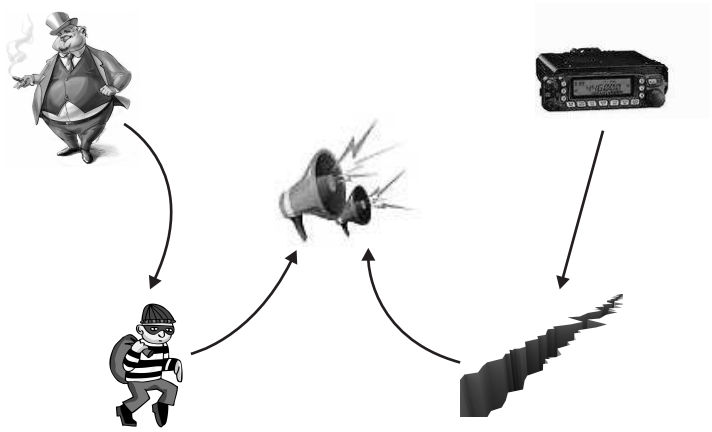
- Это новый подход к синтезу экспертных систем
- В отличие от нечеткой логики, он теоретически обоснован и математически корректен

6.3.2 Пример вероятностных рассуждений

Жизненная ситуация

Предположим, что Джон установил у себя дома сигнализацию от воров. Если к нему в дом проникает вор (событие v), Джон получает СМС на свой мобильный (событие t). Сигнализация также может срабатывать от небольших землятрясений (событие z), которые иногда происходят в городе Джона. Пусть в один из дней в обед Джон получает сигнал тревоги. За обедом он встречает своего друга (событие d), который сообщает ему, что уровень преступности в квартале Джона в 10 раз выше среднего по говоду. Закончив обедать, Джон слышит сводку новостей по радио (событие r), в которой сообщается о только что произошедшем землятрясении.

Символом $\neg$ будем обозначать событие, противоположное к исходному



Вероятностная интерпретация

- Технические характеристики сигнализации $p(t|v,z) = p(t|v,\neg z) = 1$, $p(t|\neg v,z) = 0.1$, $p(t|\neg v,\neg z) = 0$
- Статистическая информация, набранная Джоном $p(v) = 2 \cdot 10^{-4}$, $p(z) = 0.01$
- Сообщение друга $p(d) = 1$, $p(v|d) = 2 \cdot 10^{-3}$
- Мы предположим, что Джон полностью доверяет другу. Но мы легко могли бы учесть и тот факт, что друг Джона – большой шутник и мог его разыграть
- Сводка новостей по радио $p(r) = 1$, $p(r|z) = 0.5$, $p(r|\neg z) = 0$

Вероятность взлома и ложной тревоги при получении сигнала тревоги

Срабатывание сигнализации $p(t) = 1$

$$p(v|t) = \frac{1}{Z} p(t|v)p(v)$$

$$p(\neg v|t) = \frac{1}{Z} p(t|\neg v)p(\neg v)$$

$$Z = p(t|v)p(v) + p(t|\neg v)p(\neg v)$$

$$p(t|v) = p(t|v, \neg z)p(\neg z) + p(t|v, z)p(z) = p(\neg z) + p(z) = 1$$

$$p(t|\neg v) = p(t|\neg v, \neg z)p(\neg z) + p(t|\neg v, z)p(z) = p(t|\neg v, z)p(z) = 10^{-3}$$

$$Z = 1.2 \cdot 10^{-3}$$

$$p(v|t) = \frac{1}{6} \approx 16.7\%$$

$$p(\neg v|t) = \frac{5}{6} \approx 83.3\%$$

Вероятность взлома и ложной тревоги при получении сигнала тревоги и после разговора с другом

Сообщение друга $p(d) = 1$

$$p(v|t, d) = \{Cond.\ ind.\} = \frac{1}{Z} \frac{p(v|t)p(v|d)}{p(v)} = \frac{1}{Z} \frac{10}{6}$$

$$p(\neg v|t, d) = \{Cond.\ ind.\} = \frac{1}{Z} \frac{p(\neg v|t)p(\neg v|d)}{p(\neg v)} \approx \frac{1}{Z} \frac{5}{6}$$

$$Z = \frac{p(v|t)p(v|d)}{p(v)} + \frac{p(\neg v|t)p(\neg v|d)}{p(\neg v)}$$

$$Z = \frac{15}{6}$$

$$p(v|t, d) = \frac{10}{15} \approx 66.7\%$$

$$p(\neg v|t, d) = \frac{5}{15} \approx 33.3\%$$

Комментарием $\{Cond.\ ind.\}$ обозначена т.н. условная независимость событий d и t относительно v. Подробнее см. раздел 12.1.2

Вероятность взлома и ложной тревоги при получении сигнала тревоги, после разговора с другом и радиосообщения

Радиосводка $p(r) = 1$, т.к. $p(r|\neg z) = 0$, то $p(z|r) = 1$, по условию

$$p(v|t,d,r) = \frac{1}{Z}p(t|v,r,d)p(v,r,d) = \frac{1}{Z}p(v,r,d) = \{Indep.\ assump.\} = \frac{1}{Z}p(v,d)p(r) = \frac{1}{Z}p(v|d)p(d)p(r) = \frac{1}{Z}2\cdot 10^{-3}\times 1\times 1$$

$$p(\neg v|t,d,r) = \{p(t|\neg v,d,r) = p(t|\neg v,d,z)p(z|r) + p(t|\neg v,d,\neg z)p(\neg z|r)\} = \frac{1}{Z}p(t|\neg v,r,d)p(\neg v,r,d) = \frac{1}{Z}0.1\times p(\neg v,r,d) = \{Indep.\ assump.\} = \frac{1}{Z}0.1\times p(\neg v,d)p(r) = \frac{1}{Z}0.1\times p(\neg v|d)p(d)p(r) = \frac{1}{Z}0.1\times(1-2\cdot 10^{-3})\times 1\times 1$$

$$p(v|t,d,z) = \frac{20}{1018} \approx 1.9\%$$
$$p(\neg v|t,d,z) = \frac{998}{1018} \approx 98.1\%$$

Ошибка Джона

- Успокоенный Джон возвращается на работу, а вечером, придя домой, обнаруживает, что квартира «обчищена».
- Джон отлично владел байесовским аппаратом теории вероятностей, но значительно хуже разбирался в человеческой психологии
- Предположение о независимости кражи и землетрясения оказалось неверным

$$p(v,z) \neq p(v)p(z)$$

- Действительно, когда происходит землетрясение, воры проявляют значительно большую активность, достойную лучшего применения

$$p(v|z) > p(v|\neg z)$$

Глава 7

Решение задачи выбора модели по Байесу. Обоснованность модели

В главе описывается общая схема байесовского вывода. Внимание уделяется вопросам практического применения байесовского вывода с помощью сопряженных распределений. Подробно рассматривается двухуровневая схема байесовского вывода и лежащий в ее основе принцип наибольшей обоснованности. В конце главы приведен пример использования обоснованности для выбора модели, показаны отличия между оцениванием по методу максимального правдоподобия и оцениванием по максимуму апостериорной вероятности.

7.1 Ликбез: Бритва Оккама и Ad Hoc гипотезы

Бритва Оккама

- В 14 в. английский монах В.Оккам ввел принцип, ставший методологической основой современной науки
- Entia non sunt multiplicanda sine necessitate (лат. сущности не следует умножать без необходимости)
- Согласно этому принципу из всех гипотез, объясняющих некоторое явление, следует предпочесть наиболее простую

  Наполеон однажды спросил Лапласа (полушутя, полусерьёзно): «Что-то я не вижу в Вашей теории места для Бога». На что Лаплас, якобы, ответил: «Сир, у меня не было нужды в этой гипотезе».

Ad Hoc гипотезы

- Если гипотеза выдвигается специально для объяснения одного конкретного явления, ее называют ad hoc гипотезой
- В научных исследованиях ad hoc гипотезой называют поправки, вводимые в теорию, чтобы она смогла объяснить очередной эксперимент, который не укладывается в рамки теории
- Согласно принципу Оккама, ad hoc гипотезы не являются научными и не должны использоваться
- Классификатор, который в состоянии объяснить (правильно классифицировать) **только те прецеденты, которые ему предъявлялись с правильными ответами в ходе обучения** (обучающую выборку), является примером ad hoc гипотезы

7.2 Полный байесовский вывод

7.2.1 Пример использования априорных знаний

Предположим, нам необходимо оценить количество ящиков, находящихся за деревом, показанном на рисунке 7.1. С точки зрения метода максимума правдоподобия любое положительное число ящиков одинаково приемлемо (рис. 7.2). Наша же интуиция (а точнее, априорные знания о характерной ширине ящика, базирующиеся на наблюдениях ящиков справа и слева от дерева) подсказывает иной ответ (рис. 7.3)

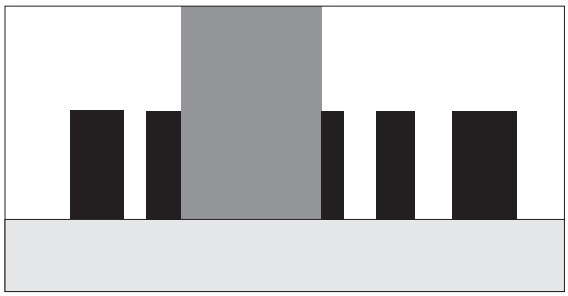


Рис. 7.1. Сколько ящиков за деревом?

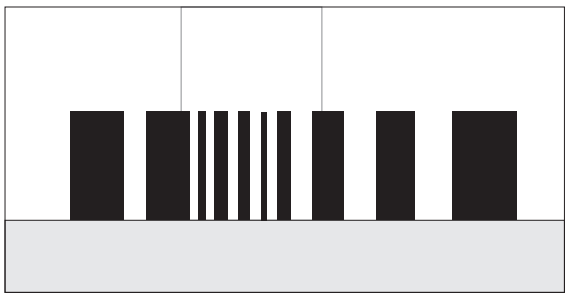


Рис. 7.2.

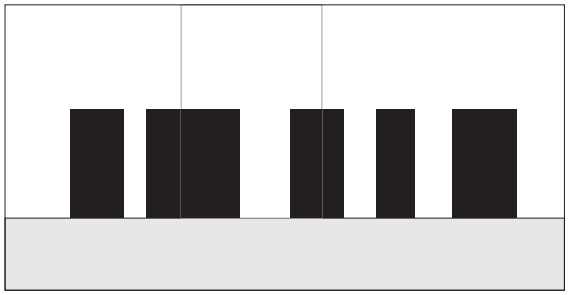


Рис. 7.3.

7.2.2 Сопряженные распределения

Получение апостериорного распределения

- Рассмотрим задачу получения апостериорного распределения на неизвестный параметр θ
- Согласно формуле Байеса

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int p(\boldsymbol{x}|\theta)p(\theta)d\theta}$$

- Таким образом, для подсчета апостериорного распределения необходимо знать значение знаменателя в формуле Байеса
- В случае, если θ является векторнозначной переменной, возникает необходимость (как правило численного) интегрирования в многомерном пространстве

Аналитическое интегрирование

- При размерности выше 5-10 численное интегрирование с требуемой точностью невозможно
- Возникает вопрос: в каких случаях можно провести интегрирование аналитически?
- Распределения $p(\theta) \sim \mathcal{A}(\alpha_0)$ и $p(\boldsymbol{x}|\theta) \sim \mathcal{B}(\beta)$ являются сопряженными, если

$$p(\theta|\boldsymbol{x}) \sim \mathcal{A}(\alpha_1)$$

- Если априорное распределение выбрано из класса распределений, сопряженных правдоподобию, то апостериорное распределение выписывается **в явном виде**

Пример

- Подбрасывание монетки n раз с вероятностью выпадания орла $q \in (0, 1)$
- Число выпавших орлов m, очевидно, имеет распределение Бернулли

$$p(m|n, q) = C_n^m q^m (1 - q)^{n-m} \sim \mathcal{B}(m|n, q)$$

- Сопряженным к распределению Бернулли является бета-распределение

$$p(q|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} q^{a-1}(1 - q)^{b-1} \sim \text{Beta}(q|a, b)$$

- Легко показать, что интеграл от произведения распределения Бернулли и бета-распределения берется аналитически

✓ Упр.

- Бета-распределение часто используется когда нужно указать распределение на вероятность какого-то события

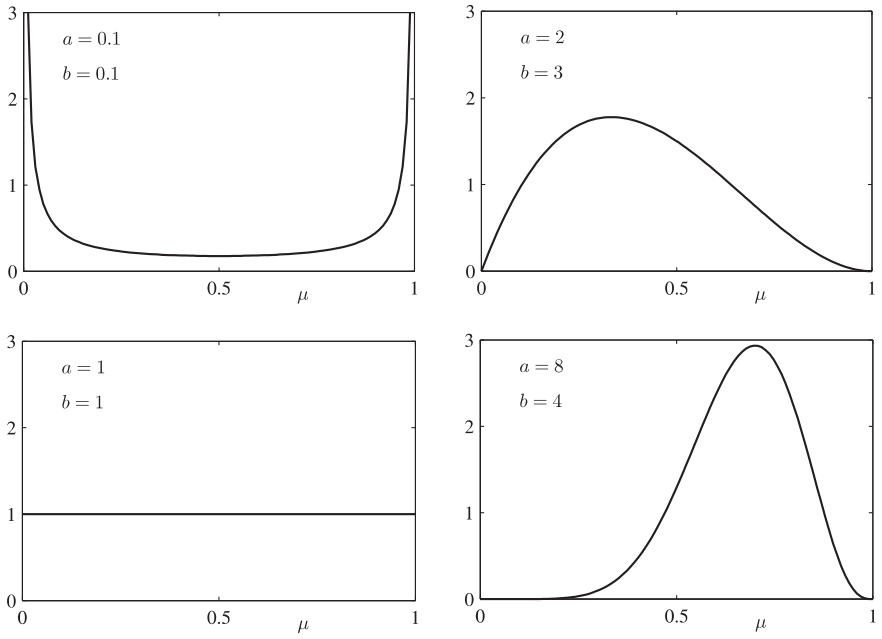


Рис. 7.4. Различные формы бета-распределения

- Применяя формулу Байеса, получаем

$$p(q|\text{«}m \text{ орлов»}) \sim \text{Beta}(q|a + m, b + n - m)$$

- Отсюда простая интерпретация параметров a и b как эффективного количества наблюдений орлов и решек
- Можно считать априорное распределение нашими прошлыми наблюдениями
- Возьмем в качестве априорного распределения равномерное (т.е. бета-распределение с параметрами $a = b = 1$). Это означает, что у нас нет никаких предпочтений относительно кривизны монеты

- В этом случае взятие мат. ожидания по апостериорному распредению на q приводит к характерной регуляризованной точечной оценке на вероятность выпадения орла

$$\hat{q}_B = \int_0^1 p(q|\text{«}m \text{ орлов»})qdq = \frac{m+1}{n+2}$$

Примеры сопряженных распределений

- Для большинства известных распределений существуют сопряженные, хотя не всегда они выписываются в простом виде
- В частности, в явном виде можно выписать сопряженные распределения для любого распределения из экспоненциального семейства, т.е. распределения вида

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = h(\boldsymbol{x})g(\boldsymbol{\alpha})\exp(\boldsymbol{\alpha}^T u(\boldsymbol{x}))$$

- К этому семейству относятся нормальное, гамма-, бета-, равномерное, Бернулли, Дирихле, Хи-квадрат, Пуассоновское и многие другие распределения
- Вывод: если правдоподобие представляет собой некоторое распределение, для которого существует сопряженное, именно его и нужно стараться взять в качестве априорного распределения. Тогда ответ (апостериорное распределение) будет выписан в явном виде

7.2.3 Иерархическая схема Байеса

Выбор априорного распределения

- Априорное распределение также может быть задано в параметрической форме $p(\theta) = p(\theta|\alpha)$
- Для того, чтобы применить формулу Байеса, необходимо сначала определить значение α
- Для оценки α можно вновь воспользоваться формулой Байеса, введя на α априорное распределение $p(\alpha)$. Тогда

$$p(\alpha|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\alpha)p(\alpha)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\alpha)p(\alpha)}{\int p(\boldsymbol{x}|\alpha)p(\alpha)d\alpha}$$

- В качестве правдоподобия относительно α выступает т.н. обоснованность $p(\boldsymbol{x}|\alpha) = \int p(\boldsymbol{x}|\theta)p(\theta|\alpha)d\theta$, полученная путем исключения (integrate out) переменной θ

Иерархическая схема байесовского вывода

- Само априорное распределение на α также может быть задано с точностью до параметра: $p(\alpha) = p(\alpha|\beta)$
- Для определения значения β можно вновь воспользоваться схемой Байеса и т.д.
- На каком-то этапе придется воспользоваться «заглушкой» в виде оценки максимума правдоподобия (рис. 7.5)

7.3 Принцип наибольшей обоснованности

7.3.1 Обоснованность модели

Обоснованность модели

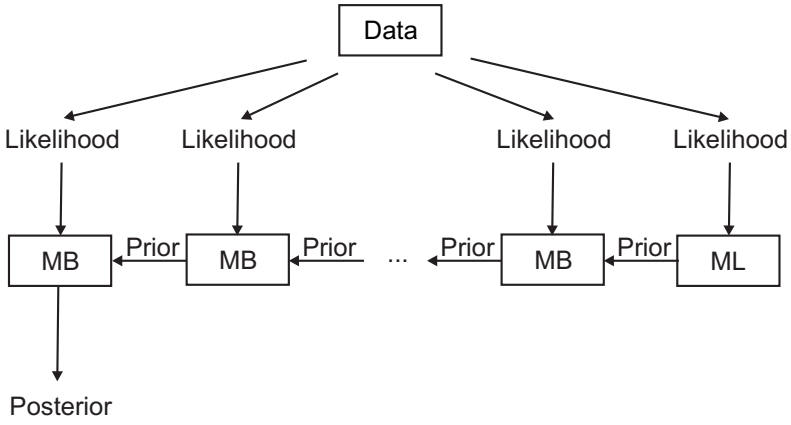


Рис. 7.5. Иерархическая схема байесовского вывода

- На практике обычно ограничиваются двумя уровнями вывода, применяя метод максимального правдоподобия для оценки гиперпараметров
- **Гиперпараметры носят более абстрактный характер, поэтому их настройка по данным не приводит к переобучению** (см. рис. 7.6)
- Функция правдоподобия гиперпараметров называется обоснованностью (evidence) модели

$$p(\boldsymbol{x}|\alpha) = \int_{\Theta} p(\boldsymbol{x}|\theta)p(\theta|\alpha)d\theta$$

- Гиперпараметры подбираются путем максимизации обоснованности

$$\alpha_{ME} = \arg\max p(\boldsymbol{x}|\alpha)$$

- Далее можно решать стандартную задачу на поиск максимума регуляризованного правдоподобия

$$\theta_{MP} = \arg\max p(\boldsymbol{x}|\theta)p(\theta|\alpha_{ME})$$

Принцип наибольшей обоснованности с точки зрения байесовского подхода

- Применение метода максимального правдоподобия на втором уровне байесовского вывода означает, что все модели для нас одинаково приемлемы, т.е. $p(\alpha) = Const$
- В этом случае легко показать, что $p(\alpha|\boldsymbol{x}) \propto p(\boldsymbol{x}|\alpha)$
- Вообще-то, это не совсем байесовский вывод...

Формальный вывод

Если действовать формально, то необходимо провести интегрирование по всем параметрам с учетом

их апостериорных распределений

$$p(x_{new}|\boldsymbol{x}) = \int\int p(x|\theta,\alpha)p(\theta,\alpha|\boldsymbol{x})d\theta d\alpha = \Big\{p(\boldsymbol{x}|\theta,\alpha) = p(\boldsymbol{x}|\theta)\Big\} =$$
$$\int\int p(x|\theta)p(\theta|\alpha,\boldsymbol{x})p(\alpha|\boldsymbol{x})d\theta d\alpha = \Big\{p(\alpha|\boldsymbol{x}) \propto p(\boldsymbol{x}|\alpha)\Big\} = \frac{1}{Z}\int\int p(x|\theta)p(\theta|\alpha,\boldsymbol{x})p(\boldsymbol{x}|\alpha)d\theta d\alpha =$$
$$\Big\{p(\boldsymbol{x}|\alpha) \approx \delta(\alpha-\alpha_{ME})\Big\} = \frac{1}{Z}\int p(x|\theta)p(\theta|\alpha_{ME},\boldsymbol{x})d\theta = \frac{1}{Zp(\boldsymbol{x}|\alpha_{ME})}\int p(x|\theta)p(\boldsymbol{x}|\theta,\alpha_{ME})p(\theta|\alpha_{ME})d\theta =$$
$$\Big\{p(\boldsymbol{x}|\theta,\alpha_{ME})p(\theta|\alpha_{ME}) \approx \delta(\theta-\theta_{MP})\Big\} \propto p(x|\theta_{MP})$$

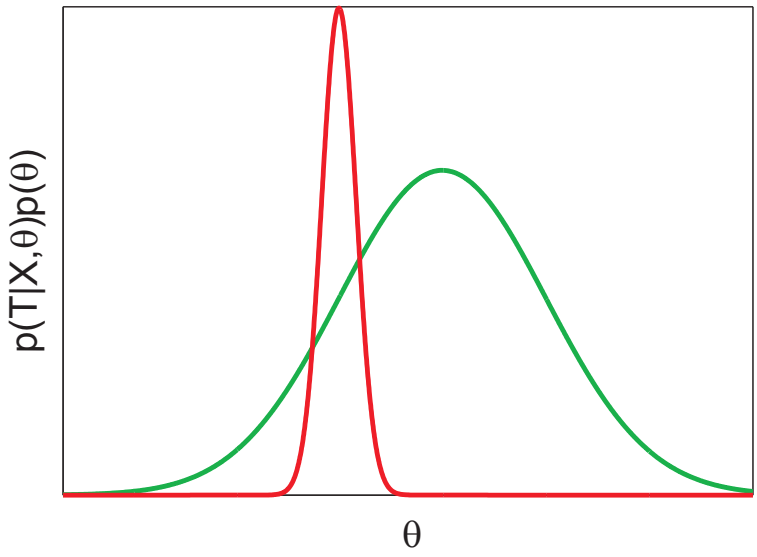


Рис. 7.6. В «пунктирной» модели присутствует небольшая доля алгоритмов, которые прекрасно объясняют обучающую выборку. В то же время, «сплошная» модель является более обоснованной, т.к. доля «хороших» алгоритмов в ней велика

7.3.2 Примеры использования

Генератор случайных чисел I

- Предположим, у нас имеется генератор случайных натуральных чисел. Мы знаем, что он может генерировать числа от 1 до N, причем $N = 10$, либо $N = 100$
- Распределение, по которому генерируются числа, нам неизвестно. Задача: оценить мат. ожидание этого распределения по выборке малой длины
- Пусть наша выборка состоит из двух наблюдений $x_1 = 8$, $x_2 = 6$ (для простоты положим порядок известным)
- Согласно принципу максимального правдоподобия легко показать, что $\mu_{ML} = \frac{1}{2}(x_1 + x_2) = 7$

Оценка максимального правдоподобия

- Пусть вероятности выпадения чисел $1, 2, \ldots, N$ равны, соответственно $q_1, q_2, \ldots, q_N$
- Тогда правдоподобие выборки $\boldsymbol{x} = (x_1, \ldots, x_n)$ равно
$$p(\boldsymbol{x}|\boldsymbol{q}) = \prod_{i=1}^{n} q_{x_i}$$
- Подставляя в формулу наши наблюдения получаем
$$p(x_1, x_2|\boldsymbol{q}) = q_8 q_6 \to \max_{\boldsymbol{q}}$$
- Учитывая, что все q_i неотрицательны и $\sum_{i=1}^{N} q_i = 1$, получаем
$$q_6^{ML} = q_8^{ML} = \frac{1}{2}, \quad q_i^{ML} = 0, \quad \forall i \neq 6, 8$$
- Отсюда мат. ожидание $\mu_{ML} = \sum_{i=1}^{N} i q_i^{ML} = \frac{1}{2}(x_1 + x_2) = 7$

Байесовская оценка вероятностей

- В отсутствие априорной информации о датчике случайных чисел, наиболее естественным является предположение о равномерности распределения вероятностей выпадения каждого числа $p(\boldsymbol{q}) = Const$
- Это частный случай распределения Дирихле
$$D(\boldsymbol{q}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha^0})} q_1^{\alpha_1^0 - 1} \ldots q_N^{\alpha_N^0 - 1}, \quad \sum_{i=1}^{N} q_i = 1, \; q_i \geq 0$$
при $\alpha_1^0 = \cdots = \alpha_N^0 = 1$
- Тогда применяя формулу Байеса, учитывая, что правдоподобие равно $p(x_1, x_2|\boldsymbol{q}) = q_8 q_6$, получаем
$$p(\boldsymbol{q}|x_1, x_2) = \frac{1}{Z} q_1^0 \ldots q_5^0 q_6^1 q_7^0 q_8^1 q_9^0 \ldots q_N^0 = D(\boldsymbol{q}|\boldsymbol{\alpha^1}),$$
где $\alpha_6^1 = \alpha_8^1 = 2$, а все остальные $\alpha_i^1 = 1$

Байесовская оценка мат. ожидания

- Чтобы получить точечные оценки вероятностей $q_1, \ldots, q_N$, возьмем мат. ожидание апостериорного распределения
- По свойству распределения Дирихле
$$\mathbb{E}q_i = \frac{\alpha_i}{\sum_{j=1}^{N} \alpha_j}$$
- Тогда, при $N = 10$ получаем
$$q_6 = q_8 = \frac{2}{12} = \frac{1}{6} \approx 0.16, \quad q_i = \frac{1}{12} \approx 0.08 \;\; \forall i \neq 6, 8$$
при $N = 100$ получаем
$$q_6 = q_8 = \frac{2}{102} = \frac{1}{51} \approx 0.02, \quad q_i = \frac{1}{102} \approx 0.01 \;\; \forall i \neq 6, 8$$
- Отсюда находим оценку мат. ожидания датчика, равную $\mu_{MP}(N = 10) = 5.75$ и $\mu_{MP}(N = 100) \approx 49.65$

Выбор наиболее обоснованной модели

- Итак, для двух различных моделей датчика мы получили два существенно разных ответа. Выберем наиболее обоснованную модель
- Обозначим обоснованность через Ev. Тогда справедливо следующее равенство

$$p(\boldsymbol{q}|\boldsymbol{x}) = \frac{q_8 q_6 \times q_1^0 \dots q_N^0}{B(\boldsymbol{\alpha^0})Ev} = \frac{q_8 q_6}{B(\boldsymbol{\alpha^1})},$$

  где $B(\boldsymbol{\alpha})$ — нормировочная константа в распределении Дирихле (многомерная бета-функция), равная

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{N} \Gamma(\alpha_i)}{\Gamma(\alpha_1 + \dots + \alpha_N)}$$

- Отсюда получаем формулу для обоснованности модели

$$Ev = \frac{B(\boldsymbol{\alpha^1})}{B(\boldsymbol{\alpha^0})}$$

- Как и следовало ожидать, $Ev(N = 10) > Ev(N = 100)$
- Зависимость обоснованности от N показана на рисунке 7.7

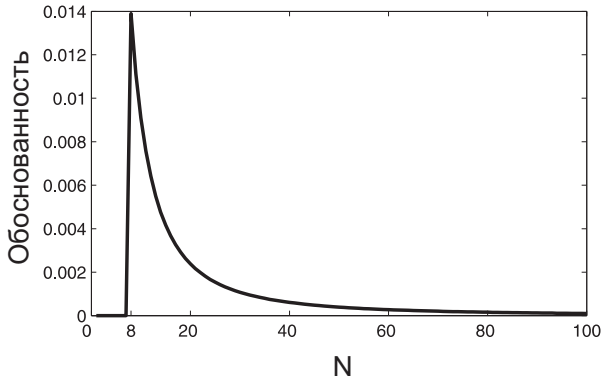


Рис. 7.7. При $N < 8$ обоснованность равна нулю, т.к. получить выборку $x_1 = 8$, $x_2 = 6$, применяя такой датчик, невозможно (функция правдоподобия всюду будет равна нулю). При больших N обоснованность падает, т.к. такие модели способны объяснить не только наши наблюдения, но и «много чего еще»

Глава 8

Метод релевантных векторов

Глава посвящена описанию метода релевантных векторов, являющегося примером успешного применения методов байесовского обучения и отправным пунктом для различных модификаций и обобщений, описанных в последующих главах. Рассматриваются задачи восстановления регрессии и классификации, показаны различия в применении метода наибольшей обоснованности для этих двух задач. Отдельное внимание уделено технике матричных вычислений, приведены основные матричные тождества.

8.1 Ликбез: Матричные тождества обращения

Матричные тождества обращения

- Эти тождества показывают, как изменяется матрица, если к ее обращению что-то добавляется.
- Тождество Шермана-Моррисона-Вудбери

$$(A^{-1} + UV^T)^{-1} = A - AU(I + V^T AU)^{-1}V^T A$$

- Лемма об определителе матрицы

$$\det(A^{-1} + UV^T) = \det(I + V^T AU)\det(A^{-1})$$

Тождество Шермана-Моррисона-Вудбери

- Тождество

$$(A^{-1} + UV^T)^{-1} = A - AU(I + V^T AU)^{-1}V^T A$$

- Доказательство

$$(A^{-1} + UV^T)(A - AU(I + V^T AU)^{-1}V^T A) = I + UV^T A - (U + UV^T AU)(I + V^T AU)^{-1}V^T A = \\ I + UV^T A - U(I + V^T AU)(I + V^T AU)^{-1}V^T A = I + UV^T A - UV^T A = I$$

Тождества для определителей матрицы

- При доказательстве многих матричных тождеств полезным оказывается следующее равенство:

$$\det\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(A)\det(D - CA^{-1}B) = \det(D)\det(A - BD^{-1}C)$$

  Здесь $A \in \mathbb{R}_{m\times m},\ B \in \mathbb{R}_{m\times n},\ C \in \mathbb{R}_{n\times m},\ D \in \mathbb{R}_{n\times n}$
- Это равенство следует из следующего тождества:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A & 0 \\ C & I \end{pmatrix}\begin{pmatrix} I & A^{-1}B \\ 0 & D - CA^{-1}B \end{pmatrix} = \begin{pmatrix} I & B \\ 0 & D \end{pmatrix}\begin{pmatrix} A - BD^{-1}C & 0 \\ D^{-1}C & I \end{pmatrix}$$

- Лемма об определителе матрицы

$$\det(A^{-1} + UV^T) = \det(I + V^T AU)\det(A^{-1})$$

- Доказательство:

$$\det\begin{pmatrix} A^{-1} & -U \\ V^T & I \end{pmatrix} = \det(A^{-1})\det(I + V^T AU) = \det(I)\det(A^{-1} + UI^{-1}V^T) = \det(A^{-1} + UV^T)$$

8.2 Метод релевантных векторов для задачи регрессии

Обобщенные линейные модели

- Рассмотрим следующую задачу восстановления регрессии: имеется выборка $(X, \boldsymbol{t}) = \{\boldsymbol{x}_i, t_i\}_{i=1}^n$, где вектор признаков $\boldsymbol{x}_i \in \mathbb{R}^d$, а целевая переменная $t_i \in \mathbb{R}$, требуется для нового объекта $\boldsymbol{x}_*$ предсказать значение целевой переменной t_*.
- Предположим, что $t = f(\boldsymbol{x}) + \varepsilon$, где $\varepsilon \sim \mathcal{N}(\varepsilon|0, \sigma^2)$, а

$$f(\boldsymbol{x}) = \sum_{j=1}^m w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$$

  Здесь $\boldsymbol{w}$ — набор числовых параметров, а $\boldsymbol{\phi}(\boldsymbol{x})$ — вектор обобщенных признаков.
- Часто в качестве обобщенных признаков выбираются следующие:
  - Обычные признаки — $\phi_j(\boldsymbol{x}) = x_j,\ j = 1, \ldots, d$
  - Ядровые функции — $\phi_j(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{x}_j),\ j = 1, \ldots, n,\ \phi_{n+1}(\boldsymbol{x}) \equiv 1$

Метод максимума правдоподобия (линейная регрессия)

- Так как шумовая компонента ε имеет независимое нормальное распределение, то можно выписать функцию правдоподобия обучающей выборки:

$$p(\boldsymbol{t}|X, \boldsymbol{w}) = \prod_{i=1}^n p(t_i|\boldsymbol{x}_i, \boldsymbol{w}) = \prod_{i=1}^n \mathcal{N}(t_i|f(\boldsymbol{x}_i, \boldsymbol{w}), \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\sigma^2}\right) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\sum_{i=1}^n (t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\sigma^2}\right)$$

- Переходя к логарифму, получаем

$$-\frac{1}{2\sigma^2}\sum_{i=1}^n (t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2 = -\frac{1}{2\sigma^2}(\boldsymbol{t} - \Phi\boldsymbol{w})^T(\boldsymbol{t} - \Phi\boldsymbol{w}) \to \max_{\boldsymbol{w}}$$

  Здесь $\Phi = [\boldsymbol{\phi}(\boldsymbol{x}_1)^T, \ldots, \boldsymbol{\phi}(\boldsymbol{x}_n)^T]^T$
- Точка максимума правдоподобия выписывается в явном виде:

$$\boldsymbol{w}_{ML} = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{t}$$

Введение регуляризации (априорного распределения)

- Следуя байесовскому подходу воспользуемся методом максимума апостериорной плотности:

$$w_{MP} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|X, \boldsymbol{t}) = \arg\max_{\boldsymbol{w}} p(\boldsymbol{t}|X, \boldsymbol{w})p(\boldsymbol{w})$$

- Выберем в качестве априорного распределения на параметры $\boldsymbol{w}$ следующее:

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}I)$$

  Такой выбор соответствует штрафу за большие значения коэффициентов $\boldsymbol{w}$ с параметром регуляризации α

- Максимизация апостериорной плотности эквивалентна следующей задаче оптимизации:

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(t_i - \boldsymbol{\phi}(\boldsymbol{x}_i))^2 - \frac{\alpha}{2}\|\boldsymbol{w}\|^2 \to \max_{\boldsymbol{w}}$$

- Решение

$$\boldsymbol{w}_{MP} = (\sigma^{-2}\Phi^T\Phi + \alpha I)^{-1}\sigma^{-2}\Phi^T\boldsymbol{t}$$

Линейная регрессия: обсуждение

- Высокая скорость обучения (достаточно сделать инверсию матрицы $\sigma^{-2}\Phi^T\Phi + \alpha I$ размера $m \times m$)
- Отсутствие способов автоматического выбора параметра регуляризации α и дисперсии шума σ^2 (параметров модели)
- Неразреженное решение (вообще говоря, все базисные функции входят в решающее правило с ненулевым весом)

Метод релевантных векторов

- Для получения разреженного решения введем в качестве априорного распределения на параметры $\boldsymbol{w}$ нормальное распределение с диагональной матрицей ковариации **с различными элементами на диагонали**:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \mathcal{N}(0, A^{-1})$$

  Здесь $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$. Такое априорное распределение соответствует независимой регуляризации вдоль каждого веса w_i со своим параметром регуляризации $\alpha_i \geq 0$
- Для подбора параметров модели $\boldsymbol{\alpha}, \sigma$ воспользуемся идеей максимизации обоснованности:

$$p(\boldsymbol{t}|\boldsymbol{\alpha}, \sigma^2) = \int p(\boldsymbol{t}|X, \boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} \to \max_{\boldsymbol{\alpha}, \sigma^2}$$

Вычисление обоснованности

- Обоснованность является сверткой двух нормальных распределений и может быть вычислена аналитически

$$p(\boldsymbol{t}|\boldsymbol{\alpha}, \sigma^2) = \int p(\boldsymbol{t}|X, \boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} = \int Q(\boldsymbol{w})d\boldsymbol{w}$$

- Рассмотрим функцию $L(\boldsymbol{w}) = \log Q(\boldsymbol{w})$. Она является квадратичной функцией и может быть представлена как:

$$L(\boldsymbol{w}) = L(\boldsymbol{w}_{MP}) + (\nabla_{\boldsymbol{w}}L(\boldsymbol{w}_{MP}))^T(\boldsymbol{w} - \boldsymbol{w}_{MP}) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{MP})^T H(\boldsymbol{w} - \boldsymbol{w}_{MP})$$

$$\boldsymbol{w}_{MP} = \arg\max_{\boldsymbol{w}} L(\boldsymbol{w}) \;\Rightarrow\; \nabla_{\boldsymbol{w}}L(\boldsymbol{w}_{MP}) = 0$$

$$H = \nabla\nabla L(\boldsymbol{w}_{MP})$$

- Тогда обоснованность может быть вычислена как

$$\int Q(\boldsymbol{w})d\boldsymbol{w} = \int \exp\left(L(\boldsymbol{w}_{MP}) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{MP})^T H(\boldsymbol{w} - \boldsymbol{w}_{MP})\right)d\boldsymbol{w} =$$

$$= Q(\boldsymbol{w}_{MP})\sqrt{(2\pi)^m}\sqrt{\det((-H)^{-1})} = \sqrt{(2\pi)^m}\frac{Q(\boldsymbol{w}_{MP})}{\sqrt{\det(-H)}}$$

✓ Упр.

Вычисление обоснованности

- Обозначив $\beta = \sigma^{-2}$, приводим подобные слагаемые в выражении для $L(\boldsymbol{w}_{MP})$

$$L(\boldsymbol{w}) = -\frac{1}{2}\beta(\boldsymbol{t} - \Phi\boldsymbol{w})^T(\boldsymbol{t} - \Phi\boldsymbol{w}) - \frac{1}{2}\boldsymbol{w}^T A\boldsymbol{w} - \frac{n}{2}\log(2\pi) - \\ - \frac{m}{2}\log(2\pi) + \frac{1}{2}\log\det(A) = -\frac{1}{2}\beta[\boldsymbol{t}^T\boldsymbol{t} - 2\boldsymbol{w}^T\Phi^T\boldsymbol{t} + \boldsymbol{w}^T\Phi^T\Phi\boldsymbol{w}] + C$$

- Приравнивая производную по $\boldsymbol{w}$ к нулю получаем значение $\boldsymbol{w}_{MP}$

$$\nabla L(\boldsymbol{w}) = -\frac{1}{2}\beta(-2\Phi^T\boldsymbol{t} + 2\Phi^T\Phi\boldsymbol{w}) - A\boldsymbol{w} = 0 \;\Rightarrow\; \boldsymbol{w}_{MP} = (\beta\Phi^T\Phi + A)^{-1}\beta\Phi^T\boldsymbol{t}$$

- Выделяем полный квадрат относительно $\boldsymbol{t}$ в выражении для $L(\boldsymbol{w}_{MP})$

$$L(\boldsymbol{w}_{MP}) = -\frac{1}{2}\left[\beta\boldsymbol{t}^T\boldsymbol{t} - 2\beta\boldsymbol{t}^T\Phi(\beta\Phi^T\Phi + A)^{-1}\beta\Phi^T\boldsymbol{t} + \boldsymbol{t}^T\Phi\beta(\beta\Phi^T\Phi + A)^{-1}\times\right. \\ \left.\times(\beta\Phi^T\Phi + A)(\beta\Phi^T\Phi + A)^{-1}\beta\Phi^T\boldsymbol{t}\right] + C = \\ - \frac{1}{2}\beta\boldsymbol{t}^T[I - 2\beta\Phi(\beta\Phi^T\Phi + A)^{-1}\Phi^T + +\Phi(\beta\Phi^T\Phi + A)^{-1}\beta\Phi^T]\boldsymbol{t} + C = \\ - \frac{1}{2}\beta\boldsymbol{t}^T[I - \beta\Phi(\beta\Phi^T\Phi + A)^{-1}\Phi^T]\boldsymbol{t} + C = \left\{(I - \beta\Phi(\beta\Phi^T\Phi + A)^{-1}\Phi^T)^{-1} = \{\text{Тож-во Вудбери}\} = \right. \\ \left. I + \beta\Phi(\beta\Phi^T\Phi\Phi^T\beta\Phi)^{-1}\Phi^T = I + \beta\Phi A^{-1}\Phi^T\right\} = \\ - 0.5\beta\boldsymbol{t}^T(I + \beta\Phi A^{-1}\Phi^T)^{-1}\boldsymbol{t} + C = -0.5\boldsymbol{t}^T(\beta^{-1}I + \Phi A^{-1}\Phi^T)^{-1}\boldsymbol{t} + C$$

- Таким образом выражение для обоснованности представляет собой гауссовское распределение относительно вектора $\boldsymbol{t}$, а значит нормализующая константа выписывается в явном виде

$$p(\boldsymbol{t}|X, \boldsymbol{\alpha}, \sigma^2) = \int p(\boldsymbol{t}|X, \sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} = \sqrt{(2\pi)^m}\frac{Q(\boldsymbol{w}_{MP})}{\sqrt{\det(-H)}} = \\ \sqrt{(2\pi)^m}\frac{\sqrt{\det A}}{(\sqrt{2\pi}\sigma)^n\sqrt{(2\pi)^m}}\exp\left(-\frac{1}{2}\boldsymbol{t}^T(\beta^{-1}I + \Phi A^{-1}\Phi^T)^{-1}\boldsymbol{t}\right)\frac{1}{\sqrt{\det(-H)}} = \\ \left\{H = -(\beta\Phi^T\Phi + A),\ \det(-H) = \det(\beta\Phi^T\Phi + A) = \det(A(I + \beta A^{-1}\Phi^T\Phi)) = \right. \\ \left. \det(A)\det(I + \beta A^{-1}\Phi^T\Phi) = \{\text{Лемма об опр-ле матр.}\} = \det(A)\det(I + \beta\Phi A^{-1}\Phi^T)\right\} = \\ \frac{1}{\sqrt{(2\pi)^n\det(\beta^{-1}I + \Phi A^{-1}\Phi^T)^{1/2}}}\exp\left(-\frac{1}{2}\boldsymbol{t}^T(\beta^{-1}I + \Phi A^{-1}\Phi^T)^{-1}\boldsymbol{t}\right)$$

Оптимизация обоснованности

✓ Упр.

- Приравнивая к нулю производные обоснованности по $\boldsymbol{\alpha}, \sigma^2$, можно получить итерационные формулы для пересчета параметров:

$$\alpha_i^{new} = \frac{\gamma_i}{w_{MP,i}^2} \qquad \gamma_i = 1 - \alpha_i^{old}\Sigma_{ii}$$

$$(\sigma^2)^{new} = \frac{\|\boldsymbol{t} - \Phi\boldsymbol{w}\|^2}{n - \sum_{i=1}^m \gamma_i}$$

Здесь $\Sigma = (\beta\Phi^T\Phi + A)^{-1}$, $\boldsymbol{w}_{MP} = \beta\Sigma\Phi^T\boldsymbol{t}$.

- Параметр γ_i может интерпретироваться как степень, в которой соответствующий вес w_i определяется данными или регуляризацией. Если α_i велико, то вес w_i существенно предопределен априорным распределением, $\Sigma_{ii} \simeq \alpha_i^{-1}$ и $\gamma_i \simeq 0$. С другой стороны для малых значений α_i значение веса w_i полностью определяется данными, $\gamma_i \simeq 1$.

Принятие решения

✓ Упр.

- Зная значения $\boldsymbol{\alpha}_{MP}, \sigma^2_{MP}$ можно вычислить распределение прогноза :

$$p(t_*|\boldsymbol{x}_*, \boldsymbol{t}, X) = \int p(t_*|\boldsymbol{x}_*, \boldsymbol{w}, \sigma^2_{MP}) p(\boldsymbol{w}|\boldsymbol{t}, X, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP}) d\boldsymbol{w} = \mathcal{N}(t_*|y_*, \sigma^2_*)$$

Здесь

$$y_* = \boldsymbol{w}^T_{MP}\boldsymbol{\phi}(\boldsymbol{x}_*)$$
$$\sigma^2_* = \sigma^2_{MP} + \boldsymbol{\phi}(\boldsymbol{x}_*)^T \Sigma \boldsymbol{\phi}(\boldsymbol{x}_*)$$

Алгоритм 1: Метод релевантных векторов для задачи регрессии

Вход: Обучающая выборка $\{\boldsymbol{x}_i, t_i\}_{i=1}^n$ $\boldsymbol{x}_i \in \mathbb{R}^d$, $t_i \in \mathbb{R}$; Матрица обобщенных признаков $\Phi = \{\phi_j(\boldsymbol{x}_i)\}_{i,j=1}^{n,m}$;
Выход: Набор весов $\boldsymbol{w}$, матрица Σ и оценка дисперсии шума β^{-1} для решающего правила $t_*(\boldsymbol{x}) = \sum_{j=1}^m w_j \phi_j(\boldsymbol{x})$, $\sigma^2_*(\boldsymbol{x}) = \beta^{-1} + \boldsymbol{\phi}^T(\boldsymbol{x}_*)\Sigma\boldsymbol{\phi}(\boldsymbol{x}_*)$;

```
инициализация: α_i := 1, i = 1,...,m, β := 1, AlphaBound := 10^12, WeightBound := 10^-6, NumberOfIterations := 100;
для k = 1,...,NumberOfIterations
    A := diag(α_1,...,α_m);
    Σ := (βΦ^TΦ + A)^-1;
    w_MP := ΣβΦ^T t;
    для j = 1,...,m
        если w_MP,j < WeightBound или α_j > AlphaBound то
            w_MP,j := 0, α_j := +∞, γ_j := 0;
        иначе
            γ_j := 1 − α_j Σ_jj, α_j := γ_j / w^2_MP,j;
    β := (n − Σ_{j=1}^m γ_j) / ||t − Φw_MP||^2
```

Метод релевантных векторов для регрессии: обсуждение

- На практике процесс обучения обычно требует 20—50 итераций. На каждой итерации вычисляется $\boldsymbol{w}_{MP}$ (это требует обращения матрицы размера $m \times m$), а также пересчитываются значения $\boldsymbol{\alpha}$, σ^2 (практически не требует времени). Как следствие, скорость обучения метода падает в 20-50 раз по сравнению с линейной регрессией.
- При использовании ядровых функций в качестве обобщенных признаков необходимо проводить скользящий контроль для различных значений параметров ядровых функций. В этом случае время обучения возрастает еще в несколько раз.
- Параметры регуляризации $\boldsymbol{\alpha}$ и дисперсии шума в данных σ^2 подбираются автоматически.
- На выходе получается разреженное решение, т.е. только небольшое количество исходных объектов входят в решающее правило с ненулевым весом.
- Кроме значения прогноза y_* алгоритм выдает также дисперсию прогноза σ^2_*.

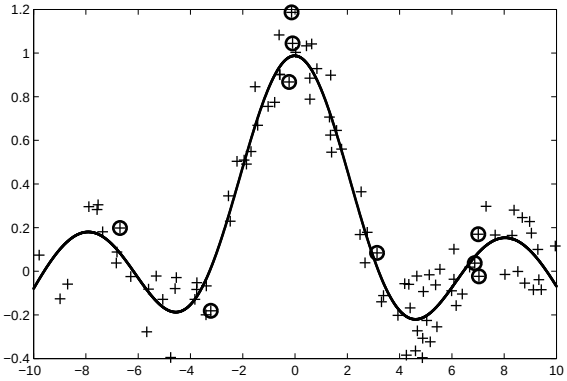


Рис. 8.1. Пример применения регрессии релевантных векторов для зашумленной функции $y(x) = \mathrm{sinc}(x)$. В качестве базисных функций использовались $\phi_j(\boldsymbol{x}) = \exp(-\beta\|\boldsymbol{x} - \boldsymbol{x}_j\|^2)$. Объекты, соответствующие релевантным базисным функциям, обведены в кружочки. В процессе обучения большинство α_j стремятся к $+\infty$. Таким образом, априорное распределение на соответствующий вес w_j становится вырожденным, что соответствует ситуации $w_j = 0$, т.е. исключению данной базисной функции из модели

8.3 Метод релевантных векторов для задачи классификации

Задача классификации

- Рассмотрим следующую задачу классификации на два класса: имеется выборка $(X, \boldsymbol{t}) = \{\boldsymbol{x}_i, t_i\}_{i=1}^n$, где вектор признаков $\boldsymbol{x}_i \in \mathbb{R}^d$, а целевая переменная $t_i \in \{+1, -1\}$, требуется для нового объекта $\boldsymbol{x}_*$ предсказать значение целевой переменной t_*.
- Воспользуемся обобщенными линейными моделями для классификации:

$$\hat{t}(\boldsymbol{x}) = \mathrm{sign}(f(\boldsymbol{x})) = \mathrm{sign}\left(\sum_{j=1}^m w_j \phi_j(\boldsymbol{x})\right) = \mathrm{sign}\left(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})\right)$$

  Здесь $\boldsymbol{w}$ — набор числовых параметров, а $\boldsymbol{\phi}(\boldsymbol{x})$ — вектор обобщенных признаков.

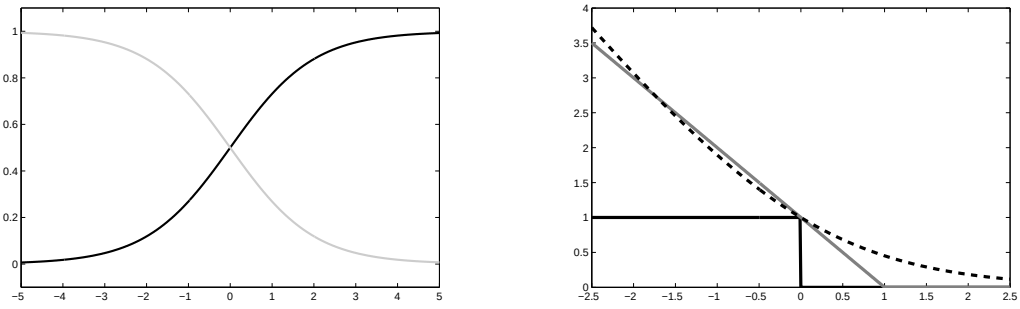
Метод максимума правдоподобия (логистическая регрессия)

- В качестве функции правдоподобия выберем произведение логистических функций (см. рис. 8.2a):

$$p(\boldsymbol{t}|X, \boldsymbol{w}) = \prod_{i=1}^n p(t_i|\boldsymbol{x}_i, \boldsymbol{w}) = \prod_{i=1}^n \frac{1}{1 + \exp(-t_i f(\boldsymbol{x}_i))}$$

- Переходя к логарифму правдоподобия, получаем (см. рис. 8.2b):

$$-\sum_{i=1}^n \log(1 + \exp(-t_i f(\boldsymbol{x}_i))) \to \max_{\boldsymbol{w}}$$



(a) (b)

Рис. 8.2. На рисунке (a) показаны логистические функции правдоподобия правильной классификации объектов из первого и второго классов. На рисунке (b) изображены различные виды функционалов, штрафующих ошибку на обучении: количество ошибок (черная кривая), функция потерь в методе опорных векторов (hinge loss, серая кривая), логарифм логистической функции (пунктирная линия)

Оптимизация функции правдоподобия (IRLS)

- Функция $-\log(1+\exp(-x))$ является вогнутой, поэтому логарифм правдоподобия как сумма вогнутых функций также является вогнутой функцией и имеет единственный максимум.
- Для поиска максимума логарифма правдоподобия L воспользуемся методом Ньютона:

$$\boldsymbol{w}^{new} = \boldsymbol{w}^{old} - H^{-1}\nabla L(\boldsymbol{w})$$

  Здесь $H = \nabla\nabla L(\boldsymbol{w})$ — гессиан логарифма правдоподобия.
- Вычисляя градиент и гессиан, получаем формулы пересчета:

$$\boldsymbol{w}^{new} = (\Phi^T R\Phi)^{-1}\Phi^T R\boldsymbol{z}$$
$$\boldsymbol{z} = \Phi\boldsymbol{w}^{old} - R^{-1}\operatorname{diag}(\boldsymbol{t})\boldsymbol{s}$$

  Здесь $s_i = \frac{1}{1+\exp(-t_i f(\boldsymbol{x}_i))}$, $\quad R = \operatorname{diag}(s_1(1-s_1),\ldots,s_n(1-s_n))$.

Введение регуляризации

- По аналогии с линейной регрессией можно рассмотреть максимум апостериорной плотности с нормальным априорным распределением с единичной матрицей ковариации, умноженной на коэффициент α^{-1}:

$$-\sum_{i=1}^{n}\log(1+\exp(-t_i f(\boldsymbol{x}_i))) - \frac{\alpha}{2}\|\boldsymbol{w}\|^2 \to \max_{\boldsymbol{w}}$$

- Метод оптимизации меняется следующим образом:

$$\boldsymbol{w}^{new} = (\Phi^T R\Phi + \alpha I)^{-1}\Phi^T R\boldsymbol{z}$$
$$\boldsymbol{z} = \Phi\boldsymbol{w}^{old} - R^{-1}\operatorname{diag}(\boldsymbol{t})\boldsymbol{s}$$

Логистическая регрессия: обсуждение

- По-прежнему довольно высокая скорость работы. На практике обучение часто требует всего 3—7 итераций.
- Отсутствие способа автоматического выбора параметра регуляризации α
- Неразреженное решение

Метод релевантных векторов

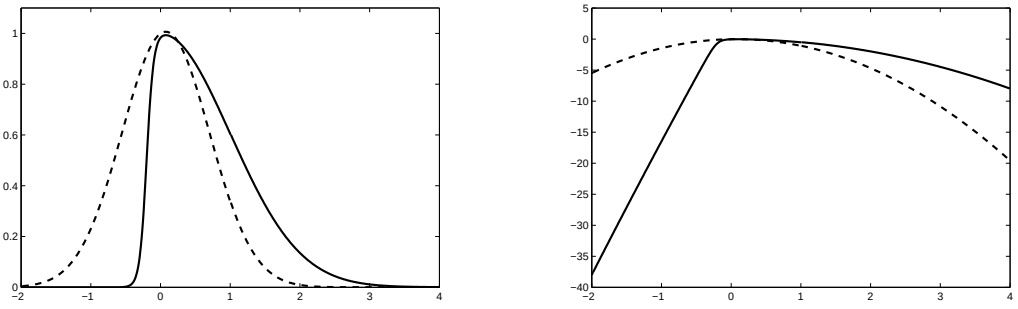
- Для получения разреженного решения введем в качестве априорного распределения на параметры $\boldsymbol{w}$ нормальное распределение с диагональной матрицей ковариации **с различными элементами на диагонали**:
$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{w}|0, A^{-1})$$
Здесь $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$.
- Для подбора параметров модели $\boldsymbol{\alpha}$ воспользуемся идеей максимизации обоснованности:
$$p(\boldsymbol{t}|\boldsymbol{\alpha}) = \int p(\boldsymbol{t}|X, \boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} \to \max_{\boldsymbol{\alpha}}$$

Приближение Лапласа

- Рассмотрим функцию $p(z) = \exp\left(-\frac{z^2}{2}\right)\frac{1}{1+\exp(-20z-4)}$ (см. рис. 8.3).
- Разложим логарифм функции в ряд Тейлора в точке максимума:
$$z_0 = \arg\max_z f(z), \quad \log f(z) \simeq \log f(z_0) + \frac{H}{2}(z - z_0)^2, \quad H = \left.\frac{d^2 f}{dz^2}\right|_{z=z_0}$$
- Тогда функцию $f(z)$ можно приблизить следующим образом (см. рис. 8.3):
$$f(z) \simeq f(z_0)\exp\left(\frac{H}{2}(z - z_0)^2\right)$$

Вычисление обоснованности

- Подынтегральное выражение в обоснованности является произведением логистических функций и нормального распределения. Такой интеграл не берется аналитически.
- Решение: приблизить подынтегральную функцию гауссианой, интеграл от которой легко берется. Для приближения воспользуемся методом Лапласа:
$$p(\boldsymbol{t}|\boldsymbol{\alpha}) = \int p(\boldsymbol{t}|X, \boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} = \int Q(\boldsymbol{w})d\boldsymbol{w} \simeq \sqrt{(2\pi)^m}\frac{Q(\boldsymbol{w}_{MP})}{\sqrt{\det(-\nabla\nabla \log Q(\boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_{MP}})}},$$
где
$$\boldsymbol{w}_{MP} = \arg\max_{\boldsymbol{w}} Q(\boldsymbol{w})$$



(a) (b)

Рис. 8.3. Функция правдоподобия (рисунок (a)) и ее логарифм (рисунок (b)) вместе с соответствующим приближением Лапласа

Оптимизация обоснованности

Приравнивая к нулю производную логарифма обоснованности по $\boldsymbol{\alpha}$, получаем:

$$\log p(\boldsymbol{t}|X,\boldsymbol{\alpha}) = \log Q(\boldsymbol{w}_{MP}) - \frac{1}{2}\log\det(-H) + C = \\ - \sum_{i=1}^{n}\log(1+\exp(-t_i f(\boldsymbol{x}_i,\boldsymbol{w}_{MP}))) - \frac{1}{2}\sum_{i=1}^{m}\alpha_i w_{MP,i}^2 - \frac{1}{2}\log\det(-H) + \frac{1}{2}\sum_{i=1}^{m}\log\alpha_i + C$$

✓ Упр. Здесь $H = -\Phi^T R\Phi - A,\ R = \mathrm{diag}(s_1(1-s_1),\ldots,s_n(1-s_n)),\ s_i = \frac{1}{1+\exp(-t_i f(\boldsymbol{x}_i,\boldsymbol{w}_{MP}))}$.

$$\frac{\partial}{\partial\alpha_j}\log p(\boldsymbol{t}|X,\boldsymbol{\alpha}) = -\frac{1}{2}w_{MP,j}^2 - \frac{1}{2}\det(\Phi^T R\Phi + A)^{-1}\det(\Phi^T R\Phi + A)\times((\Phi^T R\Phi + A)^{-1})_{jj} + \frac{1}{2\alpha_j} = 0$$

Отсюда получаем итерационные формулы пересчета $\boldsymbol{\alpha}$, аналогичные регрессии:

$$\alpha_i^{new} = \frac{1-\alpha_i^{old}\Sigma_{ii}}{w_{MP,i}^2}$$

Метод релевантных векторов: обсуждение

- На практике процесс обучения обычно требует 20—50 итераций. На каждой итерации вычисляется $\boldsymbol{w}_{MP}$ (это требует 3—7 итераций с обращениями матрицы размера $m\times m$), а также пересчитываются значения $\boldsymbol{\alpha}$ (практически не требует времени). Как следствие, скорость обучения метода падает в 20—50 раз по сравнению с логистической регрессией.
- При использовании ядровых функций в качестве обобщенных признаков необходимо проводить скользящий контроль для различных значений параметров ядровых функций. В этом случае время обучения возрастает еще в несколько раз.
- Параметры регуляризации $\boldsymbol{\alpha}$ подбираются автоматически.
- На выходе получается разреженное решение.

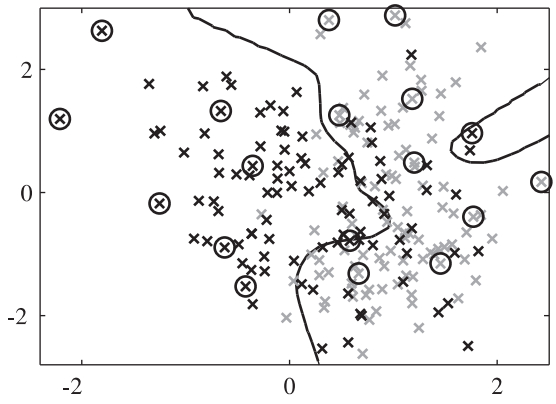


Рис. 8.4. Пример применения метода релевантных векторов для задачи классификации. В качестве базисных функций использовались $\phi_j(x) = \exp(-\beta(x-x_j)^2)$. Объекты, соответствующие релевантным базисным функциям, обведены в кружочки. В процессе обучения большинство α_j стремятся к $+\infty$. Таким образом, априорное распределение на соответствующий вес w_j становится вырожденным, что соответствует ситуации $w_j = 0$, т.е. исключению данной базисной функции из модели

- Для вычисления дисперсии прогноза необходимо проводить дополнительно аппроксимацию интеграла

$$p(t_*|\boldsymbol{x}_*, \boldsymbol{t}, X) = \int p(t_*|\boldsymbol{x}_*, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{t}, X, \boldsymbol{\alpha}_{MP}) d\boldsymbol{w}$$

Алгоритм 2: Метод релевантных векторов для задачи классификации

Вход: Обучающая выборка $\{\boldsymbol{x}_i, t_i\}_{i=1}^n$ $\boldsymbol{x}_i \in \mathbb{R}^d$, $t_i \in \{+1, -1\}$; Матрица обобщенных признаков $\Phi = \{\phi_j(\boldsymbol{x}_i)\}_{i,j=1}^{n,m}$;

Выход: Набор весов $\boldsymbol{w}$ для решающего правила $t_*(\boldsymbol{x}) = \sum_{j=1}^m w_j \phi_j(\boldsymbol{x})$;

инициализация: $\alpha_i := 1$, $i = 1, \ldots, m$, $\boldsymbol{w}_{MP} = \boldsymbol{t}$, AlphaBound $:= 10^{12}$, WeightBound $:= 10^{-6}$, NumberOfIterations $:= 100$;
для $k = 1, \ldots,$ NumberOfIterations
  $A := \operatorname{diag}(\alpha_1, \ldots, \alpha_m)$;
  повторять
    для $i = 1, \ldots, n$
      $s_i := \frac{1}{(1+\exp(-t_i \sum_{j=1}^m w_{MP,j} \phi_j(\boldsymbol{x}_i)))}$;
    $R := \operatorname{diag}(s_1(1-s_1), \ldots, s_n(1-s_n))$;
    $\boldsymbol{z} := \Phi \boldsymbol{w}_{MP} - R^{-1}(\boldsymbol{s} - \boldsymbol{t})$;
    $\Sigma := (\Phi^T R \Phi + A)^{-1}$;
    $\boldsymbol{w}_{MP} := \Sigma \Phi^T R \boldsymbol{z}$;
  пока $\|\boldsymbol{w}_{MP}^{new} - \boldsymbol{w}_{MP}^{old}\|$ меняется больше, чем на заданную величину
  для $j = 1, \ldots, m$
    если $w_{MP,j} <$ WeightBound или $\alpha_j >$ AlphaBound **то**
      $w_{MP,j} := 0$, $\alpha_j := +\infty$, $\gamma_j := 0$;
    иначе
      $\alpha_j := \frac{1-\alpha_j \Sigma_{jj}}{w_{MP,j}^2}$;

Глава 9

Недиагональная регуляризация обобщенных линейных моделей

В главе рассматриваются ограничения и недостатки метода релевантных векторов и способы их преодоления. Описана идея регуляризации степеней свободы алгоритма классификации, соответствующая использованию недиагональной матрицы регуляризации весов классификатора. Рассматривается регуляризация с помощью введения лапласовского априорного распределения и ее отличительные особенности.

9.1 Ликбез: Неотрицательно определенные матрицы и Лапласовское распределение

Неотрицательно определенные матрицы

- Матрица $A \in \mathbb{R}^{n \times n}$ называется неотрицательно определенной, если соответствующая ей квадратичная форма всегда неотрицательна, т.е. $\forall \boldsymbol{x} \in \mathbb{R}^n$

$$\langle A\boldsymbol{x}, \boldsymbol{x} \rangle \geq 0.$$

- Матрица A называется симметричной, если $A^T = A$
- В частности, множество всех n-мерных нормальных распределений с центром в нуле изоморфно множеству всех неотрицательно определенных симметричных матриц

$$\mathcal{N}(\boldsymbol{x}|\mathbf{0}, A^{-1}) = \frac{\sqrt{\det(A)}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x}\right), \quad A^T = A, \quad A \geq 0.$$

Свойство неотрицательно определенных симметричных матриц

- Из линейной алгебры известно, что любая симметричная (самосопряженная) матрица может быть приведена к диагональному виду линейным преобразованием координат, т.е. $\exists P : \ P^T = P^{-1}$, $\det(P) \neq 0$ такая, что

$$\Lambda = P^T A P = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

- Если дополнительно известно, что матрица A неотрицательно определена, то все $\lambda_i \geq 0$
- Количество ненулевых λ_i называется **рангом матрицы** A

Лапласовское распределение

- Распределением Лапласа называется двустороннее показательное распределение

$$\mathcal{L}(x|\lambda) = \frac{\lambda}{4} \exp\left(-\frac{\lambda}{2}|x|\right)$$

- Распределение Лапласа имеет сингулярность в нуле и более тяжелые хвосты, чем нормальное распределение
- В логарифмической шкале введение априорного распределения Лапласа на веса означает т.н. $L1$-регуляризацию функционала качества

$$\Phi(\theta) = \log p(x|\theta) - \frac{\alpha}{2} \sum_{j=1}^{m} |\theta_j| \to \max_{\theta}$$

Свойство лапласовского регуляризатора

При использовании лапласовского априорного распределения на веса, многие веса могут оказаться равными нулю. Это связано с тем, что использование регуляризации эквивалентно ограничению области поиска экстремума исходной (нерегуляризованной) функции (серая область на рис. 9.1) При лапласовском распределении соответствующая область имеет характерные изломы, в которых может находиться условный экстремум исходной функции. В то же время, вероятность того, что хотя бы один из весов окажется равным нулю при использовании гауссовского априорного распределения равна нулю

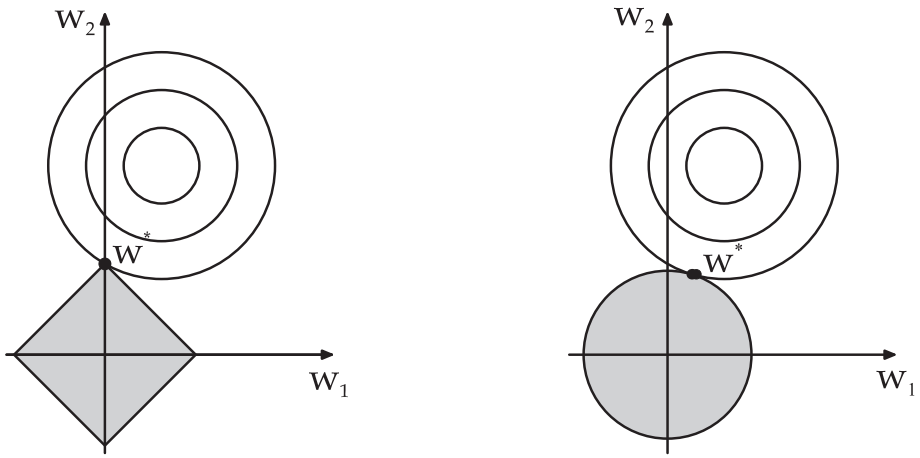


Рис. 9.1. Влияние вида регуляризатора на положение экстремума регуляризованной функции

9.2 Метод релевантных собственных векторов

9.2.1 RVM и его ограничения

Регуляризованная логистическая регрессия (напоминание)

- Рассматривается стандартная задача классификации на 2 класса с обучающей выборкой $(X, \boldsymbol{t}) = \{\boldsymbol{x}_i, t_i\}_{i=1}^n$, где $\boldsymbol{x} \in \mathbb{R}^d$ и $t \in \{-1, 1\}$
- Классификатор имеет форму

$$\hat{t}(\boldsymbol{x}) = \text{sign}(y(\boldsymbol{x}_i)) = \text{sign} \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})),$$

  где $\{\phi_j(\boldsymbol{x})\}_{j=1}^m$ — множество заранее фиксированных базисных функций, а $\boldsymbol{w}$ — вектор весов, настраиваемых в ходе обучения путем максимизации регуляризованного логарифма правдоподобия

$$\boldsymbol{w}_{MP} = \arg\max \left(\log p(\boldsymbol{t}|X, \boldsymbol{w}) - \alpha \boldsymbol{w}^T \boldsymbol{w} \right),$$

  где

$$p(\boldsymbol{t}|X, \boldsymbol{w}) = \prod_{i=1}^{n} \frac{1}{1 + \exp(-t_i y(\boldsymbol{x}_i))}.$$

- С вероятностной точки зрения это соответствует введению гауссовского априорного распределения на веса $\boldsymbol{w}$ с центром в нуле и ковариационной матрицей $\Sigma = \alpha^{-1} I$
- Назовем матрицу $A = \Sigma^{-1}$ матрицей регуляризации. В данном случае матрица регуляризации равна αI

Метод релевантных векторов (напоминание)

- В 1999г. М. Типпинг предложил установить индивидуальные коэффициенты регуляризации α_j для каждого веса w_j.
- Это соответствует использованию гауссовского априорного распределения с произвольной неотрицательно определенной диагональной матрицей регуляризации

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \sqrt{\frac{\det(A)}{(2\pi)^m}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T A \boldsymbol{w}\right),$$

  где $A = \operatorname{diag}(\alpha_1, \ldots, \alpha_m)$, $\alpha_j \geq 0$.
- Для определения значений коэффициентов регуляризации использовался принцип наибольшей обоснованности

$$\boldsymbol{\alpha}_{ME} = \arg\max p(\boldsymbol{t}|X, \boldsymbol{\alpha}) = \arg\max \int p(\boldsymbol{t}|X, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\boldsymbol{w}.$$

- Метод релевантных векторов обладает интересным свойством разреженности получаемого классификатора
- Большинство α_j уходят в $+\infty$, эффективно исключая нерелевантные базисные функции и делая классификатор разреженным

Недостатки RVM

- При обучении классификатора RVM требуется порядка 20–50 итераций для настройки $\boldsymbol{\alpha}$, на каждой из которых приходится обучать метод регуляризованной логистической регрессии
- RVM не может быть напрямую применен для лапласовского априорного распределения на веса $\boldsymbol{w}$. В то же время известно, что лапласовское априорное распределение обычно приводит к более разреженным решающим правилам
- И регуляризованная логистическая регрессия, и метод релевантных векторов не инвариантны относительно линейных преобразований базисных функций

Линейная неинвариантность

- Рассмотрим невырожденную матрицу $L \in \mathbb{R}^{m \times m}$
- Пусть $\boldsymbol{\psi}(\boldsymbol{x}) = L\boldsymbol{\phi}(\boldsymbol{x})$ — новое множество базисных функций
- Поскольку наш классификатор линеен по базисным функциям, вполне естественно требовать, чтобы классификатор, обученный по базисным функциям $\boldsymbol{\psi}(\boldsymbol{x})$, был эквивалентен классификатору, полученному при использовании базисных функций $\boldsymbol{\phi}(\boldsymbol{x})$
- К сожалению, это не так в случае RVM и регуляризованной логистической регрессии

9.2.2 Регуляризация степеней свободы

Степени свободы

- Идея: Что будет, если регуляризовывать не веса w_j, а т.н. степени свободы (направления, ассоциированные с собственными векторами гессиана логарифма правдоподобия)?
- Известно, что в большинстве случаев в окрестностях точки максимума функция правдоподобия достаточно хорошо может быть приближена гауссианой. Главные оси соответствующей матрицы ковариации определяют степени свободы классификатора

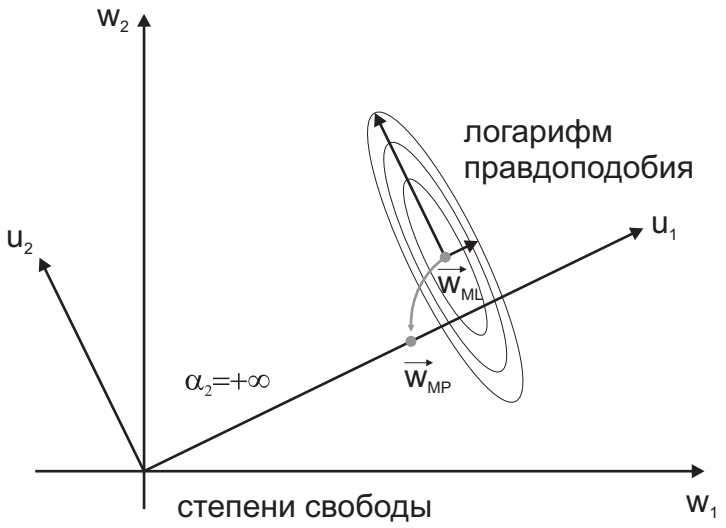


Рис. 9.2. Регуляризация степеней свободы, ассоциированных с собственными векторами гессиана правдоподобия

- Мотивация: Один и тот же вес может входить и в значимые, и в нерелевантные степени свободы
- Следствие: Такая регуляризация немедленно становится инвариантной относительно линейных преобразований множества базисных функций
- С точки зрения исходных весов $\boldsymbol{w}$ этот подход соответствует использованию недиагональной симметричной матрицы регуляризации

Приближение функции правдоподобия

- Пусть $\boldsymbol{u}$ — новые переменные, ассоциированные с каждой степенью свободы
- Приблизим правдоподобие гауссианой в точке максимума, используя приближение Лапласа

$$p(\boldsymbol{t}|X,\boldsymbol{u}) \approx \hat{p}(\boldsymbol{t}|X,\boldsymbol{u}) = p(\boldsymbol{t}|X,\boldsymbol{u}_{ML}) \exp\left(-\frac{1}{2}(\boldsymbol{u}-\boldsymbol{u}_{ML})^T Q H Q^T (\boldsymbol{u}-\boldsymbol{u}_{ML})\right),$$

  где $Q^T = Q^{-1}$ — матрица перехода от $\boldsymbol{u}$ к $\boldsymbol{w}$, т.е. $\boldsymbol{w} = Q^T\boldsymbol{u}$
- Приближенное правдоподобие $\hat{p}(\boldsymbol{t}|X,\boldsymbol{u}) = \hat{p}(\boldsymbol{t}|X,\boldsymbol{w})$ является сепарабельной функцией от $\boldsymbol{u}$, т.е. может быть представлено в виде произведения функций, зависящих от одной компоненты вектора $\boldsymbol{u}$

$$\hat{p}(\boldsymbol{t}|X,\boldsymbol{u}) = p(\boldsymbol{t}|X,\boldsymbol{u}_{ML}) \prod_{j=1}^{m} g(u_j, u_{ML,j}, h_j)$$

9.2.3 Оптимизация обоснованности для различных семейств априорных распределений

Вычисление обоснованности

- Поскольку мы ввели независимую регуляризацию каждой степени свободы u_j, регуляризатор также является сепарабельной функцией от $\boldsymbol{u}$

- Следовательно, обоснованность представима в виде произведения одномерных интегралов, каждый из которых зависит только от одного коэффициента регуляризации α_j, и можно оптимизировать все α_j одновременно и независимо.
- Обозначив собственные вектора матрицы $-H$ за $\{h_j\}$, получаем

$$p(\boldsymbol{t}|X,\boldsymbol{\alpha}) \approx \int \hat{p}(\boldsymbol{t}|X,\boldsymbol{u})p(\boldsymbol{u}|\boldsymbol{\alpha})d\boldsymbol{u} = p(\boldsymbol{t}|X,\boldsymbol{u}_{ML}) \int \prod_{j=1}^{m} g(u_j,u_{ML,j},h_j)p(u_j|\alpha_j)d\boldsymbol{u} =$$

$$p(\boldsymbol{t}|X,\boldsymbol{u}_{ML}) \prod_{j=1}^{m} \int g(u_j,u_{ML,j},h_j)p(u_j|\alpha_j)du_j = p(\boldsymbol{t}|X,\boldsymbol{u}_{ML}) \prod_{j=1}^{m} f_j(h_j,u_{ML,j},\alpha_j)$$

Гауссовское априорное распределение

В случае, когда степени свободы имеют гауссовское априорное распределение $u_j \sim \mathcal{N}(u_j|0,\alpha_j^{-1})$, одномерный интеграл и оптимальное значение для α_j могут быть получены аналитически

$$f_j^G(h_j,u_{ML,j},\alpha_j) \;=\; \sqrt{\frac{\alpha_j}{2\pi}} \int \exp\left(-\frac{h_j}{2}(u_j-u_{ML,j})^2 - \frac{\alpha_j}{2}u_j^2\right) du_j \;=\; \sqrt{\frac{\alpha_j}{h_j+\alpha_j}} \exp\left(-\frac{h_j\alpha_j u_{ML,j}^2}{2(h_j+\alpha_j)}\right),$$

$$\alpha_j^* = \begin{cases} \frac{h_j}{h_j u_{ML,j}^2 - 1}, & \text{если } h_j u_{ML,j}^2 > 1 \\ +\infty, & \text{иначе} \end{cases}$$

В частности, условие на релевантность степени свободы удается получить в явном виде.

Алгоритм 3: Метод релевантных собственных векторов

Вход: Обучающая выборка $\{\boldsymbol{x}_i,t_i\}_{i=1}^n$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $t_i \in \{+1,-1\}$; Матрица обобщенных признаков $\Phi = \{\phi_j(\boldsymbol{x}_i)\}_{i,j=1}^{n,m}$;

Выход: Набор весов $\boldsymbol{w}_{MP}$ для решающего правила $t_*(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{j=1}^m w_{MP,j}\phi_j(\boldsymbol{x})\right)$;

1: Найти $\boldsymbol{w}_{ML} = \arg\max p(\boldsymbol{t}|X,\boldsymbol{w})$;
2: Вычислить гессиан $H = \nabla\nabla \log p(\boldsymbol{t}|X,\boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_{ML}}$;
3: Вычислить собственные вектора и собственные значения гессиана $-H = Q^T\Lambda Q$, $\Lambda = \operatorname{diag}(h_1,\ldots,h_m)$;
4: Вычислить $\boldsymbol{u}_{ML} = Q\boldsymbol{w}_{ML}$;
5: **для** $j = 1,\ldots,m$
6: **если** $h_j u_{ML,j}^2 > 1$ **то**
7: $\alpha_j^* := \frac{h_j}{h_j u_{ML,j}^2 - 1}$;
8: **иначе**
9: $\alpha_j^* := +\infty$
10: Найти $\boldsymbol{w}_{MP} = \arg\max p(\boldsymbol{t}|X,\boldsymbol{w})p(Q\boldsymbol{w}|\boldsymbol{\alpha}^*)$

Лапласовское априорное распределение

- В случае, когда степени свободы имеют априорное распределение Лапласа $p(u_j|\alpha_j) = \mathcal{L}(u_j|\alpha_j^{-1})$, интеграл также может быть вычислен аналитически

- Для этого разобьем интеграл на два

$$f_j^L(h_j, u_{ML,j}, \alpha_j) = \int_{-\infty}^{+\infty} g(u_j, u_{ML,j}, h_j) p(u_j|\alpha_j) du_j =$$

$$\int_{-\infty}^{0} g(u_j, u_{ML,j}, h_j) p(u_j|\alpha_j) du_j + \int_{0}^{+\infty} g(u_j, u_{ML,j}, h_j) p(u_j|\alpha_j) du_j =$$

$$\frac{\alpha_j}{4} \int_{-\infty}^{0} \exp\left(-\frac{h_j(u_j - u_{ML,j})^2}{2} - \frac{\alpha_j}{2}|u_j|\right) du_j + \frac{\alpha_j}{4} \int_{0}^{+\infty} \exp\left(-\frac{h_j(u_j - u_{ML,j})^2}{2} - \frac{\alpha_j}{2}|u_j|\right) du_j$$

Вычисление интеграла

- Обе «половинки» представляют собой интегралы от квадратных трехчленов под экспонентой, которые легко вычисляются с помощью «интеграла ошибок»

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{+\infty} \exp\left(-\xi^2\right) d\xi$$

✓ Упр.

- Обозначим $x_1 = \sqrt{\frac{h_i}{2}}\left(\frac{\alpha_i}{2h_i} - u_{ML,i}\right)$, $x_2 = \sqrt{\frac{h_i}{2}}\left(\frac{\alpha_i}{2h_i} + u_{ML,i}\right)$. Тогда можно показать , что

$$f_j^L(h_j, u_{ML,j}, \alpha_j) = \frac{\alpha_j}{4}\sqrt{\frac{\pi}{2h_j}} \exp\left(-\frac{h_i u_{ML,i}^2}{2}\right) \times \left[\exp\left(x_1^2\right)\operatorname{erfc}(x_1) + \exp\left(x_2^2\right)\operatorname{erfc}(x_2)\right]$$

- Заметим, что при $x_{1,2}$ существенно отличных от нуля, возникают численные трудности с вычислением этих выражений

Возникающие неопределенности

- Действительно, при $x > 27$, выражение

$$\exp(x^2) > 10^{300}$$

  и большинство программ считают его равным $+\infty$
- Аналогичная ситуация с функцией $\operatorname{erfc}(x)$. При $x > 26$ выражение

$$\operatorname{erfc}(x) < 10^{-300},$$

  что является тождественным нулем, например, для MATLAB
- С другой стороны, выражение для интеграла можно преобразовать, представив его в виде произведения бесконечно больших величин на бесконечно малые

Численные хитрости

- Пусть $x_j \gg 0$, $j \in \{1, 2\}$, тогда можно показать, что

$$\operatorname{erfcx}(x_j) = \exp(x_j^2)\operatorname{erfc}(x_j) \approx 1/(\sqrt{\pi}x_j)$$

- При $x_j \ll 0$ объединяем $\exp(-h_i u_{ML,i}^2/2)$ и $\exp(x_j^2)$

$$\exp(-h_i u_{ML,i}^2/2)\exp(x_j^2) = \exp(y_j),$$

  где

$$y_{1,2} = \frac{\alpha_i^2}{8h_i} \mp \frac{\alpha_i u_{ML,i}}{2}.$$

Алгоритм 4: Метод релевантных собственных векторов с лапласовским регуляризатором

Вход: Обучающая выборка $\{\boldsymbol{x}_i, t_i\}_{i=1}^n$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $t_i \in \{+1, -1\}$; Матрица обобщенных признаков $\Phi = \{\phi_j(\boldsymbol{x}_i)\}_{i,j=1}^{n,m}$;

Выход: Набор весов $\boldsymbol{w}_{MP}$ для решающего правила $t_*(\boldsymbol{x}) = \text{sign}\left(\sum_{j=1}^m w_{MP,j}\phi_j(\boldsymbol{x})\right)$;

1: Найти $\boldsymbol{w}_{ML} = \arg\max p(\boldsymbol{t}|X, \boldsymbol{w})$;
2: Вычислить гессиан $H = \nabla\nabla \log p(\boldsymbol{t}|X, \boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_{ML}}$;
3: Вычислить собственные вектора и собственные значения гессиана $-H = Q^T \Lambda Q$, $\Lambda = \text{diag}(h_1, \ldots, h_m)$;
4: Вычислить $\boldsymbol{u}_{ML} = Q\boldsymbol{w}_{ML}$;
5: **для** $j = 1, \ldots, m$
6: $\quad \alpha_j^* := \arg\max f_j^L(h_j, u_{ML,j}, \alpha_j)$;
7: Найти $\boldsymbol{u}_{MP} = \arg\max p(\boldsymbol{t}|X, \boldsymbol{w})p(\boldsymbol{u}|\boldsymbol{\alpha}^*)$ при условии $u_{ML,i}u_i \geq 0$;
8: Найти $\boldsymbol{w}_{MP} = Q^T\boldsymbol{u}_{MP}$

Оптимизация функции $f_i^L(h_i, u_{ML,i}, \alpha_i)$

Точка максимума функции $f_i^L(h_i, u_{ML,i}, \alpha_i)$ не может быть выписана в явном виде, поэтому необходима численная оптимизация (впрочем, не слишком обременительная, т.к. функция является унимодальной (см. рис. 9.3), либо наибольшее значение достигается при $\alpha_i = +\infty$)

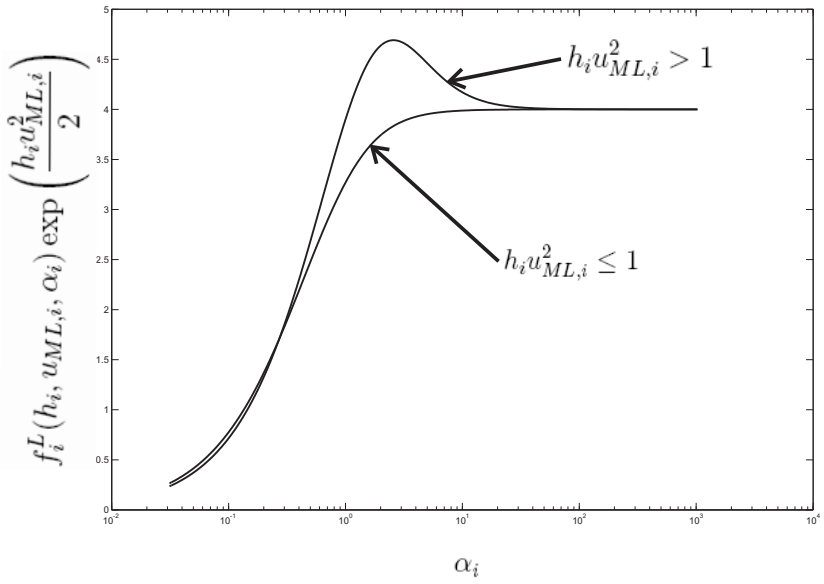


Рис. 9.3. Поведение функции $f_i^L(h_i, u_{ML,i}, \alpha_i)$ при разных значениях α_i

Глава 10

Общее решение для недиагональной регуляризации

В главе представлена схема получения наиболее обоснованного регуляризатора для обобщенных линейных моделей классификации и произвольной неотрицательной матрицей регуляризации. Подробное внимание уделено математическим преобразованиям, позволяющим свести сложную задачу условной матричной оптимизации к простому виду. Также в главе приводятся правила дифференцирования по матрице и по вектору.

10.1 Ликбез: Дифференцирование по вектору и по матрице

Дифференцирование по вектору

- Пусть $f(\boldsymbol{x})$ — некоторая скалярная функция, зависящая от вектора $\boldsymbol{x} \in \mathbb{R}^n$. Тогда ее производная по вектору по определению есть

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left(\frac{\partial f(\boldsymbol{x})}{\partial x_1}, \dots, \frac{\partial f(\boldsymbol{x})}{\partial x_n} \right) = \nabla f(\boldsymbol{x})$$

- Пусть $\boldsymbol{f}(x) = (f_1(x), \dots, f_m(x))^T$ — некоторая векторная функция от скалярной переменной $x \in \mathbb{R}$. Тогда ее производная по аргументу по определению есть

$$\frac{\partial \boldsymbol{f}(x)}{\partial x} = \left(\frac{\partial f_1(x)}{\partial x}, \dots, \frac{\partial f_m(x)}{\partial x} \right)^T$$

- Пусть $\boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \dots, f_m(\boldsymbol{x}))^T$ — некоторая векторная функция, зависящая от вектора $\boldsymbol{x} \in \mathbb{R}^n$. Тогда ее производная по вектору будет матрицей

$$\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left(\frac{\partial f_i(\boldsymbol{x})}{\partial x_j} \right) \in \mathbb{R}^{n \times m}$$

Дифференцирование матриц

- Пусть $A(x) = (a_{ij}(x)) \in \mathbb{R}^{n \times n}$ — квадратная матрица, зависящая от параметра x. Тогда ее производная по параметру по определению равна

$$\frac{\partial A(x)}{\partial x} = \left(\frac{\partial a_{ij}(x)}{\partial x} \right)$$

✓ Упр.

- В частности, выписывая выражения покоординатно можно показать , что

$$\frac{\partial AB}{\partial x} = \frac{\partial A}{\partial x} B + A \frac{\partial B}{\partial x}$$

$$\frac{\partial A^{-1}}{\partial x} = -A^{-1} \frac{\partial A}{\partial x} A^{-1}$$

$$\frac{\partial \log \det(A)}{\partial x} = \operatorname{tr} \left(A^{-1} \frac{\partial A}{\partial x} \right)$$

Дифференцирование по матрице

- Рассмотрим некоторую скалярную функцию, зависящую от матрицы $f(A)$, $A = (a_{ij}) \in \mathbb{R}^{n \times n}$
- При поиске оптимальной матрицы

$$A_* = \arg\max_A f(A)$$

возникает задача дифференцирования функции по матрице

- Производной функции по матрице назовем матрицу производных по соответствующим элементам A

$$\frac{\partial f(A)}{\partial A} = \left(\frac{\partial f(A)}{\partial a_{ij}} \right) \in \mathbb{R}^{n \times n}$$

Полезные формулы

- Производная следа матрицы
$$\frac{\partial \operatorname{tr}(AB)}{\partial A} = B^T, \quad \frac{\partial \operatorname{tr}(A^T B)}{\partial A} = B$$
- Выведем производную определителя матрицы $\frac{\partial \det(A)}{\partial A}$. Для этого распишем определитель по строке
$$\det(A) = \sum_{i=1}^{n} a_{ij} M_{ij},$$
где $M_{ij} = (-1)^{i+j-1} \det(A^{ij})$ — алгебраическое дополнение, а A^{ij} — матрица, полученная из A путем вычеркивания i-ой строки и j-го столбца. Тогда, учитывая, что M_{ik} не зависит от a_{ij} для любых $k \neq j$, получаем
$$\frac{\partial \det(A)}{\partial a_{ij}} = \frac{\partial \sum_{i=1}^{n} a_{ij} M_{ij}}{\partial a_{ij}} = M_{ij}.$$
Каждый элемент матрицы A^{-1} выражается через алгебраические дополнения матрицы A как $a_{ij}^{-1} = \frac{1}{\det(A)} M_{ji}$, отсюда
$$\frac{\partial \det(A)}{\partial A} = \det(A) A^{-1}$$

10.2 Общее решение для недиагональной регуляризации

10.2.1 Получение выражения для обоснованности с произвольной матрицей регуляризации

Гауссовское априорное распределение на веса классификатора

- Рассмотрим стандартный алгоритм логистической регрессии с произвольным гауссовским регуляризатором $p(\boldsymbol{w}|A) = \frac{\sqrt{\det(A)}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T A \boldsymbol{w}\right)$
$$\boldsymbol{w}_{MP} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{t}|X, \boldsymbol{w}) \mathcal{N}(\boldsymbol{w}|0, A^{-1}) =$$
$$\arg\max_{\boldsymbol{w}} \prod_{i=1}^{n} \frac{1}{1 + \exp(-t_i \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}_i))} \frac{\sqrt{\det(A)}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T A \boldsymbol{w}\right)$$
- Матрица регуляризации находится в результате поиска наиболее обоснованной модели
$$A = \arg\max_{A \in \mathcal{A}} p(\boldsymbol{t}|X, A) = \arg\max_{A \in \mathcal{A}} \int p(\boldsymbol{t}|X, \boldsymbol{w}) p(\boldsymbol{w}|A) d\boldsymbol{w}$$
- Классическая байесовская логистическая регрессия соответствует множеству $\mathcal{A} = \{A | A = \alpha I,\ \alpha \geq 0\}$, а метод релевантных векторов — множеству $\mathcal{A} = \{A | A = \operatorname{diag}(\alpha_1, \ldots, \alpha_m),\ \alpha_j \geq 0\}$

Общая постановка задачи

- Очевидно, что ни байесовская логистическая регрессия, ни метод релевантных векторов не покрывают все возможные гауссовские априорные распределения на множество весов $\boldsymbol{w}$
- Рассмотрим задачу поиска наиболее обоснованного распределения во всем классе нормальных распределений

$$A = \arg\max_{A \in \mathcal{A}} p(\boldsymbol{t}|X, A) = \arg\max_{A \in \mathcal{A}} \int p(\boldsymbol{t}|X, \boldsymbol{w}) p(\boldsymbol{w}|A) d\boldsymbol{w},$$

где $\mathcal{A} = \{A | A^T = A,\ A \geq 0\}$

Приближение Лапласа для правдоподобия

- Используем метод Лапласа для того, чтобы приблизить правдоподобие гауссианой
- Пусть $H = \nabla\nabla - \log p(\boldsymbol{t}|X, \boldsymbol{w})|_{\boldsymbol{w}_{ML}}$ — отрицательный гессиан логарифма правдоподобия, взятый в точке максимума, тогда

$$p(\boldsymbol{t}|X, \boldsymbol{w}) \approx \hat{p}(\boldsymbol{t}|X, \boldsymbol{w}) = p(\boldsymbol{t}|X, \boldsymbol{w}_{ML}) \exp\left(-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{ML})^T H (\boldsymbol{w} - \boldsymbol{w}_{ML})\right)$$

- Обозначим

$$Q(\boldsymbol{w}) = \hat{p}(\boldsymbol{t}|X, \boldsymbol{w}) p(\boldsymbol{w}|A) = \hat{p}(\boldsymbol{t}|X, \boldsymbol{w}) \frac{\sqrt{\det(A)}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T A \boldsymbol{w}\right),$$

✓ Упр. тогда легко показать, что выражение для обоснованности принимает вид

$$E(A) \approx \frac{Q(\boldsymbol{w}_{MP})(2\pi)^{m/2}}{\sqrt{\det(-\nabla\nabla \log Q(\boldsymbol{w})|_{\boldsymbol{w}_{MP}})}}$$

Окончательный вид оптимизируемого функционала

- Для упрощения выкладок, перейдем к рассмотрению логарифма обоснованности, очевидно, что

$$A = \arg\max_{A \in \mathcal{A}} E(A) = \arg\max_{A \in \mathcal{A}} \log E(A)$$

- Выражение для логарифма обоснованности имеет вид

$$\log E(A) \approx \log \hat{p}(\boldsymbol{t}|X, \boldsymbol{w}_{MP}) - 0.5 \boldsymbol{w}_{MP}^T A \boldsymbol{w}_{MP} + 0.5 \log\det\left((H + A)^{-1} A\right) + C \to \max_{A \in \mathcal{A}}$$

Задача поиска оптимальной матрицы в классе неотрицательно определенных (semi-definite programming) является нетривиальной и проблема разработки эффективного численного метода решения на настоящий момент является открытой

Компонента $\log\det(A)$ возникает из плотности $p(\boldsymbol{w}|A)$, являющейся множителем $Q(\boldsymbol{w})$, а $\det(H+A)$ — это определитель гессиана $\det(-\nabla\nabla \log Q(\boldsymbol{w})|_{\boldsymbol{w}_{ML}})$

10.2.2 Получение оптимальной матрицы регуляризации в явном виде

Схема последующих выкладок

- Выражение обоснованности через точку максимума правдоподобия
- Выражение обоснованности через промежуточную матрицу $M = H(H + A)^{-1} A$
- Получение явной формулы для M и произвольной симметричной матрицы A
- Получение оптимальной матрицы A с учетом ее неотрицательной определенности

Выражение $\boldsymbol{w}_{MP}$ через $\boldsymbol{w}_{ML}$

- Обоснованность зависит от точки максимума регуляризованного правдоподобия $\boldsymbol{w}_{MP}$, которая **на момент поиска наилучшего регуляризатора неизвестна**
- Учитывая, что $\boldsymbol{w}_{MP}$ зависит от выбранной матрицы регуляризации A, получим явный вид этой зависимости

$$Q(\boldsymbol{w}) = \hat{p}(\boldsymbol{t}|X, \boldsymbol{w}_{ML}) \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_{ML})^T H(\boldsymbol{w}-\boldsymbol{w}_{ML})\right) \frac{\sqrt{\det(A)}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T A\boldsymbol{w}\right)$$

$$\log Q(\boldsymbol{w}) = -0.5\left[(\boldsymbol{w}-\boldsymbol{w}_{ML})^T H(\boldsymbol{w}-\boldsymbol{w}_{ML}) + \boldsymbol{w}^T A\boldsymbol{w} - \log\det(A))\right] + \log\hat{p}(\boldsymbol{t}|X, \boldsymbol{w}_{ML}) - \frac{m}{2}\log(2\pi)$$

$$\frac{\partial \log Q(\boldsymbol{w})}{\partial \boldsymbol{w}} = -H(\boldsymbol{w}-\boldsymbol{w}_{ML}) - A\boldsymbol{w} = -(H+A)\boldsymbol{w} + H\boldsymbol{w}_{ML}$$

- В точке $\boldsymbol{w} = \boldsymbol{w}_{MP}$ производная регуляризованного правдоподобия равна нулю, отсюда

$$\boldsymbol{w}_{MP} = (H+A)^{-1}H\boldsymbol{w}_{ML}.$$

Выражение обоснованности через точку максимума правдоподобия

- Подставим формулу для $\boldsymbol{w}_{MP}$ в выражение для обоснованности

$$\log E(A) = 0.5\log\det((H+A)^{-1}A) - \frac{m}{2}\log(2\pi) + \log\hat{p}(\boldsymbol{t}|X, \boldsymbol{w}_{ML}) -$$

$$-0.5\left[(\boldsymbol{w}_{MP}-\boldsymbol{w}_{ML})^T H(\boldsymbol{w}_{MP}-\boldsymbol{w}_{ML}) + \boldsymbol{w}_{MP}^T A\boldsymbol{w}_{MP}\right]$$

- Учитывая, что матрицы H и $(H+A)$ симметричные, $\boldsymbol{w}_{MP}^T = \boldsymbol{w}_{ML}^T H(H+A)^{-1}$
- Разность $\boldsymbol{w}_{MP} - \boldsymbol{w}_{ML}$ может быть записана в матричном виде

$$\boldsymbol{w}_{MP} - \boldsymbol{w}_{ML} = \left((H+A)^{-1}H - I\right)\boldsymbol{w}_{ML}$$

- Результат подстановки в последнее слагаемое обоснованности

$$-0.5\boldsymbol{w}_{ML}^T\left[\left\{H(H+A)^{-1} - I\right\}H\left\{(H+A)^{-1}H - I\right\} + H(H+A)^{-1}A(H+A)^{-1}H\right]\boldsymbol{w}_{ML} =$$
$$-0.5\boldsymbol{w}_{ML}^T\left[H(H+A)^{-1}H(H+A)^{-1}H - 2H(H+A)^{-1}H + H + H(H+A)^{-1}A(H+A)^{-1}H\right]\boldsymbol{w}_{ML} =$$
$$-0.5\boldsymbol{w}_{ML}^T\left[H(H+A)^{-1}(H+A)(H+A)^{-1}H - -2H(H+A)^{-1}H + H\right]\boldsymbol{w}_{ML} =$$
$$-0.5\boldsymbol{w}_{ML}^T\left[H(H+A)^{-1}H - 2H(H+A)^{-1}H + H\right]\boldsymbol{w}_{ML} = -0.5\boldsymbol{w}_{ML}^T\left[-H(H+A)^{-1}H + H\right]\boldsymbol{w}_{ML}.$$

Матричная хитрость

- Воспользуемся следующим матричным тождеством

$$H - H(H+A)^{-1}H = H(H+A)^{-1}((H+A) - H) = H(H+A)^{-1}A$$

- Тогда выражение для логарифма обоснованности (не забыв добавить $0.5\log\det((H+A)^{-1}A)$) можно переписать

$$\log E(A) = \log\hat{p}(\boldsymbol{t}|X, \boldsymbol{w}_{ML}) - \frac{m}{2}\log(2\pi) +$$

$$0.5[-\boldsymbol{w}_{ML}^T H(H+A)^{-1}A\boldsymbol{w}_{ML} + \log\det\left((H+A)^{-1}A\right)]$$

- Но и в таком виде оптимизация по A крайне затруднительна

Еще одна матричная хитрость

- Сделаем замену переменной $M = H(H + A)^{-1}A$, тогда
$$\log E(A) = 0.5[-\boldsymbol{w}_{ML}^T M \boldsymbol{w}_{ML} + \log\det\left((H + A)^{-1}A\right)] + C$$
- Используя свойство определителя произведения, перепишем второе слагаемое
$$\log\det\left((H + A)^{-1}A\right) = \log\det(M) - \log\det(H)$$
- Учитывая, что H не зависит от A, получаем
$$\log E(A) = 0.5[-\boldsymbol{w}_{ML}^T M \boldsymbol{w}_{ML} + \log\det(M)] + C_1,$$
но такое выражение легко оптимизировать по матрице M!

Выражение для оптимальной матрицы M

✓ Упр.
- Продифференцируем логарифм обоснованности поэлементно по матрице M и приравняем производную к нулю
$$\frac{\partial \log E(A)}{\partial M} = 0.5\left[M^{-1} - \boldsymbol{w}_{ML}\boldsymbol{w}_{ML}^T\right] = 0,$$
- Отсюда получаем выражение для оптимальной матрицы M^{-1}
$$M^{-1} = \boldsymbol{w}_{ML}\boldsymbol{w}_{ML}^T$$
- Матрица M^{-1} имеет ранг 1, т.к. равна произведению двух ненулевых векторов (матриц ранга 1).

Выражение для оптимальной матрицы A

- Получим выражение для матрицы A
$$M = H(H + A)^{-1}A$$
$$A^{-1}(H + A) = M^{-1}H$$
$$A^{-1}H + I = M^{-1}H$$
$$A^{-1} = (M^{-1}H - I)H^{-1} = M^{-1} - H^{-1} = \boldsymbol{w}_{ML}\boldsymbol{w}_{ML}^T - H^{-1}$$

✓ Упр.
- Матрица A^{-1} симметричная
- Матрица $H > 0$, а значит A не является неотрицательной

Неотрицательная матрица регуляризации

- Для того, чтобы получить неотрицательную матрицу, приведем A^{-1} к диагональному виду с помощью ортогонального преобразования
$$D = U^T A^{-1} U = \operatorname{diag}(d_1, d_2 \leq 0, \ldots, d_m \leq 0), \quad U^T = U^{-1}$$
- Все собственные значения A^{-1} кроме, быть может, одного, заведомо неположительные. Заменим их нулями
$$D = \operatorname{diag}(d_1, +0, \ldots, +0)$$
- Тогда $D^{-1} = \operatorname{diag}(d_1^{-1}, +\infty, \ldots, +\infty)$
- Такое преобразование соответствует оптимальной неотрицательной матрице регуляризации с сохранением направлений регуляризации, задаваемых оптимальной матрицей $\boldsymbol{w}_{ML}\boldsymbol{w}_{ML}^T - H^{-1}$

Смысл оптимальной матрицы регуляризации

- У оптимальной матрицы регуляризации $A = UD^{-1}U^T$ все собственные значения, кроме одного, равны бесконечности
- Это означает, что веса $\boldsymbol{w}$ не могут меняться вдоль соответствующих собственных векторов
- Обозначим за $\boldsymbol{u}$ собственный вектор, имеющий конечное собственное значение d_1^{-1}, тогда максимум регуляризованного правдоподобия $\boldsymbol{w}_{MP} = \theta_{MP}\boldsymbol{u}$, где

$$\theta_{MP} = \arg\max_{\theta\in\mathbb{R}} p(\boldsymbol{t}|X, \theta\boldsymbol{u})p(\theta|d_1^{-1}),$$

  здесь $p(\theta|d_1^{-1}) = \frac{1}{\sqrt{2\pi d_1}}\exp\left(-\frac{\theta^2}{2d_1}\right) \sim \mathcal{N}(\theta|0, d_1)$
- Полученный классификатор имеет единственную степень свободы!

Алгоритм 5: «Идеальная» гауссовская регуляризация

Вход: Обучающая выборка $\{\boldsymbol{x}_i, t_i\}_{i=1}^n$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $t_i \in \{+1, -1\}$; Матрица обобщенных признаков $\Phi = \{\phi_j(\boldsymbol{x}_i)\}_{i,j=1}^{n,m}$;

Выход: Набор весов $\boldsymbol{w}_{MP}$ для решающего правила $t_*(\boldsymbol{x}) = \text{sign}\left(\sum_{j=1}^m w_{MP,j}\phi_j(\boldsymbol{x})\right)$;

1: Найти $\boldsymbol{w}_{ML} = \arg\max p(\boldsymbol{t}|X, \boldsymbol{w})$;
2: Вычислить $H = -\nabla\nabla \log p(\boldsymbol{t}|X, \boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_{ML}}$;
3: Вычислить собственные вектора и собственные значения $A = \boldsymbol{w}_{ML}\boldsymbol{w}_{ML}^T - H^{-1} = Q^T D^{-1} Q$, $D = \text{diag}(d_1, \ldots, d_m)$;
4: **если** $\forall j \;\; d_j \leq 0$ **то**
5: $\quad \boldsymbol{w}_{MP} := \boldsymbol{0}$;
6: **иначе**
7: $\quad$ Найти $j_0 : \; d_{j_0} > 0$;
8: $\quad \boldsymbol{u} := \boldsymbol{u}_{j_0}$
9: $\quad$ Найти $\theta_{MP} = \arg\max_{\theta\in\mathbb{R}} p(\boldsymbol{t}|X, \theta\boldsymbol{u})p(\theta|d_{j_0}^{-1})$;
10: $\quad$ Вычислить $\boldsymbol{w}_{MP} = \theta_{MP}\boldsymbol{u}$;

Глава 11

Методы оценки обоснованности

Глава посвящена двум методам оценки обоснованности, которые часто применяются при использовании байесовских методов. Описана идея вариационного подхода, при котором приближенное значение обоснованности получают путем минимизаций дивергенции Кульбака-Лейблера между подынтегральной функцией и ее приближением. Приведен пример использования вариационного метода для задачи построения линейной регрессии. Во второй части главы описаны методы Монте-Карло, позволяющие приближенно вычислять вероятностные интегралы путем генерации выборки из некоторого распределения.

11.1 Ликбез: Дивергенция Кульбака-Лейблера и Гамма-распределение

Дивергенция Кульбака-Лейблера

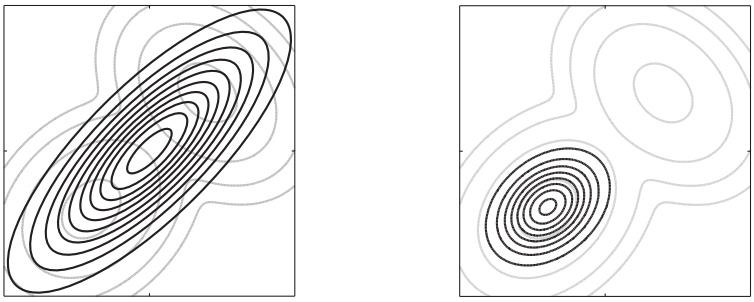
- Существует множество способов определить близость между вероятностными распределениями
- Рассмотрим распределения $p(\boldsymbol{x})$ и $q(\boldsymbol{x})$. Дивергенцией Кульбака-Лейблера называется величина

$$KL(q||p) = -\int q(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}$$

- Заметим, что дивергенция несимметрична

$$KL(q||p) \neq KL(p||q)$$

- Минимизация дивергенции Кульбака-Лейблера часто используется для приближения сложного распределения $p(\boldsymbol{x})$ более простым распределением $q(\boldsymbol{x})$ (см. рис. 11.1)



(a) (b)

Рис. 11.1. На рисунке (a) показан результат минимизации $KL(p||q)$ по $q(\boldsymbol{x})$, а на рисунке (b) — результат минимизации $KL(q||p)$ по $q(\boldsymbol{x})$. В данном случае предполагается, что бимодальное распределение $p(\boldsymbol{x})$ приближается унимодальным распределением $q(\boldsymbol{x})$

Свойства дивергенции Кульбака-Лейблера

- Неотрицательность: $KL(p||q) \geq 0$ для любых двух распределений
- Дивергенция равна нулю тогда и только тогда, когда $q(\boldsymbol{x}) = p(\boldsymbol{x})$
- Антисимметричность: $KL(p||q) \neq KL(q||p)$

Гамма-распределение

- Гамма-распределение имеет плотность

$$\mathcal{G}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda), \quad a,b > 0$$

- Характеристики гамма-распределения

$$\mathbb{E}\lambda = \frac{a}{b}, \qquad \mathbb{D}\lambda = \frac{a}{b^2}$$

- Гамма-распределение является сопряженным для обратной дисперсии (точности) нормального распределения $\lambda = \sigma^{-2}$, т.к.

$$\mathcal{N}(x|\mu,\sigma^2) = \mathcal{N}(x|\mu,\lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right)$$

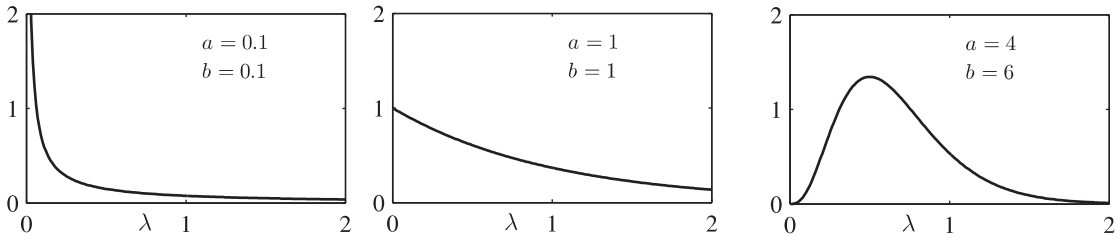


Рис. 11.2. График гамма-распределения с различными параметрами a и b

11.2 Вариационный метод

11.2.1 Идея метода

Недостатки приближения Лапласа

- Метод Лапласа хорошо приближает распределение гауссианой в точке максимума, но плохо делает приближение в целом, если распределение сильно отличается от гауссианы
- В частности, математические ожидания и дисперсии распределения и его приближения Лапласа могут сильно отличаться
- Это приводит к сильным смещениям оценки обоснованности

Приближение апостериорного распределения

- Используем общие обозначения, применявшиеся во второй главе при описании EM-алгоритма. Пусть X — совокупность наблюдаемых переменных, а Z — множество настраиваемых параметров (ненаблюдаемых переменных)

- Вероятностная модель обычно позволяет в явном виде задать совместное распределение $p(X, Z)$. Целью задачи является нахождение (или приближение) обоснованности выбранной модели $p(X) = \int P(X, Z)dZ$ и апостериорного распределения

$$p(Z|X) = \frac{p(X, Z)}{p(X)}$$

- На практике прямое интегрирование выражения $p(X, Z)$ обычно невозможно, поэтому ограничиваются приближением распределения $p(Z|X)$ с помощью некоторого распределения $q(Z)$

Разложение обоснованности

- Справедливо следующее преобразование

$$\log p(X) = \log p(X) \int q(Z)dZ = \int \log p(X)q(Z)dZ =$$
$$\int \log \frac{p(X, Z)}{p(Z|X)} q(Z)dZ = \int \log \frac{p(X, Z)q(Z)}{q(Z)p(Z|X)} q(Z)dZ =$$
$$\int \log \frac{p(X, Z)}{q(Z)} q(Z)dZ - \int \log \frac{p(Z|X)}{q(Z)} q(Z)dZ = \mathcal{L}(q) + KL(q||p)$$

- Величина $\mathcal{L}(q)$ представляет собой нижнюю границу логарифма обоснованности
- Так как $\log p(X)$ не зависит от $q(Z)$, максимизация $\mathcal{L}(q)$ эквивалентна **минимизации дивергенции Кульбака-Лейблера** $KL(q||p)$ между $q(Z)$ и апостериорным распределением $p(Z|X)$!

Факторизация $q(Z)$

- Очевидно, что максимум $\mathcal{L}(q)$ достигается при $q(Z) = p(Z|X)$. В этом случае второе слагаемое оказывается равным нулю
- Прямое вычисление $p(Z|X)$ обычно невозможно, поэтому необходимо ограничить множество $\{q(Z)\}$, в котором проводится поиск наилучшего приближения, например, классом нормальных распределений, и свести задачу к оптимизации соответствующих параметров
- Альтернативой параметрическому ограничению семейства $\{q(Z)\}$ служит его факторизация

$$q(Z) = \prod_{i=1}^{k} q_i(z_i)$$

Факторизованное приближение

- Подставим $q(Z) = \prod_{i=1}^{k} q_i(z_i) = \prod_{i=1}^{k} q_i$ в выражение для $\mathcal{L}(q)$

$$\mathcal{L}(q) = \int \prod_i q_i \left(\log p(X, Z) - \sum_i \log q_i \right) dZ =$$
$$\int q_j \left(\int \log p(X, Z) \prod_{i \neq j} q_i dz_i \right) dz_j - \int q_j \log q_j dz_j + C$$

- Обозначим $\log \tilde{p}(X, z_j) = \mathbb{E}_{i \neq j} \log p(X, Z) = \int \log p(X, Z) \prod_{i \neq j} q_i dz_i$. Тогда

$$\mathcal{L}(q) = \int q_j \log \frac{\tilde{p}(X, z_j)}{q_j} dz_j + C = -KL(q||\tilde{p}) + C$$

Основной результат

- Максимизация $\mathcal{L}(q)$ по q_j эквивалентна минимизации дивергенции между $q_j(z_j)$ и $\tilde{p}(X, z_j)$
- Отсюда оптимальное распределение $q_j^*(z_j) = \tilde{p}(X, z_j)$, т.е.

$$\log q_j^*(z_j) = \mathbb{E}_{i \neq j} \log p(X, Z) + C$$

- Заметим, что нам не пришлось делать каких-либо предположений о функциональной форме распределения $q_j(z_j)$
- Выражение для оптимального $q_j^*(z_j)$ зависит от остальных $q_i(z_i)$, поэтому необходима итерационная оптимизация

11.2.2 Вариационная линейная регрессия

Вероятностная модель линейной регрессии

- Рассмотрим стандартную задачу восстановления регрессии $(X, \boldsymbol{t})$ — обучающая выборка, $t \in \mathbb{R}$. Регрессия имеет вид $y(\boldsymbol{x}) = \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$
- Определим следующую вероятностную модель $p(\boldsymbol{t}, \boldsymbol{w}, \alpha) = p(\boldsymbol{t}|\boldsymbol{w}) p(\boldsymbol{w}|\alpha) p(\alpha)$, где

$$p(\boldsymbol{t}|\boldsymbol{w}) = \prod_{i=1}^{n} \mathcal{N}(t_i|\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1}) \qquad p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1} I)$$

$$p(\alpha) = \mathcal{G}(\alpha|a_0, b_0)$$

- В данной модели роль наблюдаемых переменных играет $\boldsymbol{t}$, а в роли Z выступают $\boldsymbol{w}$ и α
- Для простоты предположим, что значение интенсивности белого шума β известно

Вариационный вывод для α

- Будем искать приближение распредедения $p(\boldsymbol{w}, \alpha|\boldsymbol{t})$ в виде

$$q(\boldsymbol{w}, \alpha) = q(\boldsymbol{w}) q(\alpha)$$

- Используя основной результат для $q(\alpha)$ получаем

$$\log q^*(\alpha) = \mathbb{E}_{\boldsymbol{w}} \log p(\boldsymbol{t}, \boldsymbol{w}, \alpha) = \mathbb{E}_{\boldsymbol{w}} \left(\log p(\boldsymbol{w}|\alpha) p(\alpha)\right) + C = \mathbb{E}_{\boldsymbol{w}} \log p(\boldsymbol{w}|\alpha) + \log p(\alpha) + C =$$
$$\frac{m}{2} \log \alpha - \frac{\alpha}{2} \mathbb{E} \boldsymbol{w}^T \boldsymbol{w} + (a_0 - 1) \log \alpha - b_0 \alpha + C_1$$

- Но это в точности логарифм гамма-распределения с параметрами a_n и b_n, т.е. $\alpha \sim \mathcal{G}(\alpha|a_n, b_n)$, причем

$$a_n = a_0 + \frac{m}{2}, \qquad b_n = b_0 + \frac{1}{2} \mathbb{E} \boldsymbol{w}^T \boldsymbol{w}$$

Вариационный вывод для $\boldsymbol{w}$

- Проделаем аналогичную операцию для $q(\boldsymbol{w})$

$$\log q^*(\boldsymbol{w}) = \mathbb{E}_\alpha \log p(\boldsymbol{t}, \boldsymbol{w}, \alpha) = \mathbb{E}_\alpha \log \left(p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}|\alpha)p(\alpha)\right) = \log p(\boldsymbol{t}|\boldsymbol{w}) + \mathbb{E}_\alpha \log p(\boldsymbol{w}|\alpha) + C =$$
$$-\frac{\beta}{2}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i) - t_i)^2 - \frac{1}{2}\mathbb{E}\alpha \cdot \boldsymbol{w}^T\boldsymbol{w} + C_1 = -\frac{1}{2}\boldsymbol{w}^T(\mathbb{E}\alpha I + \beta\Phi^T\Phi)\boldsymbol{w} + \beta\boldsymbol{w}^T\Phi^T\boldsymbol{t} + C_2,$$

  где $\Phi = (\boldsymbol{\phi}(\boldsymbol{x}_1), \ldots, \boldsymbol{\phi}(\boldsymbol{x}_n))$
- Последовательно проведено отбрасывание слагаемых, не зависящих от $\boldsymbol{w}$, раскрытие скобок и приведение подобных слагаемых
- Выделяя полный квадрат, получаем, что $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_n, S_n)$, где

$$\boldsymbol{\mu}_n = \beta S_n \Phi^T \boldsymbol{t}, \quad S_n = (\mathbb{E}\alpha I + \beta\Phi^T\Phi)^{-1}$$

Итерационные формулы

- Окончательные формулы: $q^*(\alpha) = \mathcal{G}(\alpha|a_n, b_n)$, $q^*(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_n, S_n)$, т.е.

$$\mathbb{E}\alpha = \frac{a_n}{b_n}$$

$$\mathbb{E}\boldsymbol{w}^T\boldsymbol{w} = \operatorname{tr}(\mathbb{E}\boldsymbol{w}\boldsymbol{w}^T) = \operatorname{tr}(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^T + S_n) = \boldsymbol{\mu}_n^T\boldsymbol{\mu}_n + S_n$$

- Параметры распределений определяются по итерационным формулам

$$a_n = a_0 + \frac{m}{2}$$

$$b_n = b_0 + \mathbb{E}\boldsymbol{w}^T\boldsymbol{w} = b_0 + \boldsymbol{\mu}_n^T\boldsymbol{\mu}_n + \operatorname{tr} S_n$$

$$\boldsymbol{\mu}_n = \beta S_n \Phi^T \boldsymbol{t}$$

$$S_n = \left(\mathbb{E}\alpha I + \beta\Phi^T\Phi\right)^{-1} = \left(\frac{a_n}{b_n}I + \beta\Phi^T\Phi\right)^{-1}$$

Заключительные замечания

- Отметим, что никаких ограничений на форму апостериорных распределений не вводилось, а единственным приближением было предположение о факторизации
- Вариационный метод позволяет получать приближение обоснованности, нижней оценкой которой является выражение

$$\mathcal{L}(q) = \int q(\boldsymbol{w}, \alpha) \log \frac{p(\boldsymbol{t}, \boldsymbol{w}, \alpha)}{q(\boldsymbol{w}, \alpha)} d\boldsymbol{w} d\alpha = \mathbb{E} \log p(\boldsymbol{t}, \boldsymbol{w}, \alpha) - \mathbb{E} \log q(\boldsymbol{w}, \alpha) =$$
$$\mathbb{E}_{\boldsymbol{w}} \log p(\boldsymbol{t}|\boldsymbol{w}) + \mathbb{E}_{\boldsymbol{w},\alpha} \log p(\boldsymbol{w}|\alpha) + \mathbb{E}_\alpha \log p(\alpha) - \mathbb{E}_{\boldsymbol{w}} \log q(\boldsymbol{w}) - \mathbb{E}_\alpha q(\alpha)$$

✓ Упр.
- Все эти выражения выписываются в явном виде

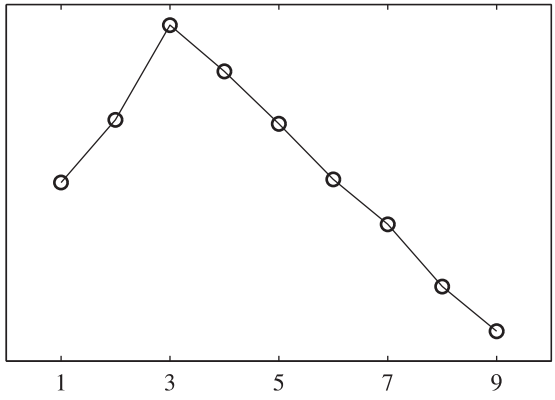


Рис. 11.3. На рисунке изображена зависимость $\mathcal{L}(q)$ от степени полинома для полиномиальной регрессии, построенной по зашумленной выборке, полученной с помощью кубического многочлена

11.3 Методы Монте-Карло

11.3.1 Простейшие методы

Идея метода Монте-Карло

- Метод Монте-Карло применяется для решения задач численного моделирования, в частности взятия интегралов
$$\int_a^b f(x)dx \approx \frac{b-a}{n}\sum_{i=1}^{n} f(x_i) = \hat{f}, \quad x_i \sim U[a,b]$$
- Можно показать, что при весьма общих предположениях $\hat{f} \to \int_a^b f(x)dx$ при $n \to \infty$
- Точность оценки интегралов **не зависит** от размерности пространства d, а определяется исключительно дисперсией самой функции
$$\mathbb{D}\hat{f} = \frac{1}{n}\left[(b-a)\int f^2(x)dx - \left(\int f(x)dx\right)^2\right]$$
- Для численной оценки вероятностных интегралов необходимы специальные методы

Вероятностные интегралы

- В дальнейшем будем рассматривать интегралы вида
$$\mathbb{E}f = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$$
- К ним сводятся многие интегралы, возникающие при байесовском обучении, в частности обоснованность
$$Evidence = \mathbb{E}_{\boldsymbol{w}}p(\boldsymbol{t}|\boldsymbol{w}) = \int p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$
и голосование по апостериорному распределению
$$p(t_{new}|\boldsymbol{t}) = \mathbb{E}_{\boldsymbol{w}}p(t_{new}|\boldsymbol{w}) = \int p(t_{new}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{t})d\boldsymbol{w}$$

Особенности вероятностных интегралов

- Классическая выборка из равномерного распределения для взятия таких интегралов, т.е. формула

$$\int_D f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \approx \frac{|D|}{n}\sum f(\boldsymbol{x}_i)p(\boldsymbol{x}_i), \quad \boldsymbol{x} \sim U(D),$$

  крайне неэффективна, так как в большей части области интегрирования плотность, а, следовательно, и подынтегральная функция близка к нулю
- Для взятия интегралов вида $\int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ нужно уметь проводить выборку из распределения $p(\boldsymbol{x})$
- В этом случае интеграл может быть оценен конечной суммой

$$\int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \approx \frac{1}{n}\sum f(\boldsymbol{x}_i), \quad \boldsymbol{x} \sim p(\boldsymbol{x})$$

Метод обратной функции

- В некоторых случаях можно свести задачу генерации выборки из некоторого распределения к генерации выборки из равномерного распределения
- Пусть $F(x) = P(X < x) = \int_{-\infty}^{x} p(\xi)d\xi$ — функция распределения случайной величины X
- ✓ Упр. Легко показать , что $Y = F(X) \sim U(0,1)$, тогда $X \sim F^{-1}(U(0,1))$
- ✓ Упр. Так удается сгенерировать выборку из показательного распределения и распределения Коши

11.3.2 Схема Метрополиса-Гиббса

Схема с весами

- В дальнейшем полагаем, что нам в каждой точке известна плотность распределения величины с точностью до множителя, т.е.

$$p(\boldsymbol{x}) = \frac{1}{Z_p}\tilde{p}(\boldsymbol{x}),$$

  причем Z_p неизвестна, а $\tilde{p}(\boldsymbol{x})$ может быть легко подсчитана в любой точке
- Введем распределение $q(\boldsymbol{x})$, из которого легко сгенерировать выборку, тогда

$$\mathbb{E}_p f = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \frac{1}{Z_p}\int f(\boldsymbol{x})\frac{\tilde{p}(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x} \approx$$

$$\frac{1}{nZ_p}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{\tilde{p}(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} = \frac{1}{n\sum_{i=1}^{n} r_i}\sum_{i=1}^{n} f(\boldsymbol{x}_i)r_i, \quad \boldsymbol{x} \sim q(\boldsymbol{x})$$

- Если распределение $q(\boldsymbol{x})$ сильно отличается от $p(\boldsymbol{x})$, большинство весов r_i близки к нулю, и метод становится неустойчивым

Марковская цепь

- Методы Монте Карло, использующие Марковские цепи (Monte Carlo Markov chain, MCMC) являются более эффективными средствами получения выборки из заданного распределения
- При использовании MCMC каждая очередная точка выборки $\boldsymbol{x}_i$ зависит некоторым образом от предыдущей точки $\boldsymbol{x}_{i-1}$
- Методы этой группы позволяют «нащупать» области с высоким значением плотности и проводить выборку из них
- Полученная выборка $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ не является выборкой независимых одинаково распределенных случайных величин, но вполне подходит для взятия интеграла

Алгоритм 6: Схема Гиббса

Вход: Многомерное распределение $p(\boldsymbol{x})$;
Выход: Выборка из распределения $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$
1: Инициализация $\boldsymbol{x}_0 = (x_1^0, \ldots, x_d^0)$;
2: **для** $i = 1, \ldots, n$
3: Сгенерировать x_1^i из распределения $p(x_1 | x_2^{i-1}, x_3^{i-1}, \ldots, x_d^{i-1})$;
4: Сгенерировать x_2^i из распределения $p(x_2 | x_1^i, x_3^{i-1}, \ldots, x_d^{i-1})$;
...
5: Сгенерировать x_d^i из распределения $p(x_d | x_2^i, x_3^i, \ldots, x_{d-1}^i)$;
6: $\boldsymbol{x}_i := (x_1^i, \ldots, x_d^i)$;

11.3.3 Гибридный метод Монте-Карло

- Гибридные методы используют информацию не только о значении плотности $p(\boldsymbol{x})$, но и о градиенте ее логарифма $\frac{\partial \log p(\boldsymbol{x})}{\partial \boldsymbol{x}}$
- Для этого используются аналогии с аналитической механикой
  Аналитическая механика была разработана в первой половине 19 в. ирландским математиком Гамильтоном. В ее основе лежит идея замены одного дифференциального уравнения второго порядка во втором законе Ньютона на систему двух дифференциальных уравнений первого порядка
- Считая $\boldsymbol{x}$ переменными состояния, введем потенциальную энергию системы

$$E(\boldsymbol{x}) = -\log p(\boldsymbol{x}) + C$$

- Здесь используется принцип минимальной потенциальной энергии, гласящий, что состояние системы тем более вероятно, чем меньше ее потенциальная энергия

Аналитическая механика

- Введем дополнительные переменные, называемые моментами

$$\boldsymbol{r} = \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t}$$

- Кинетическая энергия системы является функцией моментов $K(\boldsymbol{r}) = 0.5\|\boldsymbol{r}\|^2$, а полная энергия системы (гамильтониан) равна

$$H(\boldsymbol{x}, \boldsymbol{r}) = E(\boldsymbol{x}) + K(\boldsymbol{r})$$

- Уравнения Гамильтона являются записью второго закона Ньютона через переменные состояния и моменты

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \frac{\partial H}{\partial \boldsymbol{r}}$$
$$\frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = -\frac{\partial H}{\partial \boldsymbol{x}}$$

Интегрирование уравнений Гамильтона

- При динамическом изменении замкнутой системы гамильтониан H является постоянным по времени (закон сохранения энергии)
- Изменение системы описывается функциями $\boldsymbol{x}(t)$ и $\boldsymbol{r}(t)$, связанными уравнениями Гамильтона
- При численном решении уравнений получаем

$$\boldsymbol{r}(t + \varepsilon/2) = \boldsymbol{r}(t) - \frac{\varepsilon}{2}\frac{\partial E}{\partial \boldsymbol{x}}(\boldsymbol{x}(t))$$
$$\boldsymbol{x}(t + \varepsilon) = \boldsymbol{x}(t) + \varepsilon \boldsymbol{r}(t + \varepsilon/2)$$
$$\boldsymbol{r}(t + \varepsilon) = \boldsymbol{r}(t + \varepsilon/2) - \frac{\varepsilon}{2}\frac{\partial E}{\partial \boldsymbol{x}}(\boldsymbol{x}(t + \varepsilon))$$

- Полученные решения приблизительно описывают одну из линий уровня функции Гамильтона

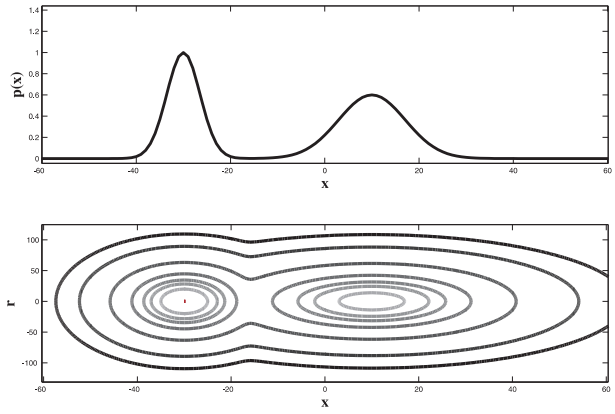


Рис. 11.4. Исходное распределение (вверху) и линии уровня соответствующего ему гамильтониана. Численное решение уравнений Гамильтона приводит к получению последовательности точек, находящихся на одной линии уровня

Схема генерации выборки

- Точки $(\boldsymbol{x}(t_1), \ldots, \boldsymbol{x}(t_n))$ представляют собой равномерную выборку из множества $\{\boldsymbol{x} | p(\boldsymbol{x}) \geq C_0\}$
- Чтобы получить выборку из распределения $p(\boldsymbol{x})$ через каждые $m \ll n$ итераций значение моментов берется из распределения $p(\boldsymbol{r}) = \frac{1}{Z_r}\exp(-K(\boldsymbol{r})) = \mathcal{N}(\boldsymbol{r}|0, I)$
- Такая схема генерации выборки позволяет быстро найти области с большим значением $p(\boldsymbol{x})$ и получить репрезентативную выборку из этих областей

Глава 12

Графические модели. Гауссовские процессы в машинном обучении

В первой части главы описываются графические модели, являющиеся основным средством анализа структурированной информации методами машинного обучения. Кратко описаны понятия условной независимости, ориентированных (байесовские сети) и неориентированных (марковские сети) графических моделей. Вторая часть главы посвящена гауссовским случайным процессам (полям) и их применению для решения задачи восстановления регрессии и классификации. Отдельное внимание уделено автоматическому подбору наиболее обоснованной ковариационной функции случайного поля по конечному множеству наблюдений.

12.1 Ликбез: Случайные процессы и условная независимость

12.1.1 Случайные процессы

Случайные процессы

- Случайным процессом будем называть индексированное множество случайных величин $\xi(\omega) = \{\xi_t(\omega) | t \in T\}$
- Иногда используется нотация $\xi(\omega, t)$
- Первоначально $T \subset \mathbb{R}$, а переменная t ассоциировалась со временем
  Случайный процесс в этом случае удобно представлять как некоторую случайную величину, меняющуюся во времени
- Если $T \subset \mathbb{R}^d$, то случайный процесс обычно называют случайным полем
  Случайный процесс в этом случае удобно представлять как некоторую случайную величину, меняющуюся в пространстве

Двойственная природа случайного процесса

- При фиксированном времени $t = t_0$ процесс представляет собой обычную случайную величину

$$X(\omega) = \xi(\omega, t_0)$$

- При фиксированном элементарном событии $\omega = \omega_0$ процесс представляет собой функцию, называемую **реализацией случайного процесса**

$$f(t) = \xi(\omega_0, t)$$

- Таким образом, случайный процесс обладает как вероятностными, так и функциональными характеристиками
- В частности, можно говорить о математическом ожидании, дисперсии процесса в фиксированный момент времени, а также рассматривать производные и интегралы от реализаций процесса

Вероятностные характеристики случайного процесса

- Среднее значение процесса

$$m(t) = \mathbb{E}\xi(\omega, t)$$

- Ковариационная функция процесса

$$C(t_1, t_2) = \mathrm{Cov}(\xi(\omega, t_1), \xi(\omega, t_2)),$$

  обладающая следующими свойствами

$$C(t, t) = \mathbb{D}\xi(\omega, t) \geq 0, \quad C(t_1, t_2) \leq \sqrt{C(t_1, t_1)C(t_2, t_2)}$$

- Процесс называется стационарным, если его вероятностные характеристики не зависят от времени, в частности

$$C(t, t + \tau) = C(0, \tau) = C(\tau), \quad \forall t$$

Большинство теорем в теории случайных процессов доказано для стационарных процессов

12.1.2 Условная независимость

Условная независимость случайных величин

- Случайные величины x и y называются условно независимыми от z, если

$$p(x, y|z) = p(x|z)p(y|z)$$

- Другими словами вся информация о взаимозависимостях между x и y содержится в z
- Заметим, что из безусловной независимости не следует условная и наоборот
- Основное свойство условно независимых случайных величин

$$p(z|x, y) = \frac{p(x, y|z)p(z)}{p(x, y)} = \frac{p(x|z)p(y|z)p(z)}{p(x, y)} =$$

$$\frac{p(x|z)p(z)p(y|z)p(z)}{p(x, y)p(z)} = \frac{p(z|x)p(z|y)}{p(z)p(x)p(y)p(x, y)} = \frac{1}{Z}\frac{p(z|x)p(z|y)}{p(z)}$$

Пример

- Рассмотрим следующую гипотетическую ситуацию: римские легионы во главе с императором атакуют вторгшихся варваров
- Легионы могут победить варваров, а могут быть разгромлены (Рим в этом случае весьма вероятно будет уничтожен). В свою очередь император может уцелеть, а может погибнуть в сражении
- События «гибель императора» и «уничтожение Рима» не являются независимыми
- Однако, если нам дополнительно известен исход битвы с варварами, эти два события становятся независимыми
- В самом деле, если легионы битву проиграли, то судьба Рима мало зависит от того, был ли император убит в сражении

12.2 Графические модели

12.2.1 Ориентированные графы

Байесовские сети

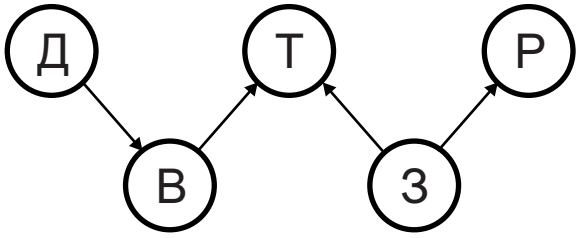


Рис. 12.1. Графическая модель, соответствующая примеру про Джона и колокольчик для воров (см. главу 6)

- Во многих задачах взаимосвязи между наблюдаемыми и скрытыми переменными носят сложный характер

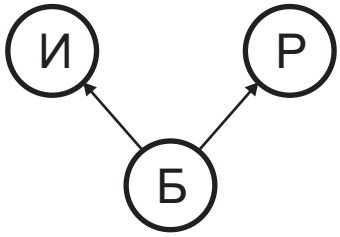


Рис. 12.2. Графическая модель «Варвары и Рим времен заката»

- В частности, между отдельными переменными существуют вероятностные зависимости
- Если удается выделить причинно-следственные связи между переменными, то такие взаимосвязи удобно изображать в виде ориентированных графов
- Ориентированные графы также часто называются байесовскими сетями

Совместное распределение переменных

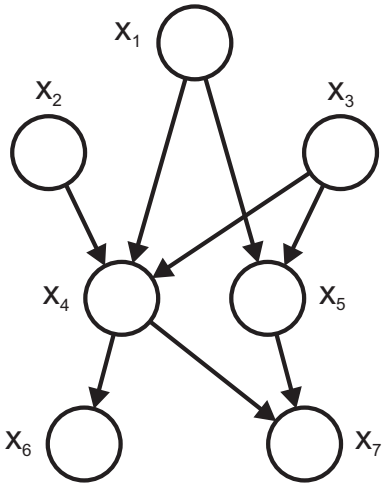


Рис. 12.3.

Рассмотрим графическую модель, изображенную на рис. 12.3. Совместное распределение системы переменных задается выражением

$$p(x_1, x_2, x_3, x_4, x_5, x_6, x_7) =$$
$$p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5).$$

Совместное и условные распределения

- В общем случае, совместное распределение для графа с n вершинами

$$p(\boldsymbol{x}) = \prod_{i=1}^{n} p(x_i|\mathrm{pa}_i),$$

где pa_i — множество вершин-родителей x_i

- Основной задачей, возникающей при работе с графическими моделями, является подсчет условных вероятностей

$$p(\text{unobs}(\boldsymbol{x})|\text{obs}(\boldsymbol{x})),$$

где $\text{obs}(\boldsymbol{x})$ — множество наблюдаемых переменных, а $\text{unobs}(\boldsymbol{x})$ — множество скрытых переменных

- При работе с графическими моделями широко используются sum- и product- rule

Вычисление условных распределений I

- Вернемся к иллюстрации графической модели из семи переменных
- Пусть нам необходимо найти распределение (x_5, x_7) при заданных значениях x_1, x_2, x_4 и неизвестных x_3, x_6 (см. рис. 12.4)

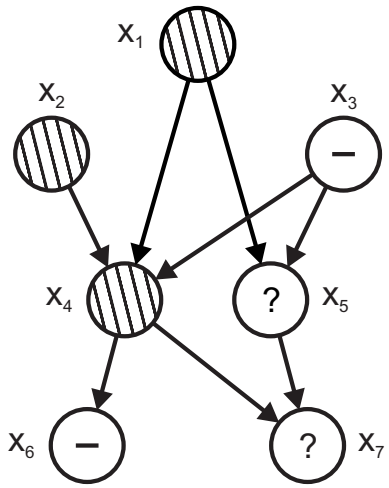


Рис. 12.4.

Вычисление условных распределений II

- По определению условной вероятности

$$p(x_5, x_7|x_1, x_2, x_4) = \frac{p(x_1, x_2, x_4, x_5, x_7)}{p(x_1, x_2, x_4)}$$

- Расписываем знаменатель

$$p(x_1, x_2, x_4) = p(x_1)p(x_2)p(x_4|x_1, x_2) = \{Sum\ rule\}$$

$$p(x_1)p(x_2)\int p(x_4|x_1, x_2, x_3)p(x_3)dx_3$$

- Аналогично числитель

$$p(x_1, x_2, x_4, x_5, x_7) = p(x_1)p(x_2)p(x_4|x_1, x_2)p(x_5|x_1)p(x_7|x_5, x_4) = \\ p(x_1)p(x_2)\left(\int p(x_4|x_1, x_2, x_3)p(x_3)dx_3\right)\left(\int p(x_5|x_1, x_3)p(x_3)dx_3\right)p(x_7|x_5, x_4)$$

- Для взятия возникающих интегралов обычно пользуются методами Монте Карло
- Таким образом, условное распределение выражено через известные атомарные распределения вида $p(x_i|pa_i)$

12.2.2 Три элементарных графа

Граф 1

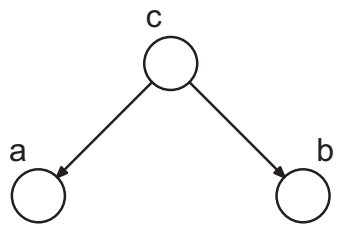


Рис. 12.5.

- Аналогия: Рим, император и варвары
- Переменные a и b условно независимы от c (см. рис. 12.5)
- Возможна маргинализация (исключение переменной)

$$p(a, b) = \int p(a|c)p(b|c)p(c)dc \neq p(a)p(b)$$

Граф 2

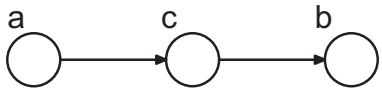


Рис. 12.6.

- Аналогия: данные $\boldsymbol{t}$, параметры алгоритма $\boldsymbol{w}$, параметры модели (гиперпараметры) $\boldsymbol{\alpha}$ в байесовском обучении
- Переменные a и b условно независимы от c (см. рис. 12.6)
- Возможна маргинализация (исключение переменной)

$$p(a, b) = p(a)\int p(b|c)p(c|a)dc \neq p(a)p(b)$$

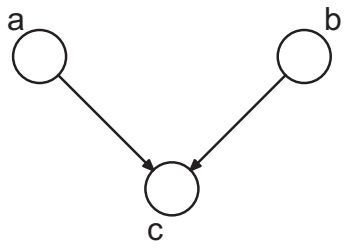


Рис. 12.7.

Граф 3

- Аналогия: Вор, землятрясение и сигнализация
- Переменные a и b независимы, т.е. $p(a,b) = p(a)p(b)$, но не условно независимы (см. рис. 12.7)!
- Зависимость $p(c|a,b)$ не может быть выражена через $p(c|a)$ и $p(c|b)$, хотя обратное верно

$$p(c|a) = \int p(c|a,b)p(b)db$$

12.2.3 Неориентированные графы

Марковские поля

- Неориентированные графические модели также называются Марковскими полями
- Ребра между узлами графа иллюстрируют взаимозависимость между переменными
- Обычно используются для анализа массива данных, имеющего структуру, например сигнала, изображения, сложного объекта

Скрытые марковские поля

- Наиболее известным примером неориентированной графической модели являются скрытые марковские поля, используемые, в частности, для анализа речевых сигналов (рис. 12.8)

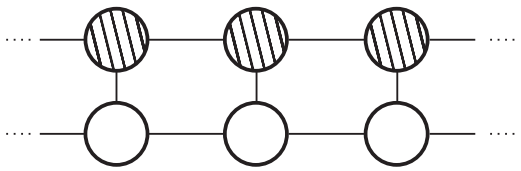
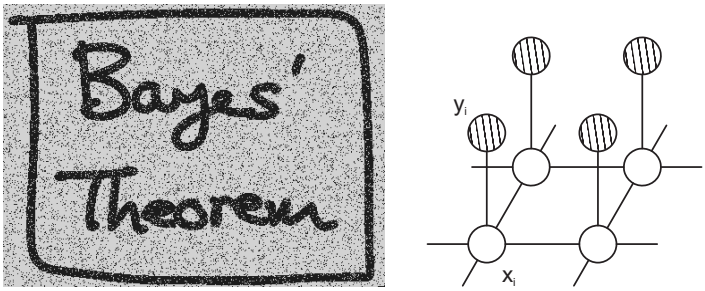


Рис. 12.8.

- Предполагается, что наблюдаемая компонента есть реализация некоторого случайного процесса, характеристики которого являются скрытыми переменными, образующими марковскую цепь

Фильтрация изображений

Примером использования неориентированных графических моделей может служить задача фильтрации изображений (см. рис. 12.9). Выбор между ориентированными и неориентированными графическими моделями зависит от решаемой задачи и определяется исключительно удобством применения, а не какими-то внутренними свойствами исследуемого процесса.



(a) (b)

Рис. 12.9. Соседние пиксели исходного изображения связаны между собой (более вероятно имеют один и тот же цвет). Эту связь можно использовать для фильтрации изображения (рисунок (a)). Соответствующая графическая модель приведена на рисунке (b)

12.3 Гауссовские процессы в машинном обучении

12.3.1 Гауссовские процессы в задачах регрессии

Гауссовские процессы

- Гауссовским процессом называется случайный процесс, все конечномерные распределения которого нормальные

$$p(\xi(\omega, x_1), \ldots, \xi(\omega, x_n)) = \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{\mu}, \Sigma)$$

  В дальнейшем символ ω будем опускать

- Гауссовский процесс является обобщением многомерной гауссианы и полностью задается функцией среднего значения и ковариационной функцией
- Далее будем рассматривать стационарные гауссовские поля $\xi(\boldsymbol{x})$

$$\mu(t) = m, \qquad C(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{y}) = C(\boldsymbol{y})$$

  Если дополнительно известно, что ковариационная функция зависит только от нормы разности $C(\boldsymbol{y}) = C(\|\boldsymbol{y}\|)$, то процесс называют изотропным

Примеры гауссовских процессов

Гауссовские процессы (ГП) являются довольно гибким средством описания данных, а степень «гладкости» процесса определяется видом ковариационной функции (см. рис. 12.10)

Использование случайных полей в задачах восстановления регрессии

- Рассмотрим задачу восстановления регрессии по обучающей выборке $(X, \boldsymbol{t})$, $t \in \mathbb{R}$
- Значения t_i можно интерпретировать как значения реализации случайного процесса (поля) в соответствующей точке $\boldsymbol{x}_i$
- Возникает задача прогноза значения поля t в новой точке $\boldsymbol{x}$ при условии, что в точках обучающей выборки поле имело значения $\boldsymbol{t}$

$$p(\xi(\boldsymbol{x})|\xi(\boldsymbol{x}_1) = t_1, \ldots, \xi(\boldsymbol{x}_n) = t_n) = ?$$

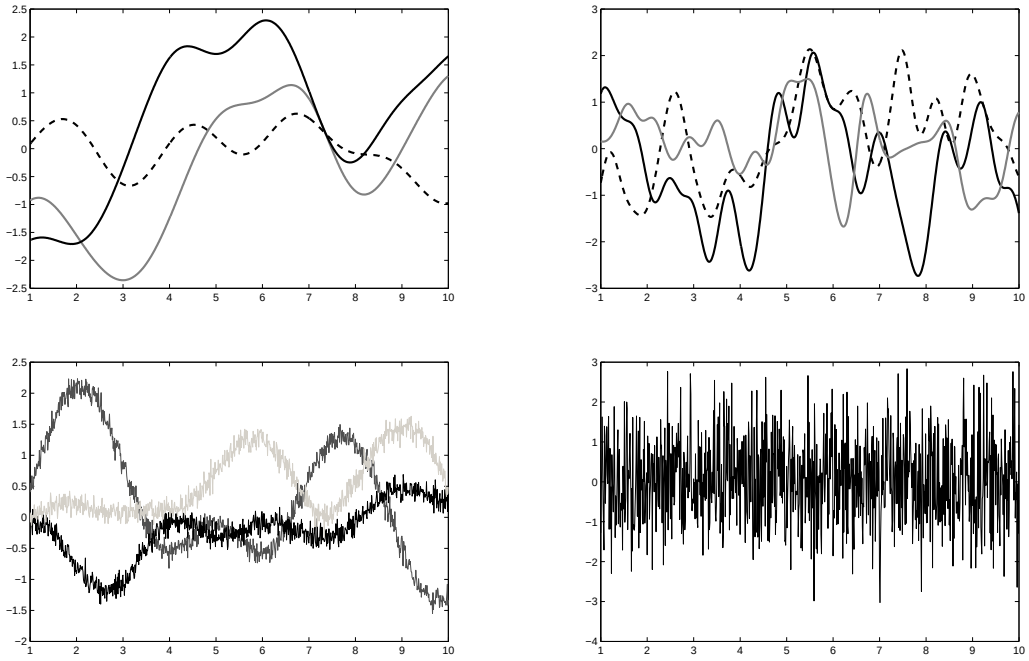


Рис. 12.10. Примеры реализаций стационарных гауссовских случайных процессов с различными ковариационными функциями

Конечномерные распределения поля

- Заметим, что по определению гауссовского случайного процесса (поля)

$$p(\xi(\boldsymbol{x}_1),\ldots,\xi(\boldsymbol{x}_n),\xi(\boldsymbol{x})) = \mathcal{N}((\boldsymbol{\xi},\xi)|\boldsymbol{0},\hat{C}),$$

  где

$$\hat{C} = \begin{pmatrix} C & \boldsymbol{k} \\ \boldsymbol{k}^T & C(\boldsymbol{x},\boldsymbol{x}) \end{pmatrix},$$

$$C = (C(\boldsymbol{x}_i,\boldsymbol{x}_j)), \quad \boldsymbol{k} = (C(\boldsymbol{x}_1,\boldsymbol{x}),\ldots,C(\boldsymbol{x}_n,\boldsymbol{x}))$$

- Также по определению $p(\xi(\boldsymbol{x}_1),\ldots,\xi(\boldsymbol{x}_n)) = \mathcal{N}(\boldsymbol{\xi}|\boldsymbol{0},C)$

Формула Андерсона

- Учитывая, что

$$p(\xi(\boldsymbol{x})|\xi(\boldsymbol{x}_1),\ldots,\xi(\boldsymbol{x}_n)) = \frac{p(\xi(\boldsymbol{x}_1),\ldots,\xi(\boldsymbol{x}_n),\xi(\boldsymbol{x}))}{p(\xi(\boldsymbol{x}_1),\ldots,\xi(\boldsymbol{x}_n))},$$

  легко показать , что

✓ Упр.

$$p(\xi(\boldsymbol{x})|\xi(\boldsymbol{x}_1),\ldots,\xi(\boldsymbol{x}_n)) = \mathcal{N}(\xi|\mu,\sigma^2)$$

- Прогноз поля имеет нормальное распределение с параметрами

$$\mu = \boldsymbol{k}^T C^{-1} \boldsymbol{t}$$

$$\sigma^2 = C(\boldsymbol{x}, \boldsymbol{x}) - \boldsymbol{k}^T C^{-1} \boldsymbol{k} = s^2 - \boldsymbol{k}^T C^{-1} \boldsymbol{k},$$

где $s^2 = \mathbb{D}\xi$ — дисперсия случайного поля

12.3.2 Гауссовские процессы в задачах классификации

Задача классификации

- В задаче классификации ситуация сложнее
- Значение реализации процесса в точках обучающей выборки неизвестно, да и интересует нас лишь знак прогноза, т.е.

$$p(\text{sign}(\xi(\boldsymbol{x}))|\text{sign}(\xi(\boldsymbol{x}_1)) = t_1, \ldots, \text{sign}(\xi(\boldsymbol{x}_n)) = t_n) = ?$$

- Решение заключается в поиске наиболее правдоподобной реализации случайного процесса с учетом информации о знаках

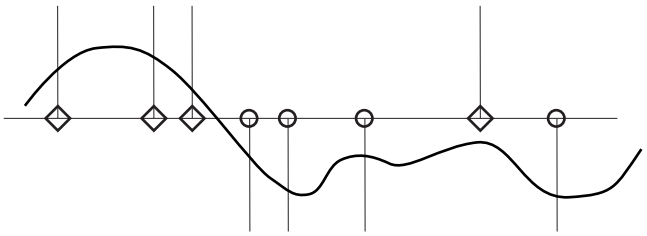


Рис. 12.11. При решении задачи классификации пользователю известен лишь знак реализации процесса в конечном числе точек

ГП классификатор

- Введем правдоподобие метки класса

$$p(\text{sign}(\xi(\boldsymbol{x}))|\xi(\boldsymbol{x})) = \frac{1}{1 + \exp(-\text{sign}(\xi(\boldsymbol{x}))\xi(\boldsymbol{x}))}$$

- Тогда обозначив $\boldsymbol{\xi} = (\xi(\boldsymbol{x}_1), \ldots, \xi(\boldsymbol{x}_n))$, получаем

$$p(\boldsymbol{\xi}|\boldsymbol{t}) \propto p(\boldsymbol{t}|\boldsymbol{\xi}) p(\boldsymbol{\xi}) =$$

$$\prod_{i=1}^{n} \frac{1}{1 + \exp(-t_i \xi(\boldsymbol{x}_i))} \frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp\left(-\frac{1}{2} \boldsymbol{\xi}^T C^{-1} \boldsymbol{\xi}\right)$$

- Отсюда находим

$$\hat{\boldsymbol{\xi}} = \arg\max p(\boldsymbol{\xi}|\boldsymbol{t})$$

Для поиска $\hat{\boldsymbol{\xi}}$ можно воспользоваться методом IRLS (см. лекцию 3)

- Окончательный вид решающего правила для ГП классификатора

$$t_{new} = \text{sign}(\boldsymbol{k} C^{-1} \hat{\boldsymbol{\xi}})$$

12.3.3 Подбор ковариационной функции

Функционал качества для ковариационной функции

- В зависимости от вида ковариационной функции могут быть найдены различные реализации ГП
- *!! Ковариационная функция является структурным параметром ГП!!*
- Запишем правдоподобие ковариационной функции при данной реализации

$$p(\boldsymbol{\xi}|C(\boldsymbol{x},\boldsymbol{y})) = \frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp\left(-\frac{1}{2}\boldsymbol{\xi}^T C^{-1}\boldsymbol{\xi}\right) \to \max_{C_{ij}=C(\boldsymbol{x}_i,\boldsymbol{x}_j)}$$

Заметим, что при этой оптимизации реализация $\boldsymbol{\xi}$ фиксирована

Обоснованность модели ГП

- Популярным параметрическим семейством ковариационных функций является

$$C_{A,\sigma,s}(\boldsymbol{x},\boldsymbol{y}) = A\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2s^2}\right) + \sigma^2 I_{\{\boldsymbol{x}=\boldsymbol{y}\}}$$

- При оптимизации $p(\boldsymbol{\xi}|C(\boldsymbol{x},\boldsymbol{y}))$ происходит поиск ковариационной функции, **наиболее адекватной данной реализации**
- Величина $p(\boldsymbol{\xi}|C(\boldsymbol{x},\boldsymbol{y}))$ является правдоподобием структурных параметров или **обоснованностью модели ГП**

Литература

[1] М. А. Айзерман, Э. М. Браверман, Л. И. Розоноэр *Метод потенциальных функций в теории обучения машин* М.: Наука, 1970

[2] Д. П. Ветров, Д. А. Кропотов *Алгоритмы выбора моделей и синтеза коллективных решений в задачах классификации, основанные на принципе устойчивости* М.: УРСС, 2006

[3] C. M. Bishop *Pattern Recognition and Machine Learning* Springer, 2006

[4] C. Burges. Tutorial on Support Vector Machines Data Mining and Knowledge Discovery, 2, 1998, 121-167.

[5] D. MacKay *Information Theory, Inference, and Learning Algorithms* Cambridge University Press, 2003

[6] V. N. Vapnik *The Nature of Statistical Learning Theory* Springer, 1995

[7] О. С. Середин Методы и алгоритмы беспризнакового распознавания образов *Дисс. к.ф.-м.н.*, Тульский гос. университет, 2001

[8] С. А. Шумский. Байесова регуляризация обучения. сб. Лекции по нейроинформатике, часть 2, 2002

[9] D. Kropotov, D. Vetrov On One Method of Non-Diagonal Regularization in Sparse Bayesian Learning. Proc. of 24th International Conference on Machine Learning (ICML'2007), 2007

[10] D. Kropotov, D. Vetrov. Optimal Bayesian Classifier with Arbitrary Gaussian Regularizer Proc. of 7th Open German-Russian Workshop on Pattern Recognition and Image Understanding (OGRW-7-2007), 2007

[11] M. Tipping. Sparse Bayesian Learning. Journal of Machine Learning Research, 1, 2001, pp. 211-244