Д. П. Ветров, Д. А. Кропотов

Байесовские методы машинного обучения

Учебное пособие

Пособие создано при поддержке программы «Формирование системы инновационного образования в МГУ им. М.В. Ломоносова»

Москва, 2007г

Цели курса

- Ознакомление с классическими методами обработки данных, особенностями их применения на практике и их недостатками
- Представление современных проблем теории машинного обучения
- Введение в байесовские методы машинного обучения
- Изложение последних достижений в области практического использования байесовских методов
- Напоминание основных результатов из смежных дисциплин (теория кодирования, анализ, матричные вычисления, статистика, линейная алгебра, теория вероятностей, случайные процессы)

Структура курса

- 1 семестр, 12 лекций, 24 аудиторных часа + 12 часов для самостоятельной работы
- В каждой лекции секция ликбеза, содержащая краткое напоминание полезных фактов из смежных областей математики
- В конце курса экзамен. Три вопроса в билете, один из секции ликбеза + задача
- Каждая лекция сопровождается показом презентации
- Методические материалы (включая презентации), а также большая часть рекомендуемой литературы доступна на сайте http://mmphome.1gb.ru
- Лекторы: Дмитрий Ветров (VetrovD@yandex.ru) и Дмитрий Кропотов (DKropotov@yandex.ru)

Оглавление

Глава 1

Различные задачи машинного обучения

В главе рассматриваются различные постановки задачи машинного обучения и вводятся основные обозначения, использующиеся в последующих главах. Также приведены основные проблемы, возникающие при обучении ЭВМ.

1.1 Некоторые задачи машинного обучения

Концепция машинного обучения

- Решение задач путем обработки прошлого опыта (case-based reasoning)
- Альтернатива построению математических моделей (model-based reasoning)
- Основное требование – наличие обучающей информации
- Как правило в качестве таковой выступает выборка **прецедентов** – ситуационных примеров из прошлого с известным исходом
- Требуется построить алгоритм, который позволял бы обобщить опыт прошлых наблюдений/ситуаций для обработки новых, не встречавшихся ранее случаев, исход которых неизвестен.

1.1.1 Задача классификации

Классификация

- Исторически возникла из задачи машинного зрения, поэтому часто употребляемый синоним – распознавание образов
- В классической задаче классификации обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором вещественнозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$
- В качестве исхода объекта $\boldsymbol{x}$ фигурирует переменная t, принимающая конечное число значений, обычно из множества $\mathcal{T} = \{1, \ldots, l\}$
- Требуется постросить алгоритм (классификатор), который по вектору признаков $\boldsymbol{x}$ вернул бы метку класса $\hat{t}$ или вектор оценок принадлежности (апотериорных вероятностей) к каждому из классов $\{p(s|\boldsymbol{x})\}_{s=1}^l$ (см. рис. 1.1)

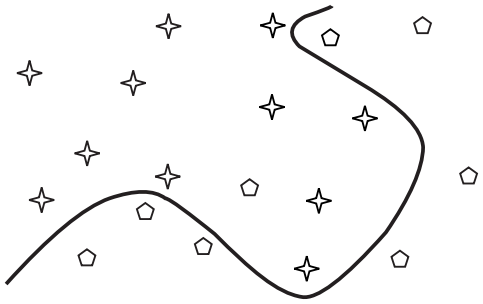


Рис. 1.1. Пример двухклассовой задачи классификации. Звездочками обозначены объекты из одного класса, пятиугольниками — объекты из другого класса. Черная линия соответствует разделяющей поверхности, обеспечивающей качественную классификацию новых объектов

Примеры задач классификации

- Медицинская диагностика: по набору медицинских характеристик требуется поставить диагноз
- Геологоразведка: по данным зондирования почв определить наличие полезных ископаемых

- Оптическое распознавание текстов: по отсканированному изображению текста определить цепочку символов, его формирующих
- Кредитный скоринг: по анкете заемщика принять решение о выдаче/отказе кредита
- Синтез химических соединений: по параметрам химических элементов спрогнозировать свойства получаемого соединения

1.1.2 Задача восстановления регрессии

Регрессия

- Исторически возникла при исследовании влияния одной группы непрерывных случайных величин на другую группу непрерывных случайных величин
- В классической задаче восстановления регрессии обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором вещественнозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$
- В качестве исхода объекта $\boldsymbol{x}$ фигурирует непрерывная вещественнозначная переменная t
- Требуется постросить алгоритм (регрессор), который по вектору признаков $\boldsymbol{x}$ вернул бы точечную оценку значения регрессии $\hat{t}$, доверительный интервал (t_-, t_+) или апостериорное распределение на множестве значений регрессионной переменной $p(t|\boldsymbol{x})$

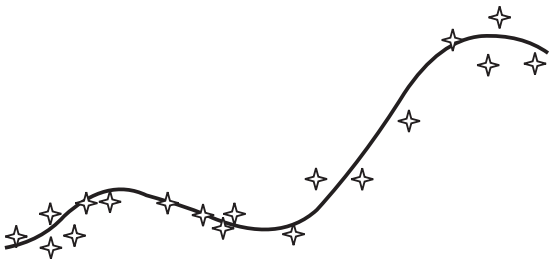


Рис. 1.2. Пример задачи восстановления регрессии. Звездочками обозначены прецеденты, черная линия показывает пример восстанавливаемой функции регрессии

Примеры задач восстановления регрессии

- Оценка стоимости недвижимости: по характеристике района, экологической обстановке, транспортной связности оценить стоимость жилья
- Прогноз свойств соединений: по параметрам химических элементов спрогнозировать температуру плавления, электропроводность, теплоемкость получаемого соединения
- Медицина: по постоперационным показателям оценить время заживления органа
- Кредитный скоринг: по анкете заемщика оценить величину кредитного лимита
- Инженерное дело: по техническим характеристикам автомобиля и режиму езды спрогнозировать расход топлива

1.1.3 Задача кластеризации (обучения без учителя)

Кластеризация

- Исторически возникла из задачи группировки схожих объектов в единую структуру (кластер) с последующим выявлением общих черт
- В классической задаче кластеризации обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором вещественнозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$
- Требуется постросить алгоритм (кластеризатор), который разбил бы выборку на непересекающиеся группы (кластеры) $X = \bigcup_{j=1}^k C_k$, $C_j \subset \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$, $C_i \cap C_j = \emptyset$
- В каждый класс должны попасть объекты в некотором смысле похожие друг на друга (см. рис. 1.3)

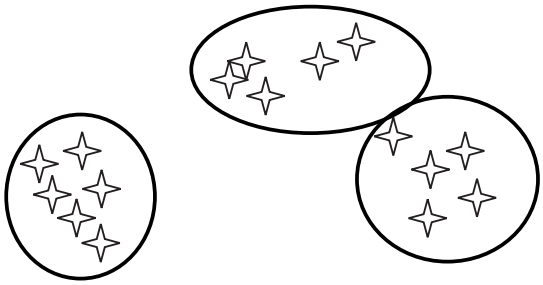


Рис. 1.3. Пример задачи кластеризации. Звездочками обозначены прецеденты. Группы объектов, обведенные кружками, образуют отдельные кластеры

Примеры задач кластерного анализа

- Экономическая география: по физико-географическим и экономическим показателям разбить страны мира на группы схожих по экономическому положению государств
- Финансовая сфера: по сводкам банковских операций выявить группы «подозрительных», нетипичных банков, сгуппировать остальные по степени близости проводимой стратегии
- Маркетинг: по результатам маркетинговых исследований среди множества потребителей выделить характерные группы по степени интереса к продвигаемому продукту
- Социология: по результатам социологических опросов выявить группы общественных проблем, вызывающих схожую реакцию у общества, а также характерные фокус-группы населения

1.1.4 Задача идентификации

Идентификация

- Исторически возникла из классификации, необходимости отделить объекты, обладающие определенным свойством, от «всего остального»
- В классической задаче идентификации обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором вещественнозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$, обладающих некоторым свойством $\chi_A(\boldsymbol{x}) = 1$
- Особенностью задачи является то, что все объекты принадлежат одному классу, причем не существует возможности сделать репрезентативную выборку из класса «все остальное»

- Требуется постросить алгоритм (идентификатор), который по вектору признаков $\boldsymbol{x}$ определил бы наличие свойства A у объекта $\boldsymbol{x}$, либо вернул оценку степени его выраженности $p(\chi_A(\boldsymbol{x}) = 1|\boldsymbol{x})$ (см. рис. 1.4)

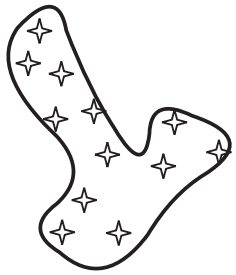


Рис. 1.4. Пример задачи идентификации. Объекты, обладающие определенным свойством, отделены от всех остальных объектов

Примеры задач идентификации

- Медицинская диагностика: по набору медицинских характеристик требуется установить наличие/отсутствие конкретного заболевания
- Системы безопасности: по камерам наблюдения в подъезде идентифицировать жильца дома
- Банковское дело: определить подлинность подписи на чеке
- Обработка изображений: выделить участки с изображениями лиц на фотографии
- Искусствоведение: по характеристикам произведения (картины, музыки, текста) определить, является ли его автором тот или иной автор

1.1.5 Задача прогнозирования

Прогнозирование

- Исторически возникла при исследовании временных рядов и попытке предсказания их значений через какой-то промежуток времени
- В классической задаче прогнозирования обучающая выборка представляет собой набор измерений $X = \{\boldsymbol{x}[i]\}_{i=1}^{n}$, представляющих собой вектор вещественнозначных величин $\boldsymbol{x}[i] = (x_1[i], \ldots, x_d[i])$, сделанных в определенные моменты времени
- Требуется постросить алгоритм (предиктор), который вернул бы точечную оценку $\{\hat{\boldsymbol{x}}[i]\}_{i=n+1}^{n+q}$, доверительный интервал $\{(\boldsymbol{x}_-[i], \boldsymbol{x}_+[i])\}_{i=n+1}^{n+q}$ или апостериорное распределение $p(\boldsymbol{x}[n+1], \ldots, \boldsymbol{x}[n+q]|\boldsymbol{x}[1], \ldots, \boldsymbol{x}[n])$ прогноза на заданную глубину q (см. рис. 1.5)
- В отличие от задачи восстановления регрессии, здесь осуществляется прогноз **по времени**, а не по признакам

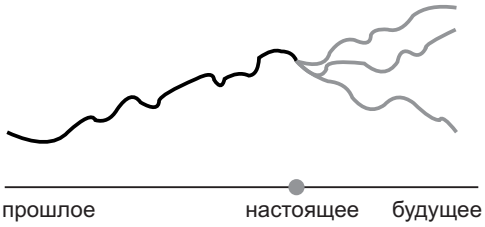


Рис. 1.5. Пример задачи прогнозирования. Черная кривая представляет собой предысторию (известное поведение характеристики). Серые кривые показывают различные варианты прогнозирования поведения характеристики в будущем

Примеры задач прогнозирования

- Биржевое дело: прогнозирование биржевых индексов и котировок
- Системы управления: прогноз показателей работы реактора по данным телеметрии
- Экономика: прогноз цен на недвижимость
- Демография: прогноз изменения численности различных социальных групп в конкретном ареале
- Гидрометеорология: прогноз геомагнитной активности

1.1.6 Задача извлечения знаний

Извлечение знаний

- Исторически возникла при исследовании взаимозависимостей между косвенными показателями одного и того же явления
- В классической задаче извлечения знаний обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором вещественнозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, \boldsymbol{x}_{i,d})$
- Требуется постросить алгоритм, генерирующий набор объективных закономерностей между признаками, имеющих место в генеральной совокупности (см. рис. 1.6)
- Закономерности обычно имеют форму предикатов «ЕСЛИ ... ТО ...» и могут выражаться как в цифровых терминах $((0.45 \leq x_4 \leq 32.1)\&(-6.98 \leq x_7 \leq -6.59) \Rightarrow (3.21 \leq x_2 \leq 3.345))$, так и в текстовых («ЕСЛИ Давление – низкое И (Реакция – слабая ИЛИ Реакция – отсутствует) ТО Пульс – нитевидный»)

Примеры задач извлечения знаний

- Медицина: поиск взаимосвязей (синдромов) между различными показателями при фиксированной болезни
- Социология: определение факторов, влияющих на победу на выборах
- Генная инженерия: выявление связанных участков генома
- Научные исследования: получение новых знаний об исследуемом процессе
- Биржевое дело: определение закономерностей между различными биржевыми показателями

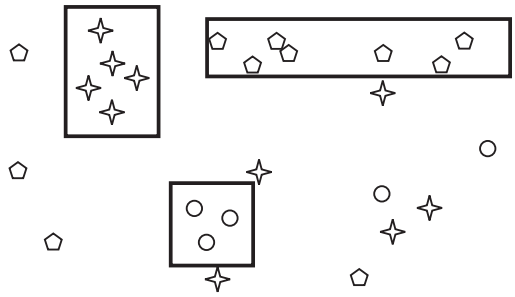


Рис. 1.6. Пример задачи извлечения знаний. В выборке представлены объекты из трех классов. Закономерности представляют собой области признакового пространства, в которых концентрация объектов из одного класса существенно превалирует над концентрациями объектов из других классов

1.2 Основные проблемы машинного обучения

1.2.1 Малый объем обучающей выборки

Объем выборки I

- Основным объектом работы любого метода машинного обучения служит обучающая выборка
- Большой объем выборки позволяет
  - Получить более надежные результаты
  - Использовать более сложные модели алгоритмов
  - Оценить точность обучения
  - **НО:** Время обучения быстро растет
- При малых выборках
  - Можно использовать **только** простые модели алгоритмов
  - Скорость обучения максимальна – можно использовать методы, требующие много времени на обучение
  - Высока вероятность переобучения при ошибке в выборе модели

Объем выборки II

- Одна и та же выборка может являться большой для простых моделей алгоритмов и малой для сложных моделей.
- Для методов с т.н. бесконечной емкостью Вапника-Червоненкиса **любая** выборка является малой.
- С ростом числа признаков увеличивается количество объектов, необходимое для корректного анализа данных
- Часто рассматривается т.н. эффективная размерность выборки $\frac{n}{d}$
- При объемах данных порядка десятков и сотен тысяч встает проблема уменьшения выборки с сохранением ее репрезентативности (active learning)

1.2.2 Некорректность входных данных

Неполнота признакового описания

- Отдельные признаки могут отсутствовать у некоторых объектов. Это может быть связано с отсутствием данных об измерении данного признака для данного объекта, а может быть связано с принципиальным отсутствием данного свойства у данного объекта
- Такое часто встречается в медицинских и химических данных
- Необходимы специальные процедуры, позволяющие корректно обрабатывать пропуски в данных
- Одним из возможных способов такой обработки является замена пропусков на среднее по выборке значение данного признака
- По возможности, пропуски следует игнорировать и исключать из рассмотрения при анализе соответствующего объекта

Противоречивость данных

- Объекты с одним и тем же признаковым описанием могут иметь разные исходы (принадлежать к разным классам, иметь отличные значения регрессионной переменной и т.п.)
- Многие методы машинного обучения не могут работать с такими наборами данных
- Необходимо заранее исключать или корректировать противоречащие объекты
- Использование вероятностных методов обучения позволяет корректно обрабатывать противоречивые данные
- При таком подходе предполагается, что исход t для каждого признакового описания $\boldsymbol{x}$ есть случайная величина, имеющая некоторое условное распределение $p(t|\boldsymbol{x})$

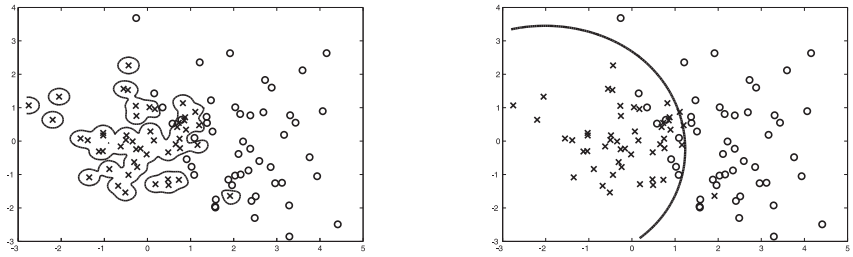
Разнородность признаков

- Хотя формально предполагается, что признаки являются вещественнозначными, они могут быть дискретными и номинальными
- Номинальные признаки отличаются особенностями метрики между значениями
- Стандартная практика состоит в замене номинальных признаков на набор бинарных переменных по числу значений номинального признака
- Текстовые признаки, признаки-изображения, даты и пр. необходимо заменить на соответствующие номинальные либо числовые значения

1.2.3 Переобучение

Идея машинного обучения

- Задача машинного обучения заключается в восстановлении зависимостей по конечным выборкам данных (прецедентов)
- Пусть $(X, \boldsymbol{t}) = (\boldsymbol{x}_i, t_i)_{i=1}^n$ – обучающая выборка, где $\boldsymbol{x}_i \in \mathbb{R}^d$ – признаковое описание объекта, а $t \in \mathcal{T}$ – значение скрытой компоненты (классовая принадлежность, значение прогноза, номер кластера и т.д.)
- При статистическом подходе к решению задачи МО предполагается, что **обучающая выборка является выборкой из некоторой генеральной совокупности** с плотностью $p(\boldsymbol{x}, t)$
- Требуется восстановить $p(t|\boldsymbol{x})$, т.е. знание о скрытой компоненте объекта по измеренным признакам



(a) (b)

Рис. 1.7. Пример двухклассовой задачи классификации. На рисунке (a) представлено решающее правило, которое способно объяснить только объекты обучающей выборки. Решающее правило на рисунке (b) улавливает общую тенденцию в данных и обладает более высокой обобщающей способностью, чем решающее правило (a)

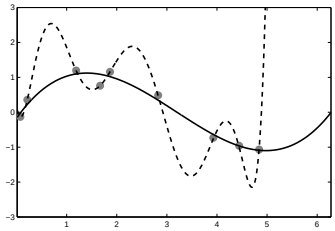


Рис. 1.8. Пример задачи восстановления регрессии. Пунктирная функция регрессии в точности предсказывает объекты обучающей выборки, однако обладает слабой экстраполирующей способностью. Функция регрессии, представленная черной линией, не так точно объясняет объекты обучения, однако хорошо улавливает общую тенденцию в данных

Проблема переобучения

Прямая минимизация невязки на обучающей выборке ведет к получению решающих правил, способных объяснить все что угодно и найти закономерности даже там, где их нет (см. рис. 1.7 и 1.8).

Способы оценки и увеличения обобщающей способности

- На сегодняшний день единственным универсальным способом оценивания обобщающей способности является кросс-валидация
- Все попытки предложить что-нибудь отличное от метода проб и ошибок пока **не привели к общепризнанному решению**. Наиболее известны из них следующие:
  - Структурная минимизация риска (В. Вапник, А. Червоненкис, 1974)
  - Минимизация длины описания (Дж. Риссанен, 1978)
  - Информационные критерии Акаике и Байеса-Шварца (Акаике, 1974, Шварц, 1978)
  - Максимизация обоснованности (МакКай, 1992)
- Последний принцип позволяет надеяться на конструктивное решение задачи выбора модели

Примеры задач выбора модели

- Определение числа кластеров в данных

- Выбор коэффициента регуляризации в задаче машинного обучения (например, коэффициента затухания весов (weight decay) в нейронных сетях)
- Установка степени полинома при интерполяции сплайнами
- Выбор наилучшей ядровой функции в методе опорных векторов (SVM)
- Определение количества ветвей в решающем дереве
- и многое другое...

1.3 Ликбез: Основные понятия мат. статистики

Краткое напоминание основных вероятностных понятий

- $X : \Omega \to \mathbb{R}$ – случайная величина
- Вероятность попадания величины в интервал (a, b) равна

$$P(a \leq X \leq b) = \int_a^b p(x)dx,$$

где $p(x)$ – плотность распределения X,

$$p(x) \geq 0, \quad \int_{-\infty}^{\infty} p(x)dx = 1$$

- Если поведение случайной величины определяется некоторым параметром, возникают условные плотности $p(x|\theta)$. Если рассматривать условную плотность как функцию от параметра

$$f(\theta) = p(x|\theta),$$

то принято говорить о т.н. функции правдоподобия

Основная задача мат. статистики

- Распределение случайной величины X известно с точностью до параметра θ
- Имеется выборка значений величины X, $\boldsymbol{x} = (x_1, \ldots, x_n)$
- Требуется оценить значение θ
- Метод максимального правдоподобия

$$\hat{\theta}_{ML} = \arg\max f(\theta) = \arg\max p(\boldsymbol{x}|\theta) = \arg\max \prod_{i=1}^{n} p(x_i|\theta)$$

- Можно показать, что ММП является ассимптотически оптимальным при $n \to \infty$
- Увы, мир несовершенен. Величина n конечна и обычно не слишком велика
- Необходима регуляризация метода

Пример некорректного использования метода максимального правдоподобия

- $X \sim w_1\mathcal{N}(x|\mu_1, \sigma_1^2) + \cdots + w_m\mathcal{N}(x|\mu_m, \sigma_m^2)$
- Необходимо определить $\boldsymbol{\theta} = (m, \mu_1, \sigma_1^2, \ldots, \mu_m, \sigma_m^2, w_1, \ldots, w_m)$
- Применяем ММП

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \quad = \quad \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) \quad = \quad \prod_{i=1}^{n}\sum_{j=1}^{m} \frac{w_j}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{\|x_i - \mu_j\|^2}{2\sigma_j^2}\right) \quad \to \quad \max_{\boldsymbol{\theta}}$$

- Решение

$$\hat{m}_{ML} = n, \quad \hat{\mu}_{j,ML} = x_j, \quad \hat{\sigma}^2_{j,ML} = 0, \quad \hat{w}_{ML,j} = \frac{1}{n}$$

Выводы

- Не все параметры можно настраивать в ходе обучения
- Существуют специальные параметры (будем называть их структурными), которые должны быть зафиксированы до начала обучения
- *!! В данном случае величина m (количество компонент смеси) является структурным параметром!!*
- Основной проблемой машинного обучения является проблема выбора структурных параметров, позволяющих избегать переобучения

Глава 2

Вероятностная постановка задачи распознавания образов

В главе вводится статистическая постановка задачи машинного обучения. Доказывается оптимальность байесовского классификатора. Подробно рассматриваются вопросы восстановления плотности распределения по выборке данных. Описываются несколько простых подходов к непараметрическому оцениванию плотности. В конце главы приведена общая схема EM-алгоритма, позволяющего восстанавливать смеси распределений с помощью введения скрытой (латентной) переменной.

2.1 Ликбез: Нормальное распределение

Нормальное распределение

- Нормальное распределение играет важнейшую роль в математической статистике

$$X \sim \mathcal{N}(x|\mu,\sigma^2) \Leftrightarrow p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
$$\mu = \mathbb{E}X, \quad \sigma^2 = \mathbb{D}X \triangleq \mathbb{E}(X-\mathbb{E}X)^2$$

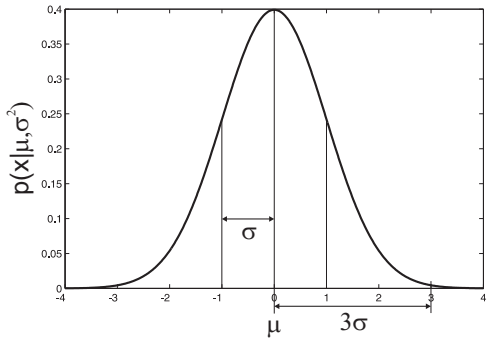


Рис. 2.1. Функция плотности нормального распределения

- Из центральной предельной теоремы следует, что сумма независимых случайных величин с ограниченной дисперсией стремится к нормальному распределению
- На практике, многие случайные величины можно считать приближенно нормальными

Многомерное нормальное распределение

- Многомерное нормальное распределение имеет вид

$$X \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) \Leftrightarrow p(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{\sqrt{2\pi}^n\sqrt{\det\Sigma}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right),$$

  где $\boldsymbol{\mu} = \mathbb{E}X$, $\Sigma = \mathbb{E}(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^T$ — вектор математических ожиданий каждой из n компонент и матрица ковариаций соответственно
- Матрица ковариаций показывает, насколько сильно связаны (коррелируют) компоненты многомерного нормального распределения

$$\Sigma_{ij} = \mathbb{E}(X_i-\mu_i)(X_j-\mu_j) = \mathrm{Cov}(X_i,X_j)$$

- Если мы поделим ковариацию на корень из произведений дисперсий, то получим коэффициент корреляции

$$\rho(X_i,X_j) \triangleq \frac{\mathrm{Cov}(X_i,X_j)}{\sqrt{\mathbb{D}X_i\mathbb{D}X_j}} \in [-1,1]$$

Особенности нормального распределения

- Нормальное распределение **полностью задается** первыми двумя моментами (мат. ожидание и матрица ковариаций/дисперсия)
- Матрица ковариаций неотрицательно определена, причем на диагоналях стоят дисперсии соответствующих компонент
- Нормальное распределение имеет очень легкие хвосты: большие отклонения от мат. ожидания практически невозможны. Это обстоятельство нужно учитывать при приближении произвольных случайных величин нормальными

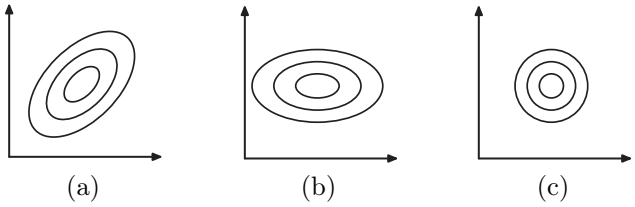


Рис. 2.2. Линии уровня для двухмерного нормального распределения. На рисунке (a) показано нормальное распределение с произвольной матрицей ковариации, случай (b) соответствует диагональной матрице ковариации, а случай (c) соответствует елиничной матрице ковариации, умноженной на некоторую константу

2.2 Статистическая постановка задачи машинного обучения

2.2.1 Вероятностное описание

Основные обозначения

- В дальнейшем будут рассматриваться преимущественно задачи классификации и восстановления регрессии
- В этих задачах обучающая выборка представляет собой набор отдельных объектов $X = \{\boldsymbol{x}_i\}_{i=1}^n$, характеризующихся вектором веществennозначных признаков $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,d})$
- Каждый объект также обладает скрытой переменной $t \in \mathcal{T}$
- Предполагается, что существует зависимость между признаками объекта и значением скрытой переменной
- Для объектов обучающей выборки значение скрытой переменной известно $\boldsymbol{t} = \{t_i\}_{i=1}^n$

Статистическая постановка задачи

- Каждый объект описывается парой $(\boldsymbol{x}, t)$
- При статистической (вероятностной) постановке задачи машинного обучения предполагается, что **обучающая выборка является набором независимых, одинаково распределенных случайных величин, взятых из некоторой генеральной совокупности**
- В этом случае уместно говорить о плотности распределения объектов $p(\boldsymbol{x}, t)$ и использовать вероятностные термины (математическое ожидание, дисперсия, правдоподобие) для описания и решения задачи
- Заметим, что это не единственная возможная постановка задачи машинного обучения

Качество обучения

- Качество обучения определяется точностью прогноза на генеральной совокупности
- Пусть $S(t, \hat{t})$ – функция потерь, определяющая штраф за прогноз $\hat{t}$ при истинном значении скрытой переменной t
- Разумно ожидать, что минимум этой функции достигается при $\hat{t} = t$
- Примерами могут служить $S_r(t, \hat{t}) = (t - \hat{t})^2$ для задачи восстановления регрессии и $S_c(t, \hat{t}) = I\{\hat{t} \neq t\}$ для задачи классификации

Абсолютный критерий качества

- Если бы функция $p(\boldsymbol{x}, t)$ была известна, задачи машинного обучения не существовало
- В самом деле абсолютным критерием качества обучения является мат. ожидание функции потерь, взятое по генеральной совокупности

$$\mathbb{E}S(t, \hat{t}) = \int S(t, \hat{t}(\boldsymbol{x}))p(\boldsymbol{x}, t)d\boldsymbol{x}dt \to \min,$$

  где $\hat{t}(\boldsymbol{x})$ – решающее правило, возвращающее величину прогноза для вектора признаков $\boldsymbol{x}$
- Вместо методов машинного обучения сейчас бы активно развивались методы оптимизации и взятия интегралов от функции потерь
- Так как распределение объектов генеральной совокупности неизвестно, то абсолютный критерий качества обучения не может быть подсчитан

2.2.2 Байесовский классификатор

Идеальный классификатор

- Итак, одна из основных задач теории машинного обучения — это разработка способов косвенного оценивания качества решающего правила и выработка новых критериев для оптимизации в ходе обучения
- Рассмотрим задачу классификации с функцией потерь вида $S_c(t, \hat{t}) = I\{\hat{t} \neq t\}$ и гипотетический классификатор $t_B(\boldsymbol{x}) = \arg\max_{t \in \mathcal{T}} p(\boldsymbol{x}, t)$
- Справделива следующая цепочка неравенств

$$\mathbb{E}S(t, \hat{t}) = \int\int S(t, \hat{t}(\boldsymbol{x}))p(\boldsymbol{x}, t)d\boldsymbol{x}dt =$$

$$\sum_{s=1}^{l} \int S(s, \hat{t}(\boldsymbol{x}))p(\boldsymbol{x}, s)d\boldsymbol{x} = 1 - \int p(\boldsymbol{x}, \hat{t}(\boldsymbol{x}))d\boldsymbol{x} \geq$$

$$\geq 1 - \int \max_t p(\boldsymbol{x}, t)d\boldsymbol{x} = 1 - \int p(\boldsymbol{x}, t_B(\boldsymbol{x}))d\boldsymbol{x} = \mathbb{E}S(t, t_B)$$

Особенности байесовского классификатора

- Таким образом, знание распределения объектов генеральной совокупности приводит к получению оптимального классификатора **в явной форме**
- Такой оптимальный классификатор называется байесовским классификатором
- Если бы удалось с высокой точностью оценить значение плотности генеральной совокупности в каждой точке пространства, задачу классификации можно было бы считать решенной
- На этом основан один из существующих подходов к машинному обучению

2.3 Методы восстановления плотностей

2.3.1 Общие замечания

Идея методов восстановления плотностей

- Восстановление плотностей является примером задачи **непараметрической** статистики
- Основной идеей методов восстановления плотностей является учет количества объектов обучающей выборки, попавших в некоторую окрестность рассматриваемой точки
- Чем больше объектов находится в окрестности рассматриваемой точки, тем выше значение плотности в этой точке
- Окрестность можно зафиксировать либо по ширине, либо по числу объектов, в нее попавших. Возникают два подхода к задаче восстановления плотности

Гистограммы

- Простейшим подходом к восстановлению плотностей является построение гистограмм
- Разбиваем область значений переменной на k областей (обычно прямоугольных) $\Delta_1 \dots \Delta_k$ и считаем сколько точек n_i попало в каждую область Δ_i (см. рис. 2.3). Оценка плотности в этом случае

$$\hat{p}(\boldsymbol{x}) = \frac{n_i}{n|\Delta_i|} I\{\boldsymbol{x} \in \Delta_i\}$$

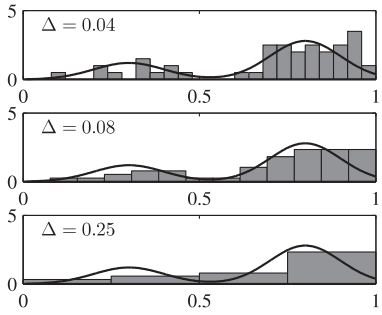


Рис. 2.3. Пример восстановления одномерной плотности с помощью гистограммы для различных значений ширины областей Δ

Недостатки гистограммного подхода

- Получающаяся оценка плотности не является непрерывной функцией
- Оценка зависит от выбора узлов гистограммы (центров областей Δ_i)
- Оценка сильно зависит от выбора ширин областей Δ_i
- *!!Ширина области является структурным параметром!!*
- С ростом размерности пространства d вычислительная сложность возрастает как k^d

2.3.2 Парзеновские окна

Парзеновское оценивание

- Рассмотрим достаточно малую область пространства $\mathcal{D}$, содержащую рассматриваемую точку $\boldsymbol{x}$. Вероятность попадания в эту область равна $P = \int_{\mathcal{D}} p(\boldsymbol{x})d\boldsymbol{x}$. Из n точек обучающей выборки, в эту область попадет k точек, приблизительно равное $k \approx nP$
- Считая, что область $\mathcal{D}$ достаточно мала, и плотность в ней постоянна, получаем $P \approx p(\boldsymbol{x})V$, где V – объем области $\mathcal{D}$, т.е.
$$\hat{p}(\boldsymbol{x}) = \frac{k}{nV}$$
- Отметим две существенные детали
  - Область $\mathcal{D}$ должна быть достаточно широка, чтобы можно было использовать формулу для приближенного оценивания k, т.е. в область $\mathcal{D}$ должно попадать достаточно много точек
  - Область $\mathcal{D}$ должна быть достаточно узка, чтобы значение плотности в ней можно было приблизить константой

Парзеновское окно

- Рассмотрим функцию $k(\boldsymbol{u}) = k(-\boldsymbol{u}) \geq 0$, такую что $\int_{-\infty}^{\infty} k(\boldsymbol{u})d\boldsymbol{u} = 1$. Такая функция называется **парзеновским окном**
- Тогда плотность может быть приближена как $\hat{p}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} k(\boldsymbol{x} - \boldsymbol{x}_i)$
- В самом деле, легко показать, что $\hat{p}(\boldsymbol{x}) \geq 0$ и
$$\int \hat{p}(\boldsymbol{x})d\boldsymbol{x} = \frac{1}{n}\sum_{i=1}^{n}\int k(\boldsymbol{x} - \boldsymbol{x}_i)d\boldsymbol{x} = 1$$
- Обычно в качестве парзеновских окон берут колоколообразные функции с максимумом в нуле, например, $k(\boldsymbol{u}) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\boldsymbol{u}^T\boldsymbol{u}}{2}\right)$
- Парзеновские окна зависят от выбранного масштаба, т.е. характерной ширины окрестности: если $k(\boldsymbol{u})$ окно, то $\frac{1}{h^d}k(\frac{\boldsymbol{u}}{h})$ – тоже окно

2.3.3 Методы ближайшего соседа

Недостатки парзеновского оценивания

- *!! Ширина парзеновского окна h является структурным параметром!!*
- Очевидно, что при фиксированной ширине в некоторые участки пространства будет попадать избыточное количество объектов, и там ширину можно было бы уменьшить и получить более точную оценку плотности
- В то же время, найдутся участки, в которых число объектов, попавших в окно, будет слишком мало, и оценка плотности будет неустойчивой
- Вывод: необходимо сделать ширину окна переменной, потребовав, чтобы в него попадало определенное количество объектов
- Методы этой группы называются методами ближайших соседей

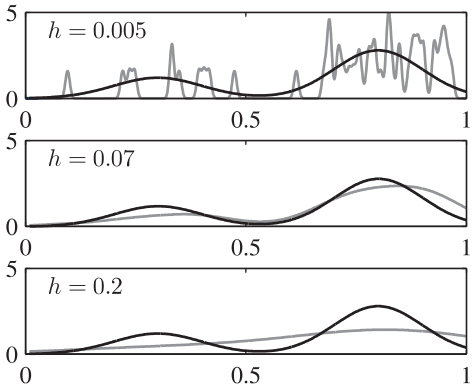


Рис. 2.4. Пример парзеновского оценивания

Идея метода

- Как было получено ранее, значение плотности может быть приближено величиной

$$\hat{p}(\boldsymbol{x}) = \frac{k}{nV},$$

  где V – объем области $\mathcal{D}$, содержащей точку $\boldsymbol{x}$
- Потребуем, чтобы $\mathcal{D}$ являлась сферой минимального радиуса с центром в точке $\boldsymbol{x}$, содержащей ровно k точек обучающей выборки
- ✓ Упр. Можно показать , что для сходимости $\hat{p}(\boldsymbol{x})$ к $p(\boldsymbol{x})$ при $n \to \infty$ необходимыми условиями являются требования

$$\lim_{n\to\infty} k = \infty \qquad \lim_{n\to\infty} \frac{k}{n} = 0$$

- В случае нарушения первого требования оценка плотности будет почти всюду ноль, а при нарушении второго – почти всюду бесконечность, т.к. $V \to 0$ при неограниченном возрастании n

Особенности использования методов ближайших соседей

- *!! Количество ближайших соседей k является структурным параметром!!*
- В большинстве приложений выбор $k \sim \sqrt{n}$ является адекватным
- Серьезным недостатком метода является тот факт, что получающаяся оценка $\hat{p}(\boldsymbol{x})$ не является плотностью, т.к. интеграл от нее, вообще говоря, расходится
- Так в одномерном случае функция $\hat{p}(x)$ имеет хвост порядка $\frac{1}{x}$
- С ростом размерности пространства и объема выборки этот недостаток становится менее критическим

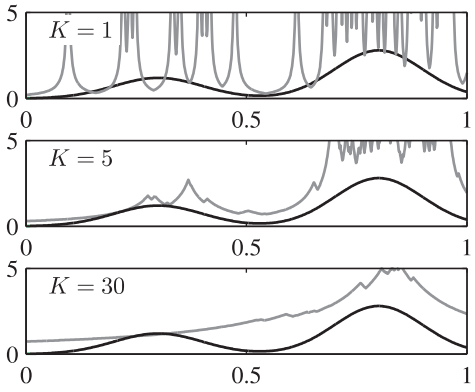


Рис. 2.5. Пример восстановления плотности с помощью метода ближайщего соседа с различным значением K

2.4 EM-алгоритм

2.4.1 Параметрическое восстановление плотностей

Параметрический подход

- При параметрическом восстановлении плотностей предполагается, что искомая плотность нам известна с точностью до параметра $p(\boldsymbol{x}|\theta)$
- Оценка плотности происходит путем оптимизации по θ некоторого функционала
- Обычно в качестве последнего используется правдоподобие или его логарифм

$$p(X|\theta) = \prod_{i=1}^{n} p(\boldsymbol{x}_i|\theta) \to \max_{\theta}$$

- В курсе математической статистики доказывается, что оценки максимального правдоподобия являются состоятельными, ассимптотически несмещенными и эффективными (т.е. имеют наименьшую возможную дисперсию)

Пример использования

- Пусть имееется выборка из нормального распределения $\mathcal{N}(x|\mu, \sigma^2)$ с неизвестными мат. ожиданием и дисперсией
- Выписываем логарифм функции правдоподобия

$$L(X|\mu,\sigma) = -\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} - n\log\sigma - \frac{n}{2}\log(2\pi) \to \max_{\mu,\sigma}$$

$$\frac{\partial L}{\partial \mu} = -\sum_{i=1}^{n} \frac{(x_i-\mu)}{2\sigma^2} = 0 \;\Rightarrow\; \mu_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\frac{\partial L}{\partial \sigma} = \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} - \frac{n}{\sigma} = 0 \;\Rightarrow\; \sigma^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

2.4.2 Задача разделения смеси распределений

Более сложная ситуация

- Часто возникают задачи, в которых генеральная совокупность является смесью нескольких элементарных распределений
- Дискретная смесь: $X \sim \sum_{j=1}^{l} w_j p(\boldsymbol{x}|\theta_j)$, $\sum_{j=1}^{l} w_j = 1$, $w_j \geq 0$
- Непрерывная смесь: $X \sim \int w(\xi) p(\boldsymbol{x}|\theta(\xi)) d\xi$, $\int w(\xi) d\xi = 1$, $w(\xi) \geq 0$
- Даже в случае, когда коэффициенты смеси известны, оптимизация правдоподобия крайне затруднительна
- Существует специальный алгоритм, позволяющий итеративно проводить оценивание коэффициентов смеси и параметров каждого из распределений

Особенности функционала качества

- Основная трудность при применении стандартного метода максимального правдоподобия заключается в необходимости оптимизации по $\boldsymbol{\theta}$ величины

$$L(X|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{l} w_j p(\boldsymbol{x}_i|\boldsymbol{\theta}_j) \right)$$

- Этот функционал имеет вид «логарифм суммы» и крайне сложен для оптимизации
- Необходимо перейти к функционалу вида «сумма логарифмов», который легко оптимизируется в явном виде

Суть EM-алгоритма

- Если бы мы знали, какой объект из какой компоненты смеси взят, т.е. функцию $j(i)$ (будем считать, что соответствующая информация содержится в ненаблюдаемой переменной z), то правдоподобие выглядело бы куда проще

$$L(X, Z|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log w_{j(i)} p(\boldsymbol{x}_i|\boldsymbol{\theta}_{j(i)})$$

- Идея EM-алгоритма заключается в введении **латентных**, т.е. ненаблюдаемых переменных Z, позволяющих упростить вычисление правдоподобия
- Оптимизация проводится итерационно методом покоординатного спуска: на каждой итерации последовательно уточняются возможные значения Z (E-шаг), а потом пересчитываются значения θ (M-шаг)

Схема EM-алгоритма

- На входе: выборка X, зависящая от набора параметров $\boldsymbol{\theta}$, $\boldsymbol{w}$
- Инициализируем $\boldsymbol{\theta}$, $\boldsymbol{w}$ некоторыми начальными приближениями
- E-шаг: Оцениваем распределение скрытой компоненты

$$p(Z|X, \boldsymbol{\theta}, \boldsymbol{w}) = \frac{p(X, Z|\boldsymbol{\theta}, \boldsymbol{w})}{\sum_Z p(X, Z|\boldsymbol{\theta}, \boldsymbol{w})}$$

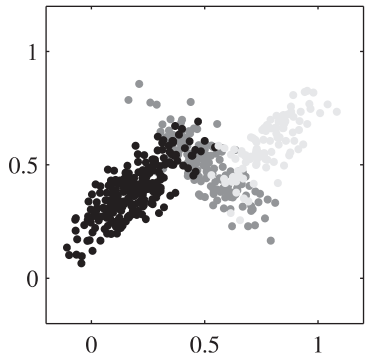


Рис. 2.6. Выборка из смеси трех нормальных распределений

- М-шаг: Оптимизируем

$$\mathbb{E}_Z \log p(X, Z|\boldsymbol{\theta}, \boldsymbol{w}) = \sum_Z p(Z|X, \boldsymbol{\theta}, \boldsymbol{w}) \log p(X, Z|\boldsymbol{\theta}, \boldsymbol{w}) \to \max_{\boldsymbol{\theta}, \boldsymbol{w}}$$

  Если бы мы **точно знали значение** $Z = Z_0$, то вместо мат. ожидания по всевозможным (с учетом наблюдаемых данных) Z, мы бы оптимизировали $\log p(X, Z_0|\boldsymbol{\theta}, \boldsymbol{w})$

- Переход к Е-шагу, пока процесс не сойдется

2.4.3 Разделение гауссовской смеси

Смесь гауссовских распределений

- Имеется выборка $X \sim \sum_{j=1}^{l} w_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \Sigma_j)) \in \mathbb{R}^d$ (см. рис. 2.6)
- Требуется восстановить плотность генеральной совокупности

EM-алгоритм

- Выбираем начальное приближение $\boldsymbol{\mu}_j$, w_j, Σ_j
- Е-шаг: Вычисляем распределение скрытых переменных $\boldsymbol{z}_i \in \{0, 1\}$, $\sum_j z_{ij} = 1$, которые определяют, к какой компоненте смеси принадлежит объект $\boldsymbol{x}_i$

$$\gamma(z_{ij}) = \frac{w_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \Sigma_j))}{\sum_{k=1}^{l} w_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \Sigma_k))}$$

- М-шаг: С учетом новых вероятностей на $\boldsymbol{z_i}$, пересчитываем параметры смеси

$$\boldsymbol{\mu}_j^{new} = \frac{1}{N_j} \sum_{i=1}^{n} \gamma(z_{ij}) \boldsymbol{x}_i, \quad w_j^{new} = \frac{N_j}{n}, \quad N_j = \sum_{i=1}^{n} \gamma(z_{ij})$$

$$\Sigma_j^{new} = \frac{1}{N_j} \sum_{i=1}^{n} \gamma(z_{ij}) (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{new})^T (\boldsymbol{x}_i - \boldsymbol{\mu}_j^{new})$$

- Переход к Е-шагу, пока не будет достигнута сходимость

Недостатки EM-алгоритма

- В зависимости от выбора начального приближения может сходиться к разным точкам
- EM-алгоритм находит локальный экстремум, в котором значение правдоподобия может оказаться намного ниже, чем в глобальном максимуме
- EM-алгоритм **не позволяет определить количество компонентов смеси** l
- *!! Величина* l *является структурным параметром!!*

Глава 3

Обобщенные линейные модели

Данная глава посвящена описанию простейших методов решения задачи восстановления регрессии и классификации. Изложение ведется в контексте т.н. обобщенных линейных моделей. Дается подробное описание метода построения линейной регрессии. Особое внимание уделено вопросам регуляризации задачи. Приводится описание метода наименьших квадратов с итеративным перевзвешиванием, используемого для обучения логистической регрессии

3.1 Ликбез: Псевдообращение матриц и нормальное псевдорешение

Псевдообращение матриц

- Предположим, нам необходимо решить СЛАУ вида $A\boldsymbol{x} = \boldsymbol{b}$
- Если бы матрица A была квадратной и невырожденной (число уравнений равно числу неизвестных и все уравнения линейно независимы), то решение задавалось бы формулой $\boldsymbol{x} = A^{-1}\boldsymbol{b}$
- Предположим, что число уравнений больше числа неизвестных, т.е. матрица A прямоугольная. Домножим обе части уравнения на A^T слева

$$A^T A\boldsymbol{x} = A^T\boldsymbol{b}$$

- В левой части теперь квадратная матрица и ее можно перенести в правую часть

$$\boldsymbol{x} = \left(A^T A\right)^{-1} A^T\boldsymbol{b}$$

- Операция $\left(A^T A\right)^{-1} A^T$ называется псевдообращением матрицы A, а $\boldsymbol{x}$ – псевдорешением

Нормальное псевдорешение

- Если матрица $A^T A$ вырождена, псевдорешений бесконечно много, причем найти их на компьютере нетривиально
- Для решения этой проблемы используется ридж-регуляризация матрицы $A^T A$

$$A^T A + \lambda I,$$

  где I – единичная матрица, а λ – коэффициент регуляризации. Такая матрица невырождена для любых $\lambda > 0$

- Величина

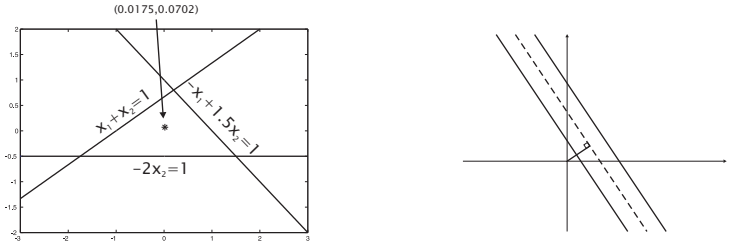
$$\boldsymbol{x} = \left(A^T A + \lambda I\right)^{-1} A^T\boldsymbol{b}$$

  называется нормальным псевдорешением. Оно всегда единственно и при небольших положительных λ определяет псевдорешение с наименьшей нормой

Графическая иллюстрация

- Псевдорешение соответствует точке, минимизирующей невязку, а нормальное псевдорешение отвечает псевдорешению с наименьшей нормой (см. рис. 3.1)

- Заметим, что псевдообратная матрица $\left(A^T A\right)^{-1} A^T$ совпадает с обратной матрицей A^{-1} в случае невырожденных квадратных матриц



(a) (b)

Рис. 3.1. На рисунке (a) показан пример единственного псевдорешения (которое одновременно является и нормальным псевдорешением), на рисунке (b) пунктирная линия обозначает множество всех псевдорешений, а нормальное псевдорешение является основанием перпендикуляра, опущенного из начала координат

3.2 Линейная регрессия

3.2.1 Классическая линейная регрессия

Задача восстановления регрессии

- Задача восстановления регрессии предполагает наличие связи между наблюдаемыми признаками $\boldsymbol{x}$ и непрерывной переменной t
- В отличие от задачи интерполяции допускаются отклонения решающего правила от правильных ответов на объектах обучающей выборки
- Уравнение регрессии $y(\boldsymbol{x}, \boldsymbol{w})$ ищется в некотором параметрическом виде путем нахождения наилучшего значения вектора весов

$$\boldsymbol{w}_* = \arg\max_{\boldsymbol{w}} F(X, \boldsymbol{t}, \boldsymbol{w})$$

Линейная регрессия

- Наиболее простой и изученной является линейная регрессия
- Главная особенность: настраиваемые параметры входят в решающее правило **линейно**
- Заметим, что линейная регрессия не обязана быть линейной по признакам
- Общее уравнение регрессии имеет вид

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

Особенность выбора базисных функций

- Общего метода выбора базисных функций $\phi_j(\boldsymbol{x})$ — не существует
- Обычно они подбираются из априорных соображений (например, если мы пытаемся восстановить какой-то периодический сигнал, разумно взять функции тригонометрического ряда) или путем использования некоторых «универсальных» базисных функций
- Наиболее распространенными базисными функциями являются
  - $\phi(\boldsymbol{x}) = x_k$

  - $\phi(\boldsymbol{x}) = x_{k_1} x_{k_2} \dots x_{k_l}$
  - $\phi(\boldsymbol{x}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{x}_0\|^p),\ \gamma, p > 0.$
- Метод построения линейной регрессии (настройки весов $\boldsymbol{w}$) **не зависит** от выбора базисных функций

Формализация задачи

- Пусть $S(t, \hat{t})$ — функция потерь от ошибки в определении регрессионной переменной t
- Необходимо минимизировать потери от ошибок на генеральной совокупности

$$\mathbb{E}S(t, y(\boldsymbol{x}, \boldsymbol{w})) = \int\int S(t, y(\boldsymbol{x}, \boldsymbol{w}))p(\boldsymbol{x}, t)d\boldsymbol{x}dt \to \min_{\boldsymbol{w}}$$

- Дальнейшие рассуждения зависят от вида функции потерь
- Во многих случаях даже не нужно восстанавливать полностью условное распределение $p(t|\boldsymbol{x})$

Важная теорема

- Теорема. Пусть функция потерь имеет вид
  - $S(t, \hat{t}) = (t - \hat{t})^2$ — «Потери старушки»;
  - $S(t, \hat{t}) = |t - \hat{t}|$ — «Потери олигарха»;
  - $S(t, \hat{t}) = \delta^{-1}(t - \hat{t})$ — «Потери инвалида».

  Тогда величиной, минимизирующей функцию $\mathbb{E}S(t, y(\boldsymbol{x}, \boldsymbol{w}))$, является следующая
  - $y(\boldsymbol{x}) = \mathbb{E}p(t|\boldsymbol{x})$;
  - $y(\boldsymbol{x}) = \text{med}\ p(t|\boldsymbol{x})$;
  - $y(\boldsymbol{x}) = \text{mod}\ p(t|\boldsymbol{x}) = \arg\max_t p(t|\boldsymbol{x})$.
- В зависимости от выбранной системы предпочтений, мы будем пытаться оценивать тот или иной функционал от апостериорного распределения **вместо того, чтобы оценивать его самого**

3.2.2 Метод наименьших квадратов

Минимизация невязки

- Наиболее часто используемой функцией потерь является квадратичная $S(t, \hat{t}) = (t - \hat{t})^2$
- Значение регрессионной функции на обучающей выборке в матричном виде может быть записано как $\boldsymbol{y} = \Phi\boldsymbol{w}$, где $\Phi = (\phi_{ij}) = (\phi_j(\boldsymbol{x}_i)) \in \mathbb{R}^{n \times m}$
- Таким образом, приходим к следующей задаче

$$\|\boldsymbol{y} - \boldsymbol{t}\|^2 = \|\Phi\boldsymbol{w} - \boldsymbol{t}\|^2 \to \min_{\boldsymbol{w}}$$

  Взяв производную по $\boldsymbol{w}$ и приравняв ее к нулю, получаем

$$\frac{\partial \|\Phi\boldsymbol{w} - \boldsymbol{t}\|^2}{\partial \boldsymbol{w}} = \frac{\partial [\boldsymbol{w}^T\Phi^T\Phi\boldsymbol{w} - 2\boldsymbol{w}^T\Phi^T\boldsymbol{t} + \boldsymbol{t}^T\boldsymbol{t}]}{\partial \boldsymbol{w}} = 2\Phi^T\Phi\boldsymbol{w} - 2\Phi^T\boldsymbol{t} = 0$$

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

Регуляризация задачи

- Заметим, что формула для весов линейной регрессии представляет собой псевдорешение уравнения $\Phi \boldsymbol{w} = \boldsymbol{t}$
- ✓ Упр. Матрица $\Phi^T \Phi \in \mathbb{R}^{m \times m}$ вырождена при $m > n$
- Регуляризуя вырожденную матрицу, получаем

$$\boldsymbol{w} = \left(\Phi^T \Phi + \lambda I\right)^{-1} \Phi^T \boldsymbol{t}$$

- Отсюда формула для прогноза объектов обучающей выборки по их правильным значениям

$$\hat{\boldsymbol{t}} = \boldsymbol{y} = \Phi \left(\Phi^T \Phi + \lambda I\right)^{-1} \Phi^T \boldsymbol{t} = H \boldsymbol{t}$$

С историческим обозначением прогноза — навешиванием шляпки связано неформальное название матрицы H, по-английски звучащее как hat-matrix

Особенности квадратичной функции потерь

- Достоинства
  - Квадратичная функция потерь гладкая (непрерывная и дифференцируемая)
  - Решение может быть получено в явном виде
  - Существует простая вероятностная интерпретация прогноза и функции потерь
- Недостатки
  - Решение неустойчиво (не робастно) относительно даже малого количества выбросов. Это связано с быстрым возрастанием квадратичной функции потерь при больших отклонениях от нуля
  - Квадратичная функция неприменима к задачам классификации

3.2.3 Вероятностная постановка задачи

Нормальное распределение ошибок

- Рассмотрим вероятностную постановку задачи восстановления регрессии. Регрессионная переменная t — случайная величина с плотностью распределения $p(t|\boldsymbol{x})$
- В большинстве случаев предполагается, что t распределена нормально относительно некоторого мат. ожидания $y(\boldsymbol{x})$, определяемого точкой $\boldsymbol{x}$

$$t = y(\boldsymbol{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\varepsilon|0, \sigma^2)$$

- Необходимо найти функцию $y(\boldsymbol{x})$, которую мы можем отождествить с уравнением регрессии
- Предположение о нормальном распределении отклонений можно обосновать ссылкой на центральную предельную теорему

Метод максимального правдоподобия для регрессии

- Используем ММП для поиска $y(\boldsymbol{x})$
- Правдоподобие задается следующей формулой

$$p(\boldsymbol{t}|\boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t_i - y_i)^2}{2\sigma^2}\right) \to \max$$

- Взяв логарифм и отбросив члены, не влияющие на положение максимума, получим

$$\sum_{i=1}^{n}(t_i - y_i)^2 = \sum_{i=1}^{n}(t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2 \to \min_{\boldsymbol{w}}$$

- Таким образом, применение метода максимального правдоподобия **в предположении о нормальности отклонений** эквивалентно методу наименьших квадратов

Вероятностный смысл регуляризации

- Теперь будем максимизировать не правдоподобие, а апостериорную вероятность
- По формуле условной вероятности

$$p(\boldsymbol{w}|\boldsymbol{t}, X) = \frac{p(\boldsymbol{t}|X, \boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{t}, X)} \to \max_{\boldsymbol{w}},$$

  знаменатель не зависит от $\boldsymbol{w}$, поэтому им можно пренебречь
- Пусть $p(\boldsymbol{w}) \sim \mathcal{N}\left(\boldsymbol{w}\,\middle|\,\boldsymbol{0}, \left(\frac{\sigma^2}{\lambda}\right) I\right)$. Тогда

$$p(\boldsymbol{w}|\boldsymbol{t}, X) \propto \frac{\lambda^{m/2}}{\left(\sqrt{2\pi}\sigma\right)^{m+n}} \exp\left(-\frac{1}{2}\left(\sigma^{-2}\|\Phi\boldsymbol{w} - \boldsymbol{t}\|^2 + \frac{\lambda}{\sigma^2}\|\boldsymbol{w}\|^2\right)\right)$$

- Логарифмируя и приравнивая производную по $\boldsymbol{w}$ к нулю, получаем

$$\boldsymbol{w} = (\Phi^T\Phi + \lambda I)^{-1}\Phi\boldsymbol{t}$$

- **Регуляризация эквивалентна введению априорного распределения, поощряющего небольшие веса**

3.3 Применение регрессионных методов для задачи классификации

3.3.1 Логистическая регрессия

Особенности задачи классификации

- Рассмотрим задачу классификации на два класса $t \in \{-1, +1\}$
- Ее можно свести к задаче регрессии, например, следующим образом

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

- Возникает вопрос: что использовать в качестве значений регрессионной переменной на этапе обучения?
- Наиболее распространенный подход заключается в использовании значения $+\infty$ для $t = +1$ и $-\infty$ для $t = -1$
- Геометрический смысл: чем дальше от нуля значение $y(\boldsymbol{x})$, тем увереннее мы в классификации объекта $\boldsymbol{x}$

Правдоподобие правильной классификации

- Метод наименьших квадратов, очевидно, неприменим при таком подходе
- Воспользуемся вероятностной постановкой для выписывания функционала качества
- Определим правдоподобие классификации следующим образом

$$p(t|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1 + \exp(-ty(\boldsymbol{x}))}$$

- Это логистическая функция (см. рис. 3.2). Легко показать, что $\sum_t p(t|\boldsymbol{x}, \boldsymbol{w}) = 1$ и $p(t|\boldsymbol{x}, \boldsymbol{w}) > 0$, а, значит, она является функцией правдоподобия

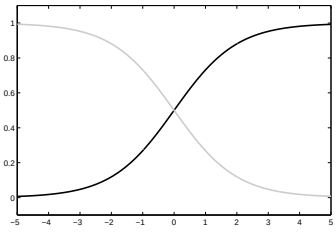


Рис. 3.2. Логистическая функция переводит выход линейной функции в вероятностные значения. Черная кривая показывает правдоподобие для случая $t_i = 1$, а серая кривая — для случая $t_i = -1$

Функционал качества в логистической регрессии

- Правдоподобие правильной классификации всей выборки имеет вид

$$p(\boldsymbol{t}|X, \boldsymbol{w}) = \prod_{i=1}^{n} p(t_i|\boldsymbol{x}_i, \boldsymbol{w}) = \prod_{i=1}^{n} \frac{1}{1 + \exp\left(-t_i \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}_i)\right)}$$

3.3.2 Метод IRLS

Особенности функции правдоподобия классификации

- Приравнивание градиента логарифма правдоподобия к нулю приводит к трансцендентным уравнениям, которые неразрешимы аналитически
- Легко показать, что гессиан логарифма правдоподобия неположительно определен

$$\frac{\partial^2 \log p(t|\boldsymbol{x}, \boldsymbol{w})}{\partial \boldsymbol{w}^2} \leq 0$$

- Это означает, что логарифм функции правдоподобия является вогнутым.
- Логарифм правдоподобия обучающей выборки $L(\boldsymbol{w}) = \log p(\boldsymbol{t}|X, \boldsymbol{w})$, являющийся суммой вогнутых функций, также вогнут, а, значит, имеет **единственный максимум**

Метод оптимизации Ньютона

Основная идея метода Ньютона — это приближение в заданной точке оптимизируемой функции параболой и выбор минимума этой параболы в качестве следующей точки итерационного процесса:

$$f(\boldsymbol{x}) \to \min_{\boldsymbol{w}}$$

$$f(\boldsymbol{x}) \simeq g(\boldsymbol{x}) = f(\boldsymbol{x}_0) + (\nabla f(\boldsymbol{x}_0))^T(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T(\nabla\nabla f(\boldsymbol{x}_0))(\boldsymbol{x} - \boldsymbol{x}_0)$$

$$\nabla g(\boldsymbol{x}_*) = \nabla f(\boldsymbol{x}_0) + (\nabla\nabla f(\boldsymbol{x}_0))(\boldsymbol{x}_* - \boldsymbol{x}_0) = 0$$

$$\boldsymbol{x}_* = \boldsymbol{x}_0 - (\nabla\nabla f(\boldsymbol{x}_0))^{-1}(\nabla f(\boldsymbol{x}_0))$$

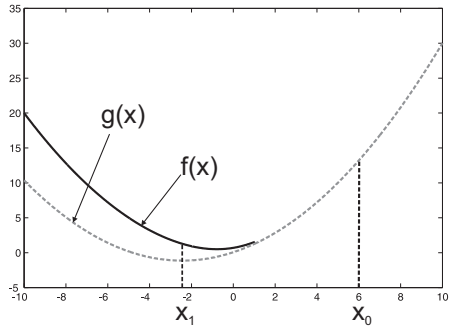


Рис. 3.3. Пример оптимизации с помощью метода Ньютона. Функция $f(x) = \log(1 + \exp(x)) + \frac{x^2}{5}$. В точке $x_0 = 6$ проведено приближение функции $f(x)$ параболой $g(x)$. Точка минимума этой параболы $x_1 = -2.4418$ является следующей точкой итерационного процесса

Итеративная минимизация логарифма правдоподобия

- Так как прямая минимизация правдоподобия невозможна, воспользуемся итерационным методом Ньютона
- Обоснованием корректности использования метода Ньютона является унимодальность оптимизируемой функции $L(\boldsymbol{w})$ и ее гладкость во всем пространстве весов
- Формула пересчета в методе Ньютона

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

где $H = \nabla\nabla L(\boldsymbol{w})$ — гессиан логарифма правдоподобия обучающей выборки

Формулы пересчета

Обозначим $s_i = \frac{1}{1+\exp(-t_i y_i)}$, тогда:

$$\nabla L(\boldsymbol{w}) = \Phi^T \operatorname{diag}(\boldsymbol{t})\boldsymbol{s}, \quad \nabla\nabla L(\boldsymbol{w}) = \Phi^T R\Phi$$

$$R = \begin{pmatrix} s_1(1-s_1) & 0 & \dots & 0 \\ 0 & s_2(1-s_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & s_n(1-s_n) \end{pmatrix}$$

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

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

Название метода (метод наименьших квадратов с итеративно пересчитываемыми весами) связано с тем, что последняя формула является формулой для взвешенного МНК (веса задаются диагональной матрицей R), причем на каждой итерации веса корректируются

Заключительные замечания

- На практике матрица $\Phi^T R \Phi$ часто бывает вырождена (всегда при $m > n$), поэтому обычно прибегают к регуляризации матрицы $(\Phi^T R \Phi + \lambda I)$
- *!! Параметр регуляризации λ является структурным параметром!!*
- *!! Базисные функции $\phi_j(\boldsymbol{x})$, а значит и матрица Φ являются структурными параметрами!!*
- С поиском методов автоматического выбора базисных функций связана одна из наиболее интригующих проблем современного машинного обучения

Глава 4

Метод опорных векторов и беспризнаковое распознавание образов

В главе подробно рассматривается метод опорных векторов для классификации и восстановления регрессии. Особое внимание уделено формулировке двойственной задачи и использованию правила множителей Лагранжа. Описывается т.н. ядровой переход, представляющий нелинейное обобщение метода опорных векторов, показана связь между этим методом и методом максимального правдоподобия с регуляризацией, а также со статистической теорией обучения Вапника-Червоненкиса. В конце главы приведены обобщения метода опорных векторов на задачи, в которых подсчет признаков невозможен или нецелесообразен, но в которых естественным образом можно ввести функцию близости между объектам.

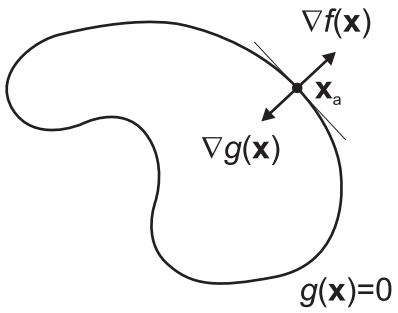


Рис. 4.1. Иллюстрация к задаче оптимизации с ограничениям в виде равенства. В оптимальной точке градиенты ∇f и ∇g должны быть параллельны друг другу

4.1 Ликбез: Условная оптимизация

Задача условной оптимизации

Пусть $f(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ — гладкая функция. Предположим, что нам необходимо найти ее экстремум:

$$f(\boldsymbol{x}) \to \operatorname*{extr}_{\boldsymbol{x}}$$

Для того, чтобы найти экстремум (решить задачу безусловной оптимизации), достаточно проверить условие стационарности:

$$\nabla f(\boldsymbol{x}) = 0$$

Предположим, что нам необходимо найти экстремум функции при ограничениях:

$$f(\boldsymbol{x}) \to \operatorname*{extr}_{\boldsymbol{x}}$$
$$g(\boldsymbol{x}) = 0$$

Поверхность ограничения (см. рис. 4.1)

Заметим, что $\nabla g(\boldsymbol{x})$ ортогонален поверхности ограничения $g(\boldsymbol{x}) = 0$. Пусть $\boldsymbol{x}$ и $\boldsymbol{x} + \boldsymbol{\varepsilon}$ — две близкие точки поверхности. Тогда

$$g(\boldsymbol{x} + \boldsymbol{\varepsilon}) \simeq g(\boldsymbol{x}) + \boldsymbol{\varepsilon}^T \nabla g(\boldsymbol{x})$$

Т.к. $g(\boldsymbol{x} + \boldsymbol{\varepsilon}) = g(\boldsymbol{x})$, то $\boldsymbol{\varepsilon}^T \nabla g(\boldsymbol{x}) \simeq 0$. При стремлении $\|\boldsymbol{\varepsilon}\| \to 0$ получаем $\boldsymbol{\varepsilon}^T \nabla g(\boldsymbol{x}) = 0$. Т.к. $\boldsymbol{\varepsilon}$ параллелен поверхности $g(\boldsymbol{x}) = 0$, то $\nabla g(\boldsymbol{x})$ является нормалью к этой поверхности.

Функция Лагранжа

Необходимым условием оптимальности является ортогональность $\nabla f(\boldsymbol{x})$ поверхности ограничения (в противном случае вектор проекции градиента $\nabla f(\boldsymbol{x})$ на поверхность ограничения имеет ненулевую длину, и можно найти большее значение функции, двигаясь вдоль вектора проекции), т.е.:

$$\nabla f + \lambda \nabla g = 0$$

Здесь $\lambda \neq 0$ — коэффициент Лагранжа. Он может быть любого знака.

Функция Лагранжа

$$L(\boldsymbol{x}, \lambda) \triangleq f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$$

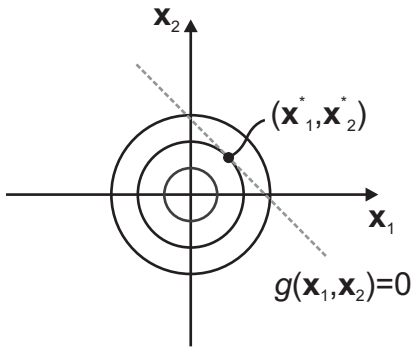


Рис. 4.2. Пример решения задачи условной оптимизации. Черные линии обозначают линии уровня оптимизируемой функции. Серая пунктирная прямая представлет собой область поиска решения. Точка $(x_1^*, x_2^*) = (1/2, 1/2)$ является решением задачи.

Тогда

$$\begin{aligned} \nabla_{\boldsymbol{x}} L = 0 \quad &\Rightarrow \text{ условие ортогональности } \nabla f + \lambda \nabla g = 0 \\ \frac{\partial}{\partial \lambda} L = 0 \quad &\Rightarrow \; g(\boldsymbol{x}) = 0 \end{aligned}$$

Функция Лагранжа. Пример (см. рис. 4.2)

$$\begin{aligned} f(x_1, x_2) = 1 - x_1^2 - x_2^2 &\to \max_{x_1, x_2} \\ g(x_1, x_2) = x_1 + x_2 - 1 &= 0 \end{aligned}$$

Функция Лагранжа:

$$L(\boldsymbol{x}, \lambda) = 1 - x_1^2 - x_2^2 + \lambda(x_1 + x_2 - 1)$$

Условия стационарности:

$$\begin{aligned} -2x_1 + \lambda &= 0 \\ -2x_2 + \lambda &= 0 \\ x_1 + x_2 - 1 &= 0 \end{aligned}$$

Решение: $(x_1^*, x_2^*) = (\frac{1}{2}, \frac{1}{2}),\ \lambda = 1.$

Ограничение в виде неравенства (см. рис. 4.3)

Задача условной оптимизации

$$\begin{aligned} f(\boldsymbol{x}) &\to \max_{\boldsymbol{x}} \\ g(\boldsymbol{x}) &\geq 0 \end{aligned}$$

Решение	Ограничение	Условие стационарности
Внутри области $g(\boldsymbol{x}) > 0$	неактивно	$\nabla f(\boldsymbol{x}) = 0$, $\nabla_{\boldsymbol{x}} L = 0$, $\lambda = 0$
На границе $g(\boldsymbol{x}) = 0$	активно	$\nabla f(\boldsymbol{x}) = -\lambda \nabla g(\boldsymbol{x})$, $\nabla_{\boldsymbol{x},\lambda} L = 0$, $\lambda > 0$

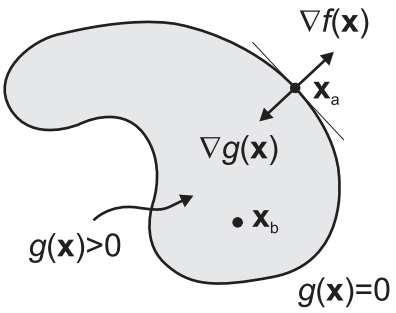


Рис. 4.3. Иллюстрация к задаче оптимизации с ограничениями в виде неравенства. В случае, если оптимальная точка лежит вне области $g(\boldsymbol{x}) \geq 0$, то в точке оптимума градиенты ∇f и ∇g должны быть коллинеарны и направлены в разные стороны

Условие дополняющей нежесткости:

$$\lambda g(\boldsymbol{x}) = 0$$

Теорема Каруша-Куна-Таккера

Пусть $f_i : X \to \mathbb{R},\ i = 0, 1, \ldots, m$ — выпуклые функции, отображающие нормированное пространство X в прямую, $A \in X$ — выпуклое множество. Рассмотрим следующую задачу оптимизации:

$$f_0(\boldsymbol{x}) \to \min; \quad f_i(\boldsymbol{x}) \leq 0,\ i = 1, \ldots, m,\ \boldsymbol{x} \in A \tag{P}$$

Теорема 1.

1. Если $\hat{\boldsymbol{x}} \in \operatorname{absmin}(P)$ *— решение задачи, то найдется ненулевой вектор множителей Лагранжа* $\boldsymbol{\lambda} \in \mathbb{R}^{m+1}$ *такой, что для функции Лагранжа* $L(\boldsymbol{x}) = \sum_{i=0}^{m} \lambda_i f_i(\boldsymbol{x})$ *выполняются условия:*

a) стационарности $\min_{\boldsymbol{x} \in A} L(\boldsymbol{x}) = L(\hat{\boldsymbol{x}})$

b) дополняющей нежесткости $\lambda_i f_i(\hat{\boldsymbol{x}}) = 0,\ i = 1, \ldots, m$

c) неотрицательности $\lambda_i \geq 0$

2. Если для допустимой точки $\hat{\boldsymbol{x}}$ *выполняются условия a)–c) и* $\lambda_0 \neq 0$*, то* $\hat{\boldsymbol{x}} \in \operatorname{absmin}(P)$

3. Если для допустимой точки $\hat{\boldsymbol{x}}$ *выполняются условия a)–c) и* $\exists \tilde{\boldsymbol{x}} \in A : f_i(\tilde{\boldsymbol{x}}) < 0, i = 0, \ldots, m$ *(условие Слейтера), то* $\hat{\boldsymbol{x}} \in \operatorname{absmin}(P)$

4.2 Метод опорных векторов для задачи классификации

Задача классификации

- Рассматривается задача классификации на два класса. Имеется обучающая выборка $(X, \boldsymbol{t}) = \{\boldsymbol{x}_i, t_i\}_{i=1}^{n}$, где $\boldsymbol{x} \in \mathbb{R}^d$, а метка класса $t \in \mathcal{T} = \{-1, 1\}$.
- Необходимо с использованием обучающей выборки построить отображение $A : \mathbb{R}^d \to \mathcal{T}$, которое для каждого нового входного объекта $\boldsymbol{x}_*$ выдает его метку класса t_*.

4.2.1 Метод потенциальных функций

Метод потенциальных функций [Айзерман и др., 1964]

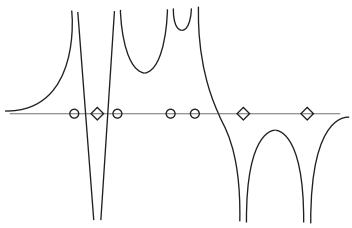


Рис. 4.4. Иллюстрация к методу потенциальных функций.

В каждом объекте $\boldsymbol{x}_i$ помещен электрический заряд $t_i q_i$. В качестве разделяющей функции используется потенциал создаваемого поля:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} t_i q_i K(\boldsymbol{x}, \boldsymbol{x}_i)$$

Функция $K(\boldsymbol{x}, \boldsymbol{y})$ называется потенциальной функцией и имеет смысл потенциала в точке $\boldsymbol{y}$, создаваемого зарядом в точке $\boldsymbol{x}$. Предполагается, что

$$K(\boldsymbol{x}, \boldsymbol{y}) \to 0 \text{ при } \|\boldsymbol{x} - \boldsymbol{y}\| \to +\infty$$
$$K(\boldsymbol{x}, \boldsymbol{y}) \to \max \text{ при } \|\boldsymbol{x} - \boldsymbol{y}\| \to 0$$

Алгоритм обучения:

$$f^{new}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) + K(\boldsymbol{x}, \boldsymbol{x}_k), & \text{если } t_k = 1 \text{ и } f(\boldsymbol{x}_k) \le 0 \\ f(\boldsymbol{x}) - K(\boldsymbol{x}, \boldsymbol{x}_k), & \text{если } t_k = -1 \text{ и } f(\boldsymbol{x}_k) \ge 0 \\ f(\boldsymbol{x}), & \text{иначе} \end{cases}$$

4.2.2 Случай линейно разделимых данных

Разделяющая гиперплоскость

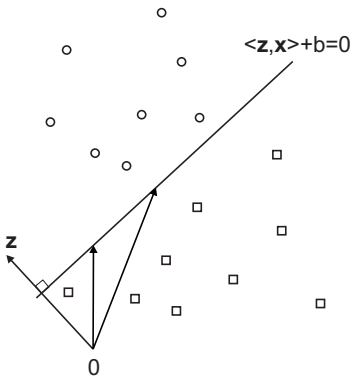


Рис. 4.5. Разделяющая гиперплоскость для объектов двух классов. Направление гиперплоскости задается вектором нормали $\boldsymbol{z}$

- Гиперплоскость задается направляющим вектором гиперплоскости $\boldsymbol{z}$ и величиной сдвига b (см. рис. 4.5):
$$\{\boldsymbol{x} \in \mathbb{R}^d | \langle \boldsymbol{z}, \boldsymbol{x} \rangle + b = 0\}, \quad \boldsymbol{z} \in \mathbb{R}^d, \ b \in \mathbb{R}$$
- Если вектор $\boldsymbol{z}$ имеет единичную длину, то величина $\langle \boldsymbol{z}, \boldsymbol{x} \rangle$ определяет длину проекции вектора $\boldsymbol{x}$ на направляющий вектор $\boldsymbol{z}$. В случае произвольной длины скалярное произведение нормируется на $\|\boldsymbol{z}\|$.

Каноническая гиперплоскость

- Если величину направляющего вектора $\boldsymbol{z}$ и величину сдвига b умножить на одно и то же число, то соответствующая им гиперплоскость не изменится.
- Пара $(\boldsymbol{z}, b)$ задает **каноническую гиперплоскость** для набора объектов $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, если
$$\min_{i=1,\ldots,n} |\langle \boldsymbol{z}, \boldsymbol{x}_i \rangle + b| = 1 \qquad (*)$$
- Условие (*) означает, что ближайший вектор к канонической гиперплоскости находится от нее на расстоянии $1/\|z\|$:
$$\boldsymbol{x}_i : |\langle \boldsymbol{z}, \boldsymbol{x}_i \rangle + b| = 1, \ \ \boldsymbol{x} : \langle \boldsymbol{z}, \boldsymbol{x} \rangle + b = 0$$
$$|\langle \boldsymbol{z}, \boldsymbol{x}_i - \boldsymbol{x} \rangle| = 1 \Rightarrow \left| \left\langle \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|}, \boldsymbol{x}_i - \boldsymbol{x} \right\rangle \right| = \frac{1}{\|\boldsymbol{z}\|}$$
- Каноническая гиперплоскость определена однозначно с точностью до знака $\boldsymbol{z}$ и b

Классификатор

- Будем искать классификатор в виде разделяющей гиперплоскости (см. рис. 4.6):
$$\hat{t}(\boldsymbol{x}) = \text{sign}(y(\boldsymbol{x})) = \text{sign}(\langle \boldsymbol{z}, \boldsymbol{x} \rangle + b)$$
- Предположим, что исходные данные являются линейно разделимыми, т.е.
$$\exists (\boldsymbol{z}, b) : \hat{t}(\boldsymbol{x}_i) = t_i \ \forall i = 1, \ldots, n$$

Оптимальная разделяющая гиперплоскость

- Зазором гиперплоскости данной точки $(\boldsymbol{x}, t)$ называется величина:
$$\rho_{(\boldsymbol{z},b)}(\boldsymbol{x}, t) = \frac{t(\langle \boldsymbol{z}, \boldsymbol{x} \rangle + b)}{\|z\|}$$
Зазором гиперплоскости называется величина:
$$\rho_{(\boldsymbol{z},b)} = \min_{i=1,\ldots,n} \rho_{(\boldsymbol{z},b)}(\boldsymbol{x}_i, t_i)$$
- Для корректно распознаваемого объекта его величина зазора соответствует расстоянию от этого объекта до гиперплоскости. Отрицательная величина зазора соответствует ошибочной классификации объекта.
- **Оптимальная разделяющая гиперплоскость — гиперплоскость с максимальной величиной зазора**

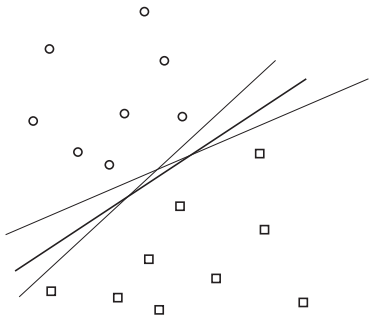


Рис. 4.6. Для линейно разделимой выборки существует бесконечно большое число гиперплоскостей, корректно разделяющих данные. Жирной линией показана оптимальная разделяющая гиперплоскость.

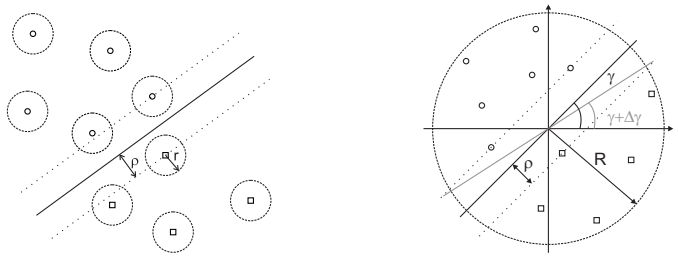


Рис. 4.7. Иллюстрация к оптимальной разделяющей гиперплоскости.

Оптимальная разделяющая гиперплоскость. Примеры

- Предположим, что все объекты тестовой выборки получены путем небольших смещений относительно обучающих объектов, т.е. для объекта обучения $(\boldsymbol{x}, t)$ тестовый объект может быть представлен как $(\boldsymbol{x} + \Delta\boldsymbol{x}, t)$, причем $\|\Delta\boldsymbol{x}\| \leq r$.
- Гиперплоскость с величиной зазора $\rho > r$ корректно классифицирует выборку (см. рис. 4.7, a).
- Если объекты располагаются на достаточном расстоянии от гиперплоскости, то небольшое изменение параметров $(\boldsymbol{z}, b)$ не меняет корректного разделения данных (см. рис. 4.7, b).

Оптимальная разделяющая гиперплоскость. Стат. теория Вапника-Червоненкиса

- Пусть имеется некоторое объективное распределение $p(\boldsymbol{x}, t)$. Обучающая и тестовая совокупности являются н.о.р. выборками из этого распределения.
- Пусть имеется семейство классификаторов $\{f(\boldsymbol{x}, \boldsymbol{w}) : \mathbb{R}^d \to \mathcal{T} | \boldsymbol{w} \in \Omega\}$ и функция потерь $L : \mathcal{T} \times \mathcal{T} \to \mathbb{R}_+$.
- **Средним риском** называется мат.ожидание функции потерь:

$$R(\boldsymbol{w}) = \mathbb{E}_{p(\boldsymbol{x},t)} L(\cdot) = \int L(t, f(\boldsymbol{x}, \boldsymbol{w})) p(\boldsymbol{x}, t) d\boldsymbol{x} dt$$

- **Эмпирическим риском** называется следующая величина:

$$R_{emp}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{n} L(t_i, f(\boldsymbol{x}_i, \boldsymbol{w}))$$

Теорема 2 (Vapnik, 1995)**.** *С вероятностью* $0 < \eta \leq 1$ *выполняется следующее неравенство:*

$$R(\boldsymbol{w}) \leq R_{emp}(\boldsymbol{w}) + \sqrt{\frac{h(\ln(2n/h)+1) - \ln(\eta/4)}{n}} \qquad (*)$$

Здесь h — положительная величина, называемая ВЧ-размерностью.

Теорема 3 (Vapnik, 1995)**.** *Допустим, что вектора* $\boldsymbol{x} \in \mathbb{R}^d$ *принадлежат сфере радиуса* R*. Тогда для семейства разделяющих гиперплоскостей с величиной зазора* ρ *верно следующее:*

$$h \leq \min\left(\left[\frac{R^2}{\rho^2}\right], d\right) + 1 \qquad (**)$$

Для гиперплоскости, корректно разделяющей данные, эмпирический риск равен нулю. Корень в оценке на средний риск (*) является монотонно возрастающей функцией по h. Поэтому минимизация оценки эквивалентна минимизации ВЦ-размерности, что согласно формуле (**) эквивалентно максимизации зазора ρ.

Постановка задачи оптимизации

- Предположим, что в выборке присутствуют объекты двух классов, т.е. $\exists i, j : t_i = 1,\ t_j = -1$.
- Для построения канонической гиперплоскости с максимальной величиной зазора, корректно разделяющей данные, необходимо решить следующую задачу оптимизации:

$$\frac{1}{2}\|z\|^2 \to \min_{\boldsymbol{z},b}$$
$$t_i(\langle \boldsymbol{z}, \boldsymbol{x}_i\rangle + b) \geq 1,\ \forall i = 1, \ldots, n$$

Функция Лагранжа

Выпишем функцию Лагранжа

$$L(\boldsymbol{z}, b, \boldsymbol{w}) = \frac{1}{2}\|z\|^2 - \sum_{i=1}^{n} w_i(t_i(\langle \boldsymbol{z}, \boldsymbol{x}_i\rangle + b) - 1) \to \min_{\boldsymbol{z},b}\max_{\boldsymbol{w}}$$

Коэффициенты Лагранжа $w_i \geq 0,\ \forall i = 1, \ldots, n$.

$$\frac{\partial}{\partial \boldsymbol{z}} L(\boldsymbol{z}, b, \boldsymbol{w}) = 0 \Rightarrow \boldsymbol{z}_* = \sum_{i=1}^{n} w_i t_i \boldsymbol{x}_i$$
$$\frac{\partial}{\partial b} L(\boldsymbol{z}, b, \boldsymbol{w}) = 0 \Rightarrow \sum_{i=1}^{n} w_i t_i = 0$$

$$L(\boldsymbol{z}_*, b_*, \boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j t_i t_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle - \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j t_i t_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle +$$
$$+ \sum_{i=1}^{n} w_i = \sum_{i=1}^{n} w_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j t_i t_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle \to \max_{\boldsymbol{w}}$$

Двойственная задача оптимизации

$$\sum_{i=1}^{n} w_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j t_i t_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \to \max_{\boldsymbol{w}}$$
$$\sum_{i=1}^{n} t_i w_i = 0$$
$$w_i \geq 0,\ \forall i = 1, \ldots, n$$

Решение

$$\hat{t}(\boldsymbol{x}) = \operatorname{sign}\left(\langle \boldsymbol{z}_*, \boldsymbol{x} \rangle + b_*\right) = \operatorname{sign}\left(\sum_{i=1}^{n} w_i^* t_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b_*\right)$$

Опорные вектора

Условие дополняющей нежесткости:

$$w_i^*(t_i(\langle \boldsymbol{z}_*, \boldsymbol{x}_i \rangle + b_*) - 1) = 0$$

Из этого условия следует, что объекты обучающей выборки, для которых $w_i^* > 0$, лежат точно на границе гиперплоскости. Они называются **опорными векторами** (см. рис. 4.8). Значение b_* может быть получено

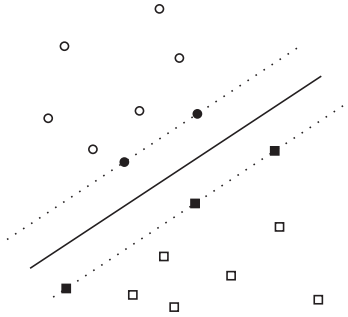


Рис. 4.8. Оптимальная разделяющая гиперплоскость. Объекты, определяющие положение гиперплоскости, называются опорными (обозначены черным цветом). Они располагаются ближе всего к гиперплоскости.

из условия дополняющей нежесткости для любого опорного вектора. При этом с вычислительной точки зрения более устойчивой процедурой является усреднение по всем таким объектам.

Разреженность решения

Решающее правило

$$\hat{t}(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{n} w_i t_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b\right)$$

В решающее правило входят только те объекты обучения, для которых $w_i > 0$ (опорные векторы). Такое правило называется **разреженным (sparse model)**. Разреженные модели обладают высокой скоростью распознавания в больших объемах данных, а также «проливают свет» на структуру обучающей совокупности, выделяя наиболее релевантные с точки зрения классификации объекты.

4.2.3 Случай линейно неразделимых данных

Ослабляющие коэффициенты

На практике данные не являются, как правило, линейно разделимыми. Кроме того, даже линейно разделимые выборки могут содержать помехи, ошибочные метки классов и проч. Практический метод распознавания должен учитывать подобные ситуации.

Предположим, что $\forall(\boldsymbol{z}, b)\ \exists \boldsymbol{x}_i : \rho_{(\boldsymbol{z},b)}(\boldsymbol{x}_i, t_i) < 0$. Позволим некоторым из ограничений не выполняться путем введения ослабляющих коэффициентов:

$$t_i(\langle \boldsymbol{z}, \boldsymbol{x}_i\rangle + b) \geq 1\ \forall i = 1, \ldots, n \ \longrightarrow\ \begin{cases} t_i(\langle \boldsymbol{z}, \boldsymbol{x}_i\rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0,\ \forall i = 1, \ldots, n \end{cases}$$

При этом потребуем, чтобы количество нарушений (количество ошибок на обучении) было бы как можно меньшим:

$$\frac{1}{2}\|\boldsymbol{z}\|^2 + C\sum_{i=1}^{n}\xi_i \to \min_{\boldsymbol{z},b,\boldsymbol{\xi}}$$

Постановка задачи оптимизации

$$\begin{aligned} &\frac{1}{2}\|\boldsymbol{z}\|^2 + C\sum_{i=1}^{n}\xi_i \to \min_{\boldsymbol{z},b} \\ &t_i(\langle \boldsymbol{z}, \boldsymbol{x}_i\rangle + b) \geq 1 - \xi_i \quad \forall i = 1, \ldots, n \\ &\xi_i \geq 0 \end{aligned}$$

Здесь $C \geq 0$ — некоторый действительный параметр, играющий роль параметра регуляризации

Функция Лагранжа

Выпишем функцию Лагранжа

$$L(\boldsymbol{z}, b, \boldsymbol{\xi}, \boldsymbol{w}, \boldsymbol{v}) = \frac{1}{2}\|\boldsymbol{z}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n} w_i[t_i(\langle \boldsymbol{z}, \boldsymbol{x}_i\rangle + b) - 1 + \xi_i] - \sum_{i=1}^{n} v_i\xi_i \to \min_{\boldsymbol{z},b,\boldsymbol{\xi}}\max_{\boldsymbol{w},\boldsymbol{v}}$$

Коэффициенты Лагранжа $w_i \geq 0,\ v_i \geq 0$.

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{z}}L(\boldsymbol{z}, b, \boldsymbol{\xi}, \boldsymbol{w}, \boldsymbol{v}) = 0 && \Rightarrow \boldsymbol{z}_* = \sum_{i=1}^{n} w_i t_i \boldsymbol{x}_i \\ &\frac{\partial}{\partial b}L(\boldsymbol{z}, b, \boldsymbol{\xi}, \boldsymbol{w}, \boldsymbol{v}) = 0 && \Rightarrow \sum_{i=1}^{n} w_i t_i = 0 \\ &\frac{\partial}{\partial \boldsymbol{\xi_i}}L(\boldsymbol{z}, b, \boldsymbol{\xi}, \boldsymbol{w}, \boldsymbol{v}) = 0 && \Rightarrow w_i + v_i = C \end{aligned}$$

Двойственная задача оптимизации

$$\sum_{i=1}^{n} w_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j t_i t_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \to \max_{\boldsymbol{w}}$$
$$\sum_{i=1}^{n} w_i t_i = 0$$
$$0 \leq w_i \leq C$$

Решающее правило остается без изменений:

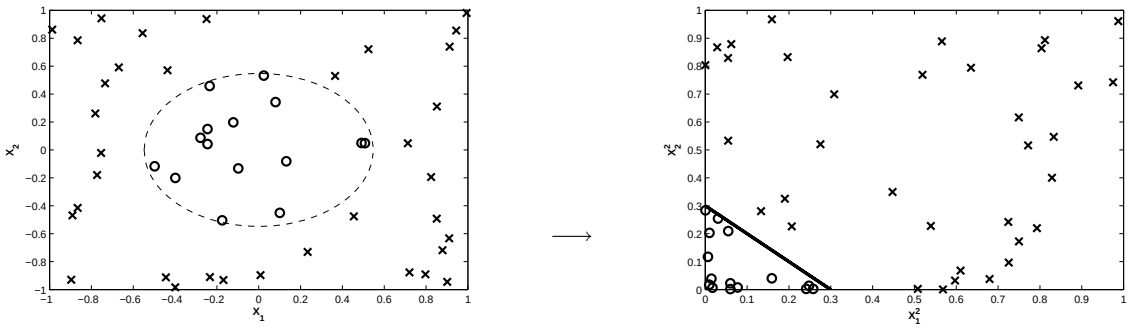
$$\hat{t}(\boldsymbol{x}) = \operatorname{sign}\left(\langle \boldsymbol{z}_*, \boldsymbol{x} \rangle + b_*\right) = \operatorname{sign}\left(\sum_{i=1}^{n} w_i^* t_i \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b^*\right)$$

4.2.4 Ядровой переход

Ядровой переход

- На практике часто встречается ситуация, когда данные порождаются нелинейной разделяющей поверхностью.
- Для обобщения метода на нелинейный случай заметим, что объекты обучающей выборки входят в двойственную задачу оптимизации только в виде попарных скалярных произведений $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$.
- Предположим, что исходное признаковое пространство было подвергнуто некоторому нелинейному преобразованию (см. рис. 4.9):

$$\Phi : \mathbb{R}^d \to \mathcal{H}$$



(a) (b)

Рис. 4.9. Пример преобразования признакового пространства. На рис. (a) данные разделяются нелинейной поверхностью — эллипсом. Переход из пространства (x_1, x_2) к новому пространству (x_1^2, x_2^2) делает данные линейно разделимыми.

Ядровой переход

- Для того, чтобы построить гиперплоскость с максимальным зазором в новом пространстве $\mathcal{H}$ необходимо знать лишь $\langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle_{\mathcal{H}}$.
- Допустим, что существует некоторая «ядровая функция» $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, такая что

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) \rangle_{\mathcal{H}}$$

- **Для построения гиперплоскости с максимальным зазором в пространстве $\mathcal{H}$ нет необходимости задавать преобразование Φ в явном виде, достаточно лишь знать K!**
- Задача оптимизации зависит только от попарных скалярных произведений, а решающее правило может быть представлено как

$$\hat{t}(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{n} w_i t_i \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}) \rangle_{\mathcal{H}} + b \right) = \operatorname{sign}\left(\sum_{i=1}^{n} w_i t_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \right)$$

Требования к ядровой функции

Очевидно, что не для любой функции двух переменных K найдутся такие $(\mathcal{H}, \Phi)$, для которых K будет определять скалярное произведение. Необходимыми и достаточными требованиями являются:

- Симметричность

$$K(\boldsymbol{x}, \boldsymbol{y}) = K(\boldsymbol{y}, \boldsymbol{x})$$

- Неотрицательная определенность (условие Мерсера)

$$\forall g(\boldsymbol{x}) : \int g^2(\boldsymbol{x}) d\boldsymbol{x} < \infty$$
$$\int K(\boldsymbol{x}, \boldsymbol{y}) g(\boldsymbol{x}) g(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \geq 0$$

Для фиксированной функции K евклидово пространство $\mathcal{H}$ и преобразование Φ определено не однозначно.

Примеры ядровых функций

- Линейная ядровая функция

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \theta, \ \theta \geq 0$$

- Полиномиальная ядровая функция

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + \theta)^d, \ \theta \geq 0, d \in \mathbb{N}$$

- Гауссиана

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\sigma^2} \right), \ \sigma > 0$$

- Сигмоидная ядровая функция

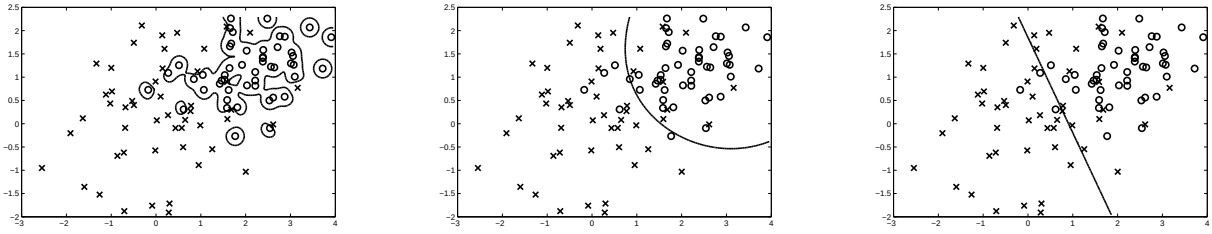
$$K(\boldsymbol{x}, \boldsymbol{y}) = \tanh(\langle \boldsymbol{x}, \boldsymbol{y} \rangle + r), \ r \in \mathbb{R}$$

Это семейство не удовлетворяет условию Мерсера!

4.2.5 Заключительные замечания

Пример использования метода опорных векторов

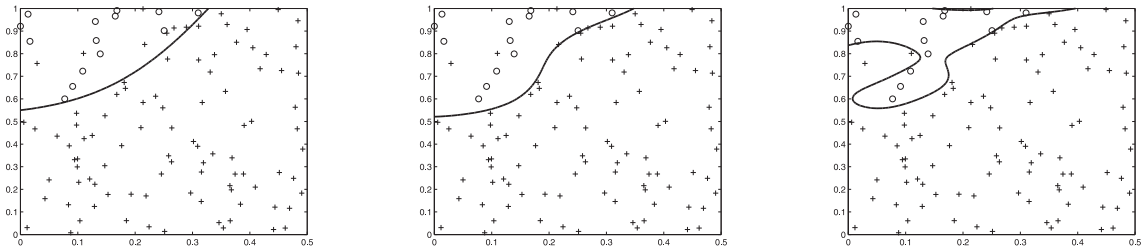
Зависимость от ширины гауссианы (см. рис. 4.10)



(a) (b) (c)

Рис. 4.10. Примеры решения двухклассовой задачи классификации с помощью метода опорных векторов. На рисунке (a) показана разделяющая поверхность для случая использования в качестве ядровой функции гауссианы с параметрами $C = 1,\ \sigma^2 = 0.1$, (b) соответствует $C = 1,\ \sigma^2 = 2$, (c) — $C = 1,\ \sigma^2 = 1000$

Зависимость от штрафного коэффициента (см. рис. 4.11)



(a) (b) (c)

Рис. 4.11. Примеры решения двухклассовой задачи классификации с помощью метода опорных векторов. На рисунке (a) показана разделяющая поверхность для случая $C = 10^{-2}$, (b) соответствует $C = 1$, (c) — $C = 10^5$

Глобальность и единственность решения

- Задача обучения SVM — задача квадратичного программирования
- Известно, что для любой задачи выпуклого программирования (в частности, квадратичного) любой локальный максимум является и глобальным. Кроме того, решение будет единственным, если целевая функция строго вогнута (гессиан отрицательно определен).
- Для обучения SVM можно воспользоваться любым стандартным методом решения задачи квадратичного программирования, однако лучше использовать специальные алгоритмы, учитывающие особенности задачи квадратичного программирования в SVM (например, SMO или SVM^{light}).
- Подробнее см. http://www.kernel-machines.org

Задача обучения SVM как задача максимума регуляризованного правдоподобия

$$\frac{1}{2}\|\boldsymbol{z}\|^2 + C\sum_{i=1}^{n}\xi_i \to \min_{\boldsymbol{z},b,\boldsymbol{\xi}}$$
$$t_i(\langle \boldsymbol{z}, \boldsymbol{x}_i\rangle + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

Если $t_i y(\boldsymbol{x}_i) \geq 1$, то $\xi_i = 0$. Для остальных точек $\xi_i = 1 - t_i y(\boldsymbol{x}_i)$. Следовательно, оптимизируемую функцию можно переписать в виде

$$\sum_{i=1}^{n} E_{SV}(t_i y(\boldsymbol{x}_i)) + \lambda\|\boldsymbol{z}\|^2$$

Здесь $\lambda = (2C)^{-1}$, а $E_{SV}(\cdot)$ — функция потерь, определенная как

$$E_{SV}(s) = [1 - s]_+$$

SVM vs. Логистическая регрессия

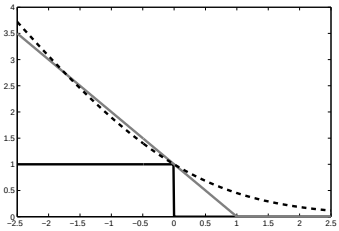


Рис. 4.12. Различные функции ошибок. Черная кривая соответствует индикаторной функции ошибки, серая кривая — функция ошибки в методе опорных векторов, пунктирная линия — функция ошибок в логистической регрессии

Задача оптимизации в логистической регрессии:

$$\sum_{i=1}^{n} E_{LR}(t_i y(\boldsymbol{x}_i)) + \lambda\|\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w}}$$

Здесь $E_{LR}(s) = \log(1 + \exp(-s))$.

Задача оптимизации в SVM:

$$\sum_{i=1}^{n} E_{SV}(t_i y(\boldsymbol{x}_i)) + \lambda\|\boldsymbol{z}\|^2 \to \min_{\boldsymbol{z}}$$

Достоинства и недостатки SVM

+ Высокое качество распознавания за счет построения нелинейных разделяющих поверхностей, максимизирующих зазор
+ Глобальность и в ряде случаев единственность получаемого решения
— Низкая скорость обучения и большие требования к памяти для задач больших размерностей
— Необходимость грамотного выбора штрафного коэффициента C и параметров ядровой функции

4.3 Метод опорных векторов для задачи регрессии

Линейная регрессия vs. SVR

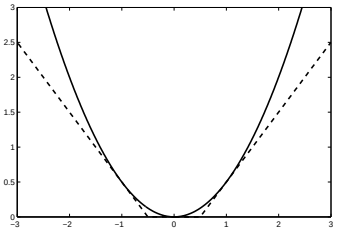


Рис. 4.13. Различные функции ошибок. Черная кривая — квадратичная функция ошибок линейной регрессии, пунктирная линия — функция ошибок в методе опорных векторов для регрессии

Задача оптимизации в линейной регрессии

$$\frac{1}{2}\sum_{i=1}^{n}(t_i - y(\boldsymbol{x}_i))^2 + \frac{1}{2}\|\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w}}$$

Для того, чтобы добиться разреженного решения, заменим квадратичную функцию потерь на ε-нечувствительную:

$$E_\varepsilon(t - y(\boldsymbol{x})) = \begin{cases} 0, & \text{если } |t - y(\boldsymbol{x})| < \varepsilon \\ |t - y(\boldsymbol{x})| - \varepsilon, & \text{иначе} \end{cases}$$

Тогда мы приходим к следующей оптимизационной задаче:

$$C\sum_{i=1}^{n} E_\varepsilon(y(\boldsymbol{x}_i) - t_i) + \frac{1}{2}\|\boldsymbol{z}\|^2 \to \min_{\boldsymbol{z}}$$

Ослабляющие коэффициенты

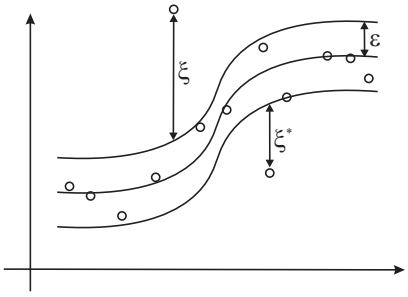


Рис. 4.14. Иллюстрация к введению ослабляющих коэффициентов. Объекты, лежащие выше ε-трубки, имеют положительное значение ξ_i, а объекты, лежащие ниже ε-трубки, имеют положительное значение ξ_i^*

$$C\sum_{i=1}^{n}(\xi_i+\xi_i^*)+\frac{1}{2}\|\boldsymbol{z}\|^2\to\min_{\boldsymbol{z},b,\boldsymbol{\xi},\boldsymbol{\xi}^*}$$
$$t_i\le y(\boldsymbol{x}_i)+\varepsilon+\xi_i$$
$$t_i\ge y(\boldsymbol{x}_i)-\varepsilon-\xi_i^*$$
$$\xi_i,\xi_i^*\ge 0$$

✓ Упр.

Двойственная задача

$$-\frac{1}{2}\sum_{i,j=1}^{n}(w_i-w_i^*)(w_j-w_j^*)K(\boldsymbol{x}_i,\boldsymbol{x}_j)-\varepsilon\sum_{i=1}^{n}(w_i+w_i^*)+\sum_{i=1}^{n}t_i(w_i-w_i^*)\to\max_{\boldsymbol{w},\boldsymbol{w}^*}$$
$$\sum_{i=1}^{n}(w_i-w_i^*)=0$$
$$0\le w_i,w_i^*\le C$$

Функция регрессии

$$y(\boldsymbol{x})=\sum_{i=1}^{n}(w_i-w_i^*)K(\boldsymbol{x},\boldsymbol{x}_i)+b$$

Условия дополняющей нежесткости

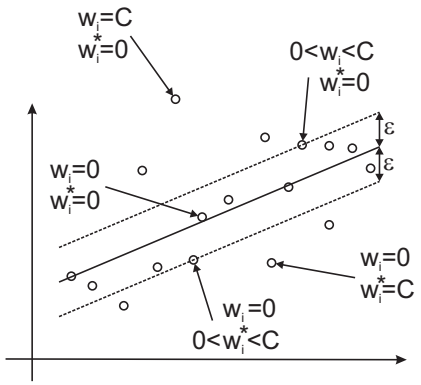


Рис. 4.15. Иллюстрация к опорным объектам в задаче регрессии

Запишем условия дополняющей нежесткости

$$w_i(\varepsilon+\xi_i-t_i+\langle\boldsymbol{z},\boldsymbol{x}_i\rangle+b)=0$$
$$w_i(\varepsilon+\xi_i^*+t_i-\langle\boldsymbol{z},\boldsymbol{x}_i\rangle-b)=0$$
$$(C-w_i)\xi_i=0,\qquad(C-w_i^*)\xi_i^*=0$$

Из них следует, что опорные объекты лежат за пределами или на границе ε-трубки, определяемой функцией $\langle\boldsymbol{z},\boldsymbol{x}\rangle+b$

4.4 Беспризнаковое распознавание образов

4.4.1 Основная методика беспризнакового распознавания образов

Задачи беспризнакового распознавания образов

Существует ряд задач распознавания образов, в которых трудно выбрать признаковое пространство, однако, относительно легко ввести меру сходства или несходства между парами объектов. Примеры:

- Задача распознавания личности по фотопортрету
- Задача идентификации личности по подписи в процессе ее формирования
- Задача распознавания классов пространственной структуры белков по последовательностям составляющих их аминокислот

Беспризнаковое распознавание образов: основная идея

- Предположим, что объекты выборки $\omega_1, \ldots, \omega_n \in \Omega$
- Пространство Ω является гильбертовым, т.е. на нем определены операции суммы и произведения на число

$$\forall \alpha_1, \alpha_2 \in \Omega \; \exists \alpha = \alpha_1 + \alpha_2 \in \Omega$$
$$\forall \alpha_1 \in \Omega, c \in \mathbb{R} \; \exists \alpha = c\alpha_1 \in \Omega,$$

  удовлетворяющие аксиомам линейного пространства:

1. $\alpha_1 + \alpha_2 = \alpha_2 + \alpha_1$
2. $\alpha_1 + (\alpha_2 + \alpha_3) = (\alpha_1 + \alpha_2) + \alpha_3$
3. $\exists \phi \in \Omega : \alpha + \phi = \phi + \alpha = \alpha$
4. $\forall \alpha \; \exists (-\alpha) : \alpha + (-\alpha) = \phi$
5. $c(\alpha_1 + \alpha_2) = c\alpha_1 + c\alpha_2$
6. $(c + d)\alpha = c\alpha + d\alpha$
7. $(cd)\alpha = c(d\alpha)$
8. $1\alpha = \alpha$

- Существует функция $K : \Omega \times \Omega \to \mathbb{R}$, определяющая скалярное произведение в пространстве Ω:

1. $K(\alpha_1, \alpha_2) = K(\alpha_2, \alpha_1)$
2. $K(\alpha, \alpha) \geq 0, \; = 0 \Leftrightarrow \alpha = \phi$
3. $K(\alpha_1 + \alpha_2, \alpha) = K(\alpha_1, \alpha) + K(\alpha_2, \alpha)$
4. $K(c\alpha_1, \alpha_2) = cK(\alpha_1, \alpha_2)$

Решающая функция	$\hat{t}(\omega) = \text{sign}(K(\vartheta, \omega) + b)$
Задача оптимизации	$\frac{1}{2}K(\vartheta,\vartheta) + C\sum_{i=1}^{n}\xi_i \to \min\limits_{\vartheta,b,\boldsymbol{\xi}}$ $t_i(K(\vartheta,\omega_i) + b) \geq 1 - \xi_i$ $\xi_i \geq 0$
Двойственная задача оптимизации	$\sum_{i=1}^{n} w_i - \frac{1}{2}\sum_{i,j=1}^{n} t_i t_j w_i w_j K(\omega_i,\omega_j) \to \max\limits_{\boldsymbol{w}}$ $\sum_{i=1}^{n} t_i w_i = 0$ $0 \leq w_i \leq C$
Решение	$\hat{t}(\omega) = \text{sign}\left(\sum_{i=1}^{n} t_i w_i K(\omega_i,\omega) + b\right)$

- Для решения задачи нет необходимости явно задавать линейные операции в пространстве Ω, достаточно лишь потребовать их существование в пространстве, где функция K задает скалярное произведение
- Для применения данного подхода достаточно задать функцию K. Эта функция может быть построена непосредственно, либо с использованием меры сходства

4.4.2 Построение функции, задающей скалярное произведение

Явное задание функции K. Пример. [Середин, 2001]

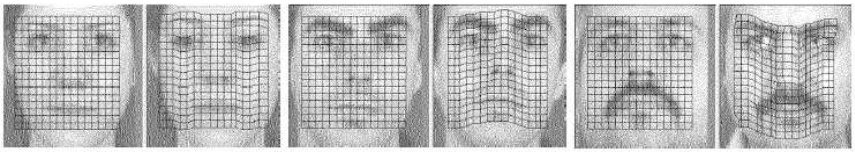


Рис. 4.16. Примеры фотографий лиц в задаче идентификации личности по фотопортрету. На фотографиях также представлено эластичное преобразование растра при сравнении двух фотопортретов, получающееся в результате решения задачи минимизации искажений при условии минимального отклонения точек растра от исходного положения

- Рассмотрим задачу распознавания личности по фотопортрету.
- Пусть два изображения заданы векторами яркости в точках растра: $\boldsymbol{y}' = \boldsymbol{y}(\omega') = (y'_{\boldsymbol{t}}, \boldsymbol{t} \in T)$ и $\boldsymbol{y}'' = \boldsymbol{y}(\omega'') = (y''_{\boldsymbol{t}}, \boldsymbol{t} \in T)$, $T = \{\boldsymbol{t} = (t_1, t_2), t_1 = \overline{1, n_1}, t_2 = \overline{1, n_2}\}$
- Потенциальная функция может быть задана как:
  - $K(\boldsymbol{y}', \boldsymbol{y}'') = \langle \boldsymbol{y}', \boldsymbol{y}'' \rangle = \sum_{\boldsymbol{t}\in T} y'_{\boldsymbol{t}} y''_{\boldsymbol{t}}$
  - $K(\boldsymbol{y}', \boldsymbol{y}'') = [\langle \boldsymbol{y}', \boldsymbol{y}'' \rangle + 1]^{\alpha}$
  - $K(\boldsymbol{y}', \boldsymbol{y}'') = \exp\left(-\alpha\|\boldsymbol{y}' - \boldsymbol{y}''\|^2\right) = \exp\left(-\alpha[\langle \boldsymbol{y}', \boldsymbol{y}' \rangle + \langle \boldsymbol{y}'', \boldsymbol{y}'' \rangle - 2\langle \boldsymbol{y}', \boldsymbol{y}'' \rangle]\right)$

- Пусть задана эластичная деформация растра $\boldsymbol{t} \to \boldsymbol{t} + \boldsymbol{x_t}$ (см. рис. 4.16). Эластичная потенциальная функция:
$$K(\boldsymbol{y}', \boldsymbol{y}'') = \sum_{\boldsymbol{t} \in T} y'_{\boldsymbol{t}} y'_{\boldsymbol{t}+\boldsymbol{x_t}}$$

Построение функции K с использованием метрики

- Пусть задано метрическое пространство A с метрикой $\rho : A \times A \to \mathbb{R}$:
  1. $\rho(\alpha_1, \alpha_2) \geq 0, \ = 0 \Leftrightarrow \alpha_1 = \alpha_2$
  2. $\rho(\alpha_1, \alpha_2) = \rho(\alpha_2, \alpha_1)$
  3. $\rho(\alpha_1, \alpha_3) \leq \rho(\alpha_1, \alpha_2) + \rho(\alpha_2, \alpha_3)$
- Общностью двух элементов из A относительно некоторого центра $\phi \in A$ назовем следующую величину:
$$\mu_\phi(\alpha_1, \alpha_2) = \frac{1}{2}\left[\rho^2(\alpha_1, \phi) + \rho^2(\alpha_2, \phi) - \rho^2(\alpha_1, \alpha_2)\right]$$
- Свойства общности
  1. $\mu_\phi(\alpha_1, \alpha_2) = \mu_\phi(\alpha_2, \alpha_1)$
  2. $\mu_\phi(\alpha, \alpha) \geq 0 \ = 0 \Leftrightarrow \alpha = \phi$
  3. $\forall \alpha \in A \ \mu_\phi(\alpha, \phi) = 0$
  4. $\mu_\phi(\alpha, \alpha) = \rho^2(\alpha, \phi)$
  5. $\rho(\alpha_1, \alpha_2) = (\mu_\phi(\alpha_1, \alpha_1) + \mu_\phi(\alpha_2, \alpha_2) - 2\mu_\phi(\alpha_1, \alpha_2))^{1/2}$
  6. $\mu_\phi(\alpha_1, \alpha_2) \leq \frac{1}{2}[\mu_\phi(\alpha_1, \alpha_1) + \mu_\phi(\alpha_2, \alpha_2)]$
  7. $|\mu_\phi(\alpha_1, \alpha_2)| \leq \sqrt{\mu_\phi(\alpha_1, \alpha_1)}\sqrt{\mu_\phi(\alpha_2, \alpha_2)}$
  8. $\mu_{\tilde{\phi}}(\alpha_1, \alpha_2) = \mu_\phi(\alpha_1, \alpha_2) - \mu_\phi(\alpha_1, \tilde{\phi}) - \mu_\phi(\alpha_2, \tilde{\phi}) + \mu_\phi(\phi, \tilde{\phi})$
- По своим свойствам общность очень похожа на скалярное произведение!

Матрица общностей

- Пусть $\{\alpha_1, \ldots, \alpha_q\} \subset A$ — конечная совокупность элементов метрического пространства с центром $\phi \in A$. Составим матрицу общностей этих элементов:
$$M_\phi = (\mu_\phi(\alpha_i, \alpha_j), i, j = 1, \ldots, q)$$
- Матрица общностей является симметричной, диагональные элементы неотрицательны.
- В отличие от матрицы скалярных произведений, которая всегда неотрицательно определена, матрица общностей может иметь собственные значения любого знака

Теорема 4. *Если M_ϕ является неотрицательно определенной для любой конечной совокупности элементов, то матрица $M_{\tilde{\phi}}$ относительно другого центра $\tilde{\phi}$ также является неотрицательно определенной.*

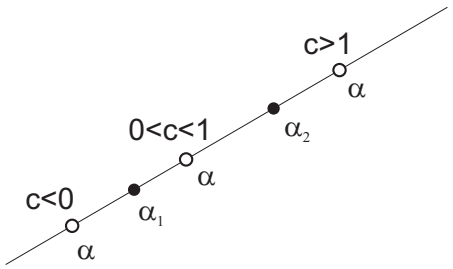


Рис. 4.17. Иллюстрация к понятию соосных элементов

Соосность элементов

Элемент $\alpha \in A$ называется **соосным элементом** для упорядоченной пары $[\alpha_1, \alpha_2]$ с коэффициентом $c \in \mathbb{R}$ и обозначается $\alpha = \text{coax}([\alpha_1, \alpha_2]; , c)$, если

$$\rho(\alpha_1, \alpha) = |c|\rho(\alpha_1, \alpha_2), \quad \rho(\alpha_2, \alpha) = |c-1|\rho(\alpha_1, \alpha_2)$$

Очевидно, что $\alpha_1 = \text{coax}([\alpha_1, \alpha_2]; 0)$, $\alpha_2 = \text{coax}([\alpha_1, \alpha_2]; 1)$. Если $\alpha = \text{coax}([\alpha_1, \alpha_2]; c)$, то $\alpha = \text{coax}([\alpha_2, \alpha_1]; 1-c)$. Для элементов $[\alpha_1, \alpha_2]$ и $\alpha = \text{coax}([\alpha_1, \alpha_2]; c)$ неравентсво треугольника переходит в равенство:

$$\begin{aligned}
&\rho(\alpha_1, \alpha_2) + \rho(\alpha_2, \alpha) = \rho(\alpha_1, \alpha), \text{ если } c > 1 \\
&\rho(\alpha_1, \alpha) + \rho(\alpha, \alpha_2) = \rho(\alpha_1, \alpha_2), \text{ если } 0 < c \leq 1 \\
&\rho(\alpha, \alpha_1) + \rho(\alpha_1, \alpha_2) = \rho(\alpha, \alpha_2), \text{ если } c \leq 0
\end{aligned}$$

Введение линейных операций

Метрическое пространство A называется **евклидовым метрическим пространством**, если $\forall [\alpha_1, \alpha_2],\ \alpha_1, \alpha_2 \in A$ и $\forall c \in \mathbb{R}\ \exists \alpha \in A : \alpha = \text{coax}([\alpha_1, \alpha_2]; c)$ и матрица общности всякой конечной совокупности элементов из A является неотрицательно определенной.

Введем операции суммы и умножения на число:

$$c\alpha \triangleq \text{coax}([\phi, \alpha]; c) \quad \alpha_1 + \alpha_2 = 2\,\text{coax}\left([\alpha_1, \alpha_2]; \frac{1}{2}\right)$$

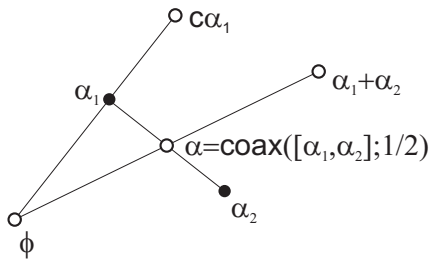


Рис. 4.18. Введение линейных операций в евклидовом метрическом пространстве

Свойства введенных операций

Теорема 5. *В евклидовом метрическом пространстве введенные операции сложения и умножения на число, а также скалярное произведение понимаемое как общность элементов* $\langle \alpha_1, \alpha_2 \rangle = \mu_\phi(\alpha_1, \alpha_2)$ *удовлетворяют всем требованиям гильбертова пространства:*

- $\alpha_1 + \alpha_2 = \alpha_2 + \alpha_1,\ (\alpha_1 + \alpha_2) + \alpha_3 = \alpha_1 + (\alpha_2 + \alpha_3)$
- $\alpha + \phi = \alpha,\ c\phi = \phi$
- $\forall \alpha \exists (-\alpha) : (-\alpha) + \alpha = \phi$
- $c_1(c_2\alpha) = (c_1c_2)\alpha$
- $1\alpha = \alpha$
- $(c_1 + c_2)\alpha = c_1\alpha + c_2\alpha,\ c(\alpha_1 + \alpha_2) = c\alpha_1 + c\alpha_2$
- $\langle \alpha_1, \alpha_2 \rangle = \langle \alpha_2, \alpha_1 \rangle, \langle c_1\alpha_1 + c_2\alpha_2, \alpha_3 \rangle = c_1\langle \alpha_1,\ \alpha_3 \rangle + c_2\langle \alpha_2, \alpha_3 \rangle$
- $\langle \alpha, \alpha \rangle \geq 0,\ = 0 \Leftrightarrow \alpha = \phi$
- $\|\alpha_1 - \alpha_2\| = \sqrt{\langle \alpha_1 - \alpha_2, \alpha_1 - \alpha_2 \rangle} = \rho(\alpha_1, \alpha_2)$

Глава 5

Задачи выбора модели

В главе рассматриваются вопросы выбора модели при обучении ЭВМ. Изложена суть проблемы и ее методологический характер. Приведены многочисленые примеры задач выбора модели, с которыми приходится сталкиваться при решении конкретных прикладных проблем. Рассмотрено несколько общих (не зависящих от эвристик) методов выбора модели, проанализированы их преимущества и недостатки.

5.1 Ликбез: Оптимальное кодирование

Оптимальное кодирование

- Рассматривается задача кодирования алфавита $\mathcal{A}$ словами из алфавита $\mathcal{B}$, как правило содержащего меньше символов
- Пусть каждый символ $a \in \mathcal{A}$ встречается в текстах с вероятностью $p(a)$. Обозначим $l(a)$ длину его кодирования в $\mathcal{B}$
- Задача построить схему кодирования, обеспечивающую минимальную среднюю длину кодированных сообщений, т.е.

$$\mathbb{E}_{\mathcal{A}} l(a) = \sum_{a \in \mathcal{A}} p(a) l(a) \to \min$$

Теорема Шеннона

- Теорема Шеннона. Если кодируемые элементы a могут встречаться с разными вероятностями $p(a)$, то существует оптимальное кодирование данного множества элементов и длина описания элемента a равна $l(a) = -\log_B p(a)$
- Основание логарифма B — это мощность кодирующего алфавита. Если алфавит $\mathcal{B}$ состоит из двух символов (например, ноль и единица), то логарифм двоичный, а длина описания измеряется в битах
  Если логарифм натуральный, то длина описания измеряется в натах, хотя довольно трудно представить себе алфавит из 2.7182... символов :)
- Теорема Шеннона согласуется со здравым смыслом: чем чаще встречается символ, тем короче должна быть длина его описания, и наоборот
  Лень Морзе дорого обошлась (и обходится) человечеству из-за того, что пропусканая способность каналов связи оказалась ниже оптимальной

5.2 Постановка задачи выбора модели

5.2.1 Общий характер проблемы выбора модели

Определение модели

- Пусть $\mathcal{A}(\boldsymbol{w})$ — решающее правило, полученное в результате настройки весов $\boldsymbol{w} \in \Omega$ в ходе обучения
- Модель — это совокупность всех решающих правил, которые получаются путем присваивания весам всех возможных допустимых значений $\mathcal{A}(\Omega)$
- Модель определяется множеством допустимых весов Ω, априорными распределениями весов на этом множестве $P(\boldsymbol{w})$ и структурой решающего правила $\mathcal{A}(.)$

Задача выбора модели

- Выбрать модель — значит определить множество Ω, указать на нем априорные вероятности весов $P(\boldsymbol{w})$ и определить структуру (схему, по которой настраиваемые веса будут образовывать композицию с входными данными) решающего правила $\mathcal{A}$
- Поскольку рассмотреть всевозможные множества, распределения и структуры невозможно, их обычно ограничивают некоторым параметрическим семейством, зависящим от **структурных параметров**
- Под задачей выбора модели будем понимать проблему автоматической настройки всех структурных параметров для данного алгоритма машинного обучения

Смысл проблемы выбора модели

- Нетрудно придумать алгоритм, блестяще работающий на обучающей выборке (например, запомнить для нее правильные ответы). Вот только на новых объектах такой алгоритм, скорее всего, будет работать плохо
- Возникает идея ограничить множество допустимых решающих правил, чтобы в процессе обучения мы не могли бы получить «плохие» решения
- Ценой неизбежно становится ухудшение работы алгоритма на обучающей выборке
- **Процесс выбора модели в машинном обучении — это поиск компромисса между точностью на обучении и его «надежностью» на произвольных объектах генеральной совокупности**

Философский смысл проблемы выбора модели

- Данная проблема проявляется в различных областях математики в частности, и человеческой деятельности вообще
- Не будет преувеличением сказать, что она носит общенаучный и даже философский характер
- Это проблема выбора средств: гайку можно выточить на старом добром станке (в котором из механизмов только электродвигатель времен Александра II), а можно использовать новейшее оборудование с ЧПУ (три микропроцессора, система калибровки, связанная со спутником и 20 кг технической документации)
- Гайка, сделанная на новейшем станке, будет иметь более высокое качество. Разумеется, если не сломается система калибровки, не сгорит хотя бы один из трех процессоров, и за станок не сядет какой-нибудь «естествоиспытатель»...

5.2.2 Примеры задач выбора модели

Задача классификации методом опорных векторов

- Решающее правило имеет вид

$$y(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{i=1}^{n} w_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b\right)$$

- Оптимизационная задача для поиска весов

$$\sum_{i=1}^{n} w_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} t_i t_j w_i w_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) \to \max$$
$$\sum_{i=1}^{n} t_i w_i = 0 \quad 0 \le w_i \le C$$

- Коэффициент регуляризации C и ядровая функция $K(\boldsymbol{x}', \boldsymbol{x}'')$ определяют модель метода опорных векторов (см. рис. 5.1, 4.10, 4.11)

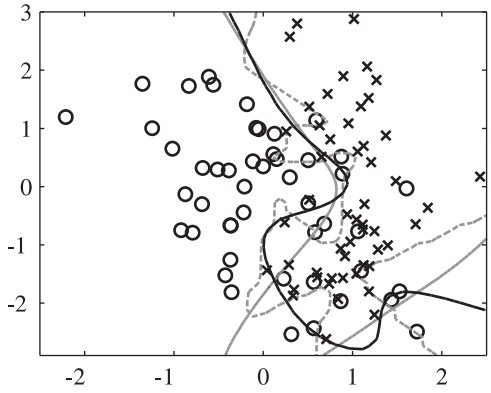


Рис. 5.1. Результаты классификации SVM с различными структурными параметрами

Задача регрессии

- Обобщенная линейная регрессия

$$y(\boldsymbol{x}) = \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x})$$

- Веса регрессии вычисляются по следующей формуле

$$\boldsymbol{w} = \left(\Phi^T \Phi + \lambda I\right)^{-1} \Phi^T \boldsymbol{t}$$

- Параметр регуляризации $\lambda \geq 0$, система базисных функций $\{\phi_j(\boldsymbol{x})\}_{j=1}^{m}$ и их количество m определяют модель

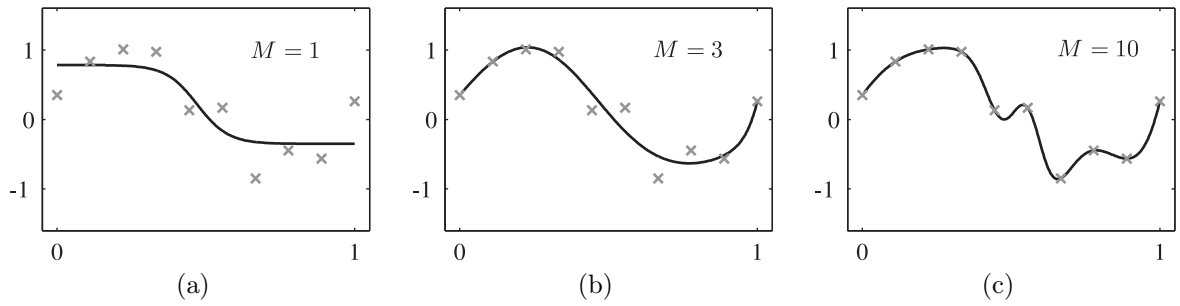
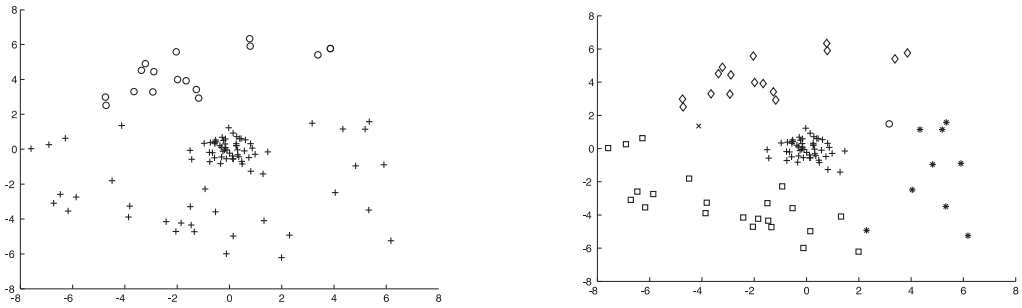


Рис. 5.2. Результаты восстановления регрессии с различным числом базисных функций

Задача кластеризации

- Большинство методов кластеризации предполагают задание пользователем количества кластеров, на которые будут разбиваться входные данные (см. рис. 5.3)



(a) (b)

Рис. 5.3. Решение задачи кластеризации с помощью метода K средних с двумя кластерами (a) и с шестью кластерами (b)

Нейронные сети

- Выбор архитектуры нейронной сети (количество нейронов на каждом уровне) и функции активации определяют модель нейронной сети

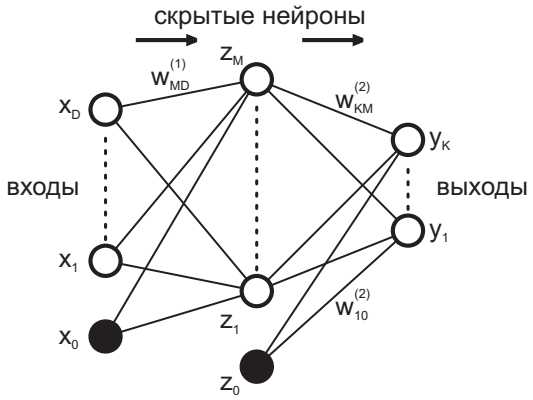


Рис. 5.4. Архитектура нейронной сети. Качества классификации нейронной сети существенно зависит от выбора ее структуры

5.3 Общие методы выбора модели

5.3.1 Кросс-валидация

Что такое общие методы выбора модели?

- Под общими методами выбора модели будем понимать алгоритмы, позволяющие проводить автоматическую настройку **любых** структурных параметров для широкого множества задач машинного обучения (например, для всех задач классификации)

- На практике такой метод может использоваться для настройки лишь некоторых структурных параметров, но гипотетически он должен позволять определять любые характеристики наилучшей модели

Скользящий контроль

- Процедура скользящего контроля (кросс-валидации) заключается в последовательном исключении части объектов из обучающей выборки, обучении на оставшихся объектах и распознавании исключенных объектов
- Тем самым эмулируется наличие тестовой выборки, которая не участвует в обучении, но для которой известны правильные ответы
- Структурные параметры настраиваются путем минимизации ошибки на скользящем контроле
- Процедура скользящего контроля является на сегодняшний день **самым лучшим средством настройки структурных параметров**

Схема k-fold cross validation

- Выборка разбивается на k непересекающихся (одинаковых по объему) частей. На каждой итерации обучение проводится по $k-1$ части, а тестирование на исключенных объектах

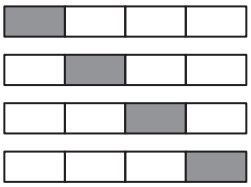


Рис. 5.5. Процедура 4-fold cross validation. Каждый раз из выборки исключается четверть объектов, на остальной части проводится обучение, а затем исключенные объекты распознаются

- На рисунке 5.5 приведена процедура 4-fold cross validation
- При $k = n$ процедура называется leave-one-out
- Наилучшим режимом скользящего контроля считается 5×2-fold cross validation.

Особенности скользящего контроля

- Ошибка на скользящем контроле является довольно точной оценкой ошибки на генеральной совокупности (обобщающей способности)
- Проведение скользящего контроля требует значительного времени на многократное повторное обучение алгоритмов и применимо лишь для «быстрых» методов машинного обучения
- С помощью скользящего контроля можно настраивать не более двух-трех структурных параметров, т.к. настройка производится путем **полного перебора** всевозможных сочетаний параметров
- При его использовании для выбора модели ошибку на скользящем контроле **нельзя** рассматривать как оценку ошибки на генеральной совокупности, т.к. она получается заниженной
- Скользящий контроль неприменим в задачах кластерного анализа и прогнозирования временных рядов

5.3.2 Теория Вапника-Червоненкиса

Идея теории структурной минимизации риска

- Теория Вапника-Червоненкиса использует косвенные характеристики для оценки обобщающей способности (среднего риска)
- Ключевым понятием является т.н. емкость (размерность Вапника-Червоненкиса, VC-dimension) модели
- Идея данного подхода к выбору модели (структурной минимизации риска) заключается в следующем: **Чем более «гибкой» является модель, тем хуже ее обобщающая способность**
- В самом деле, «гибкое» решающее правило способно настроиться на малейшие шумы, содержащиеся в обучающей выборке

Понятие емкости

- Рассмотрим задачу классификации на два класса
- (Несколько упрощая,) емкостью данной модели будем называть максимальное число объектов обучающей выборки, для которых при **любой** их разметке на классы найдется хотя бы один алгоритм из модели, безошибочно их классифицирующий
- По аналогии вводятся определения емкости для других задач машинного обучения
- Важный пример модели, для которой известна емкость — классификатор, строящий линейную гиперплоскость. Емкость линейного классификатора равна $h = d+1$, где d — размерность пространства признаков
- Следствие: $n \leq d + 1$ объектов **всегда** можно безошибочно разделить гиперплоскостью

Формула Вапника

- Очевидно, что чем больше емкость, тем хуже. Значит нужно добиваться минимально возможного количества ошибок на обучении при минимальной возможной емкости
- Ошибку на обучении (эмпирический риск) $P_{train}(\boldsymbol{w})$, емкость $h(\Omega)$ и ошибку на генеральной совокупности (средний риск) $P_{test}(\boldsymbol{w})$ связывает известная формула Вапника

$$P_{test}(\boldsymbol{w}) \leq P_{train}(\boldsymbol{w}) + \sqrt{\frac{h(\Omega)(\log(2n/h(\Omega)) + 1) - \log(\eta/4)}{n}}$$

  Неравенство верно с вероятностью $1 - \eta$ для $\forall \boldsymbol{w} \in \Omega$
- Последовательно анализируя модели с увеличивающейся емкостью, согласно теории ВЧ, необходимо выбирать модель с наименьшей верхней оценкой тестовой ошибки

Достоинства и недостатки теории ВЧ

- Достоинства
  - Серьезное теоретическое обоснование, связь с ошибкой на генеральной совокупности
  - Теория продолжает развиваться и в наши дни (эффективная емкость, локальная емкость, комбинаторный подход и т.д.)
- Недостатки
  - Оценки сильно завышены
  - Для большинства моделей емкость не поддается оценке
  - Многие модели с бесконечной емкостью показывают хорошие результаты на практике

Пути развития теории ВЧ (и не только ее)

- Емкость вводится для всевозможных положений объектов выборки, в то время как реально приходится иметь дело с одной конкретной выборкой и опять-таки имеет смысл рассматривать степень адаптируемости модели под эту конкретную выборку, а не под абстрактно возможную
- В процессе обучения поиск алгоритма в модели ведется не по всем ее представителям, а лишь по конечному числу, которое и имеет смысл рассматривать при интерпретации емкости, как степени адаптируемости модели под данные
- Искомая закономерность может обладать рядом дополнительных свойств, которые сокращают объем допустимых алгоритмов модели

5.3.3 Принцип минимальной длины описания

Предпосылка метода

- Из пункта A в пункт B передается закодированное сообщение о классификации обучающей выборки. Нужно добиться минимально возможного размера сообщения
- Стратегия 1: передаем каждый объект и его метку класса $\{(\boldsymbol{x}_i, t_i)\}_{i=1}^n$
- Стратегия 2: передаем длинное описание сложного алгоритма, который можно использовать для правильной классификации всей обучающей выборки $Descr(A)$
- Стратегия 3: передаем короткое описание простого алгоритма $Descr(A')$, который правильно классифицирует большинство объектов обучающей выборки, а классификацию неправильно распознанных объектов передаем отдельным списком$\{\boldsymbol{x}_{i_k}, t_{i_k}\}_{k=1}^p$, $p < n$

Смысл метода

- Чем точнее на обучающей выборке алгоритм, тем он сложнее, а значит тем длиннее будет его описание...
- ... но тем меньше будет список неправильно распознанных объектов (см. рис. 5.6)
- Принцип минимальной длины описания (minimum decription length MDL, Rissanen, 1978) штрафует излишнюю алгоритмическую сложность решающего правила

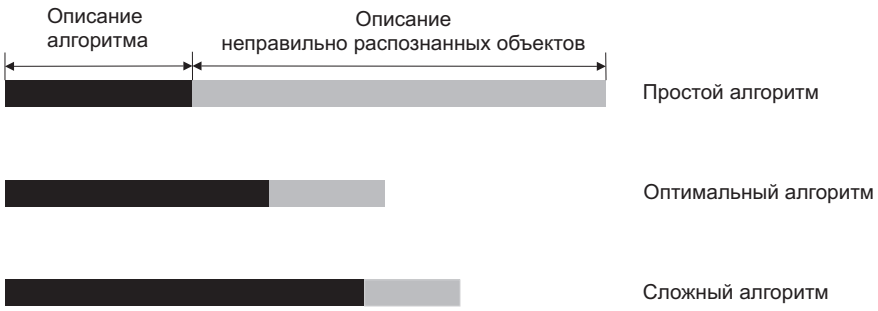


Рис. 5.6. Иллюстрация метода минимальной длины описания

Особенности подсчета длины описания

- Существует множество подходов к оценке длины описания алгоритма вплоть до длины кода реализующей его программы
- Необходимо отметить, что кодирование должно быть эффективным, т.к. даже самый простой алгоритм можно закодировать в очень длинное сообщение
- Согласно теореме Шеннона, при оптимальном кодировании длина описания структуры пропорциональна логарифму ее вероятности, взятому с противоположным знаком

Эквивалентность MDL и максимизации апостериорной вероятности

- Пусть на множестве алгоритмов задано априорное распределение $p(\boldsymbol{w})$

$$l(\boldsymbol{w}) = -\log p(\boldsymbol{w})$$

- Длина описания данных тем меньше, чем выше вероятность данной классификации при использовании данного алгоритма, т.е. чем выше правдоподобие $p(\boldsymbol{t}|X, \boldsymbol{w})$

$$l(\boldsymbol{t}|\boldsymbol{w}) = -\log p(\boldsymbol{t}|X, \boldsymbol{w})$$

- Отсюда получаем выражение, объединяющее точность на обучении и сложность алгоритма **в единое выражение**

$$l(\boldsymbol{t}, \boldsymbol{w}) = -\log p(\boldsymbol{t}|X, \boldsymbol{w}) - \log p(\boldsymbol{w})$$
$$\arg\min_{\boldsymbol{w}} l(\boldsymbol{t}, \boldsymbol{w}) = \arg\max_{\boldsymbol{w}} p(\boldsymbol{t}|X, \boldsymbol{w}) p(\boldsymbol{w})$$

- Таким образом, MDL обосновывает идею максимизации регуляризованного правдоподобия

Отличительные особенности MDL

- MDL позволяет обосновать корректность регуляризации правдоподобия
- Область применения MDL шире, чем у статистических методов обучения, т.е. MDL можно применять и там, где вводить вероятности некорректно или бессмысленно
- При использовании MDL предполагается, **что чем сложнее алгоритм, тем хуже его обобщающая способность.** Современные исследования (в частности, boosting) показывают, что это далеко не всегда так

5.3.4 Информационные критерии

Информационный критерий Акаике

- В 1973г. Акаике установил связь между правдоподобием (ключевое понятие статистики) и дивергенцией Кульбака-Лейблера (ключевое понятие в теории информации)
- Ему удалось получить приблизительное соотношение между правдоподобием генеральной совокупности и правдоподобием обучающей выборки (т.е. данных, по которым с помощью ММП производится настройка параметров решающего правила)

$$AIC = \log p(\boldsymbol{t}|X, \boldsymbol{w}_{ML}) - M,$$

  где M — число настраиваемых параметров
- Пример использования: задача восстановления регрессии с известным гауссовским шумом в одномерном пространстве при помощи полинома степени k

$$k = \arg\min \left(\frac{\sum_{i=1}^{n} (t_i - y_k(\boldsymbol{x}_i))^2}{2\sigma^2} + k + 1 \right)$$

Информационный критерий Шварца

- Критерий Шварца (часто именуемый Байесовским информационным критерием) представляет собой простейшее приближение обоснованности, широко использующейся в байесовском обучении

$$BIC \approx \int p(\boldsymbol{t}|X, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$

- Используя приближение интеграла гауссианой и сильно огрубляя, получаем

$$BIC = \log p(\boldsymbol{t}|X, \boldsymbol{w}_{MP}) - \frac{1}{2}M \log n$$

- Пример использования: Задача восстановления регрессии с известным гауссовским шумом в одномерном пространстве при помощи полинома степени k

$$k = \arg\min \left(\frac{\sum_{i=1}^{n}(t_i - y_k(\boldsymbol{x}_i))^2}{2\sigma^2} + (k+1)\log\sqrt{n} \right)$$

Особенности информационных критериев

- Оба критерия являются (весьма грубыми) приближениями более сложных выражений, часто не поддающихся аналитическому вычислению. Значение критерия может расцениваться лишь как приблизительная характеристика обобщающей способности полученного решающего правила
- Критерии разумно использовать когда **все M настраиваемых параметров оказывают примерно одинаковое влияние** на вид решающего правила, например, входят в него линейно.
- Байесовский критерий сильнее штрафует сложные (с точки зрения дополнительных параметров) модели
- В байесовский критерий входит значение правдоподобия в точке $\boldsymbol{w}_{MP}$, а в критерий Акаике — в точке $\boldsymbol{w}_{ML}$

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

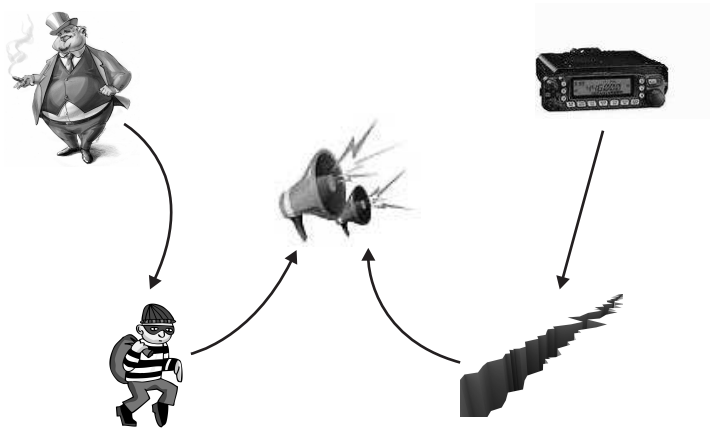
- Это новый подход к синтезу экспертных систем
- В отличие от нечеткой логики, он теоретически обоснован и математически корректен

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)

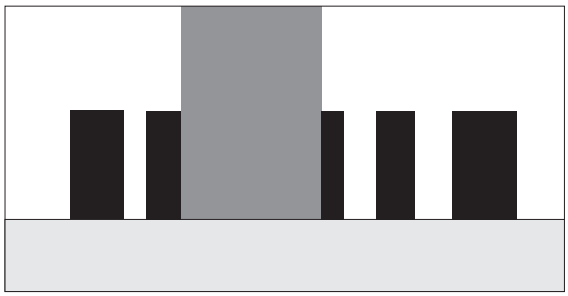


Рис. 7.1. Сколько ящиков за деревом?

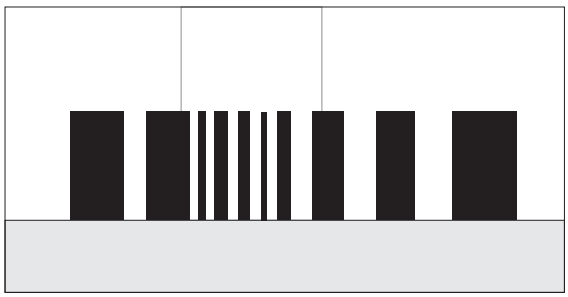


Рис. 7.2.

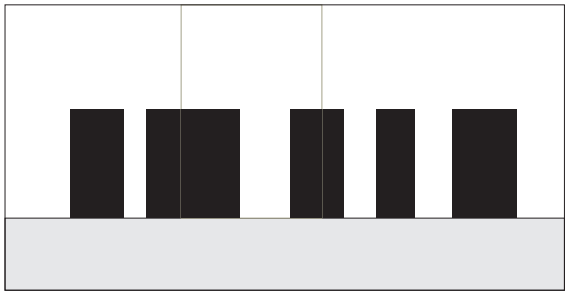


Рис. 7.3.

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

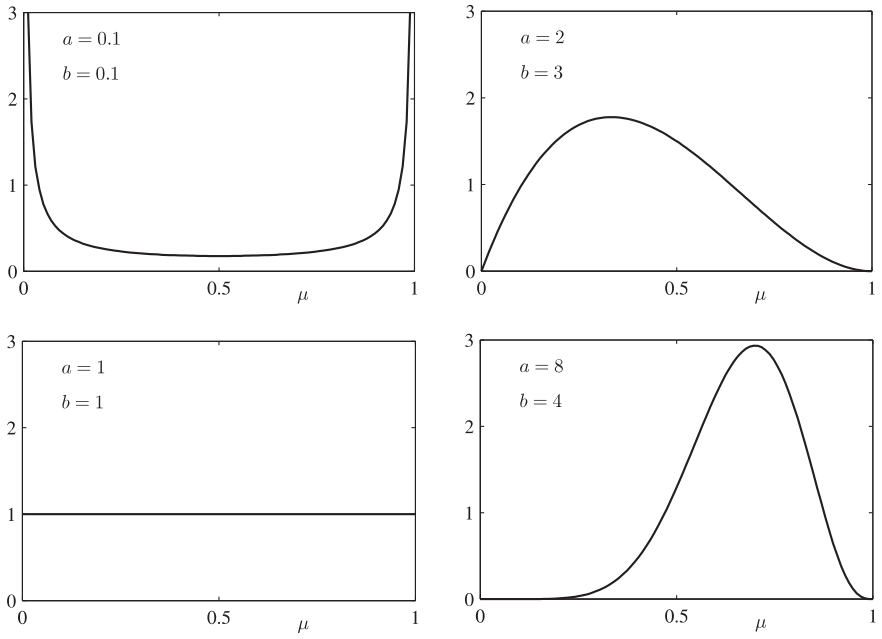


Рис. 7.4. Различные формы бета-распределения

- 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

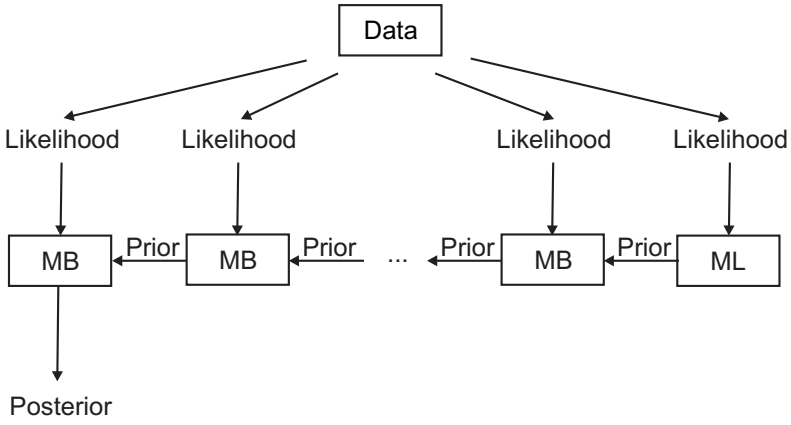


Рис. 7.5. Иерархическая схема байесовского вывода

- 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

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

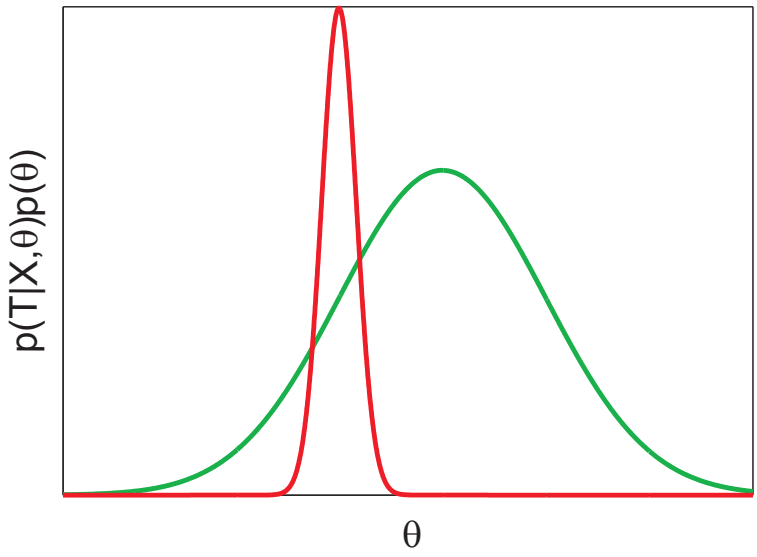


Рис. 7.6. В «пунктирной» модели присутствует небольшая доля алгоритмов, которые прекрасно объясняют обучающую выборку. В то же время, «сплошная» модель является более обоснованной, т.к. доля «хороших» алгоритмов в ней велика

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

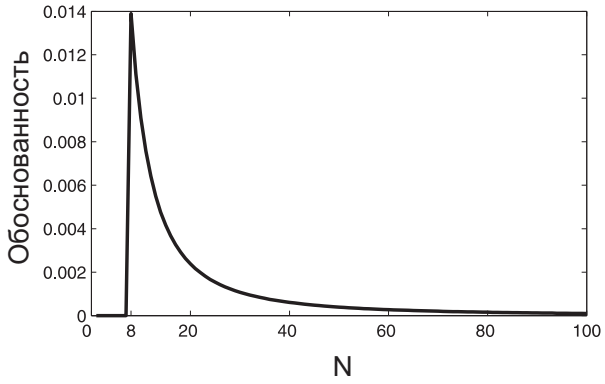


Рис. 7.7. При $N < 8$ обоснованность равна нулю, т.к. получить выборку $x_1 = 8$, $x_2 = 6$, применяя такой датчик, невозможно (функция правдоподобия всюду будет равна нулю). При больших N обоснованность падает, т.к. такие модели способны объяснить не только наши наблюдения, но и «много чего еще»

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_*.

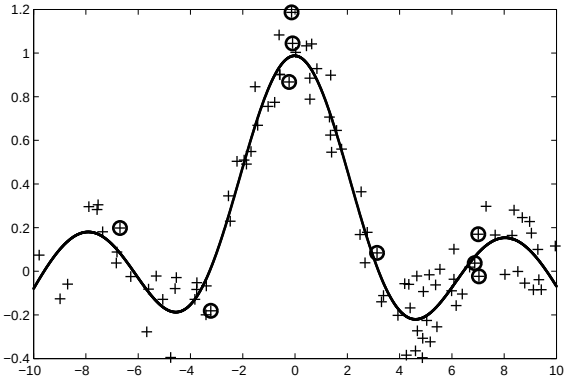


Рис. 8.1. Пример применения регрессии релевантных векторов для зашумленной функции $y(x) = \mathrm{sinc}(x)$. В качестве базисных функций использовались $\phi_j(\boldsymbol{x}) = \exp(-\beta\|\boldsymbol{x} - \boldsymbol{x}_j\|^2)$. Объекты, соответствующие релевантным базисным функциям, обведены в кружочки. 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.

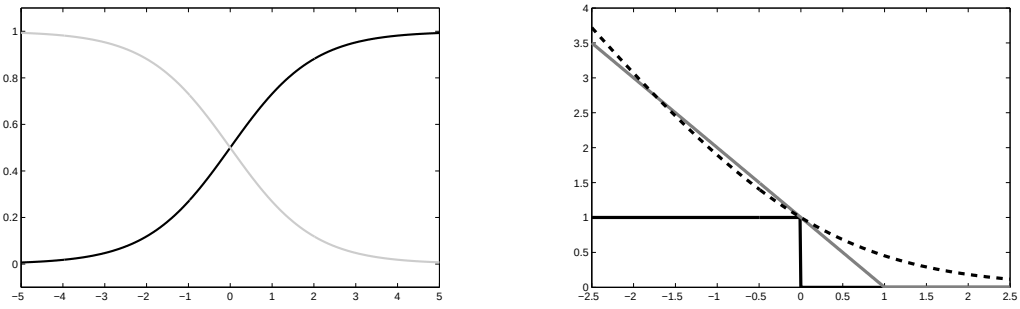
Метод максимума правдоподобия (логистическая регрессия)

- 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

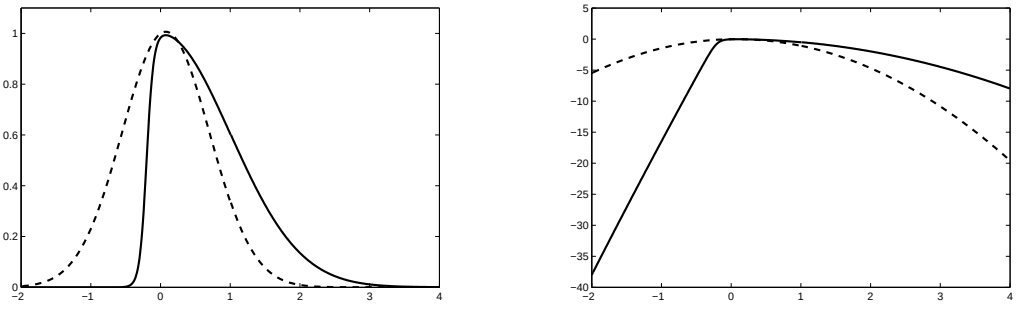
- Для получения разреженного решения введем в качестве априорного распределения на параметры $\boldsymbol{w}$ нормальное распределение с диагональной матрицей ковариации **с различными элементами на диагонали**:
$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{w}|0, A^{-1})$$
Здесь $A = \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$.
- 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.

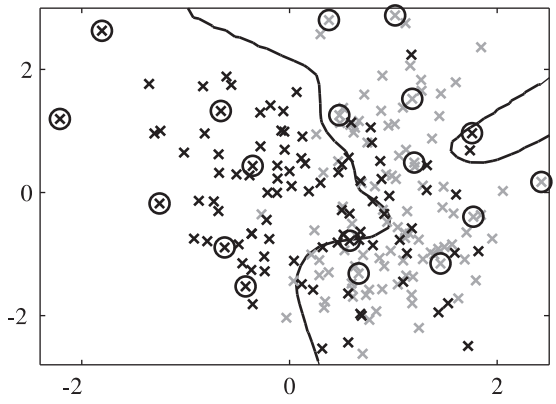


Рис. 8.4. Пример применения метода релевантных векторов для задачи классификации. В качестве базисных функций использовались $\phi_j(x) = \exp(-\beta(x-x_j)^2)$. Объекты, соответствующие релевантным базисным функциям, обведены в кружочки. В процессе обучения большинство α_j стремятся к $+\infty$. Таким 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

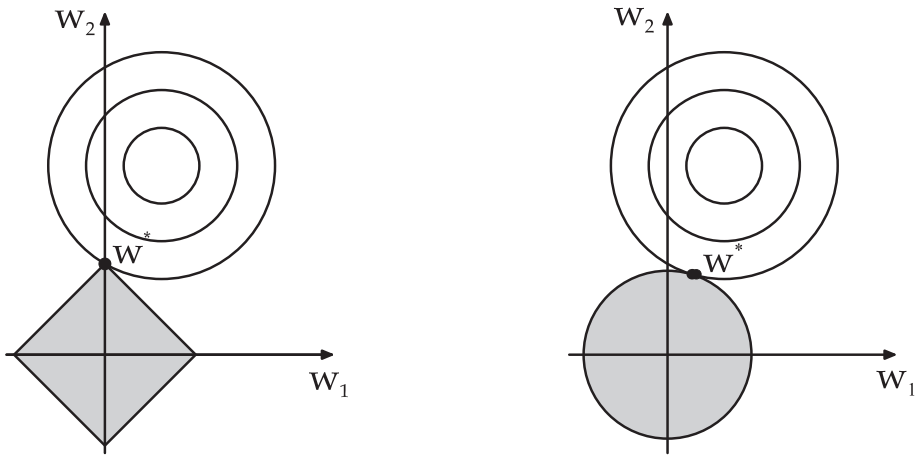


Рис. 9.1. Влияние вида регуляризатора на положение экстремума регуляризованной функции

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

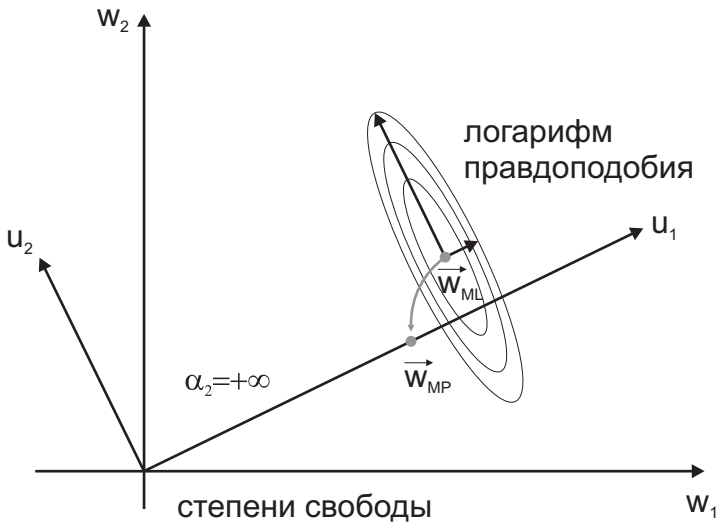


Рис. 9.2. Регуляризация степеней свободы, ассоциированных с собственными векторами гессиана правдоподобия

- 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 $\alpha_i = +\infty$)

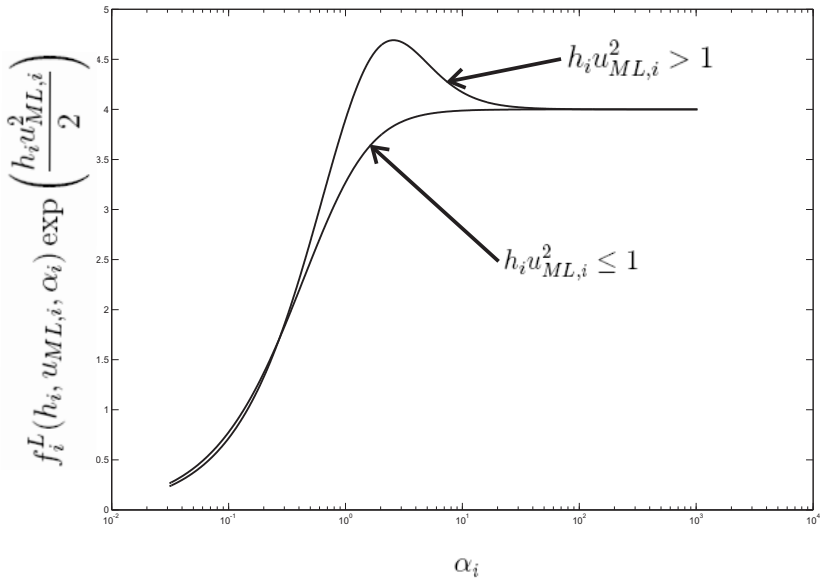


Рис. 9.3. Поведение функции $f_i^L(h_i, u_{ML,i}, \alpha_i)$ при разных значениях α_i

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

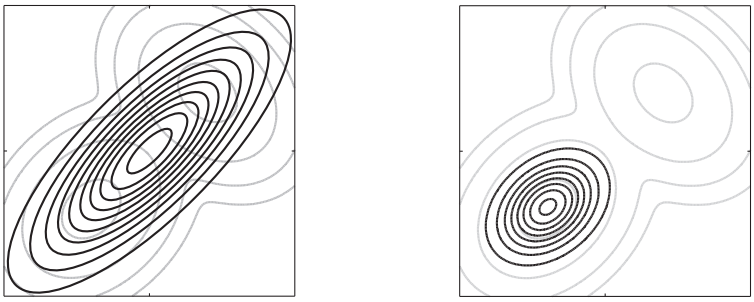
- Существует множество способов определить близость между вероятностными распределениями
- Рассмотрим распределения $p(\boldsymbol{x})$ и $q(\boldsymbol{x})$. Дивергенцией Кульбака-Лейблера называется величина

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

- 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.

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

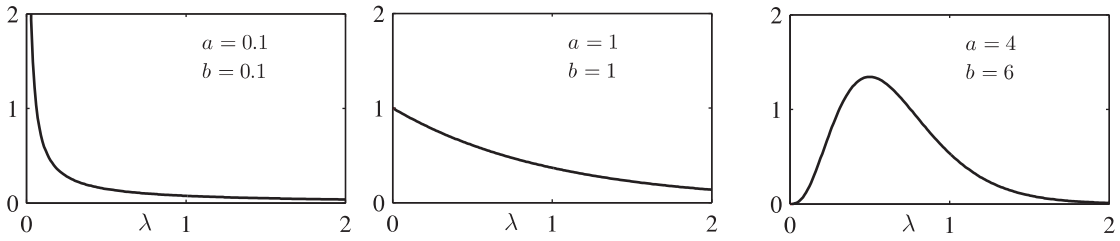


Рис. 11.2. График гамма-распределения с различными параметрами a и b

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

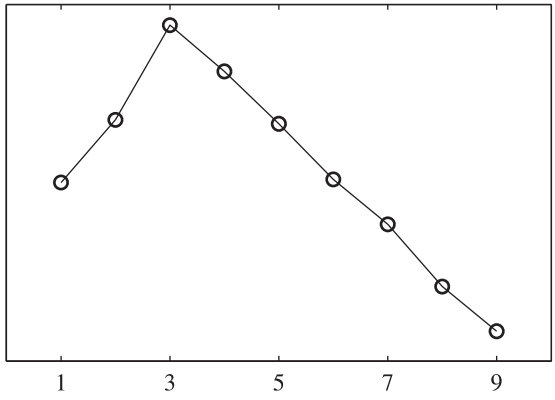


Рис. 11.3. На рисунке изображена зависимость $\mathcal{L}(q)$ от степени полинома для полиномиальной регрессии, построенной по зашумленной выборке, полученной с помощью кубического многочлена

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

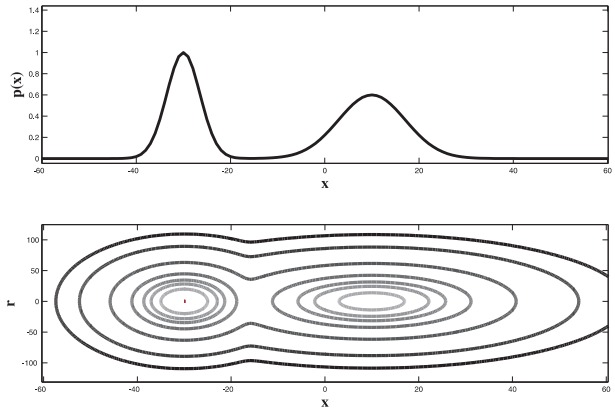


Рис. 11.4. Исходное распределение (вверху) и линии уровня соответствующего ему гамильтониана. Численное решение уравнений Гамильтона приводит к получению последовательности точек, находящихся на одной линии уровня

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

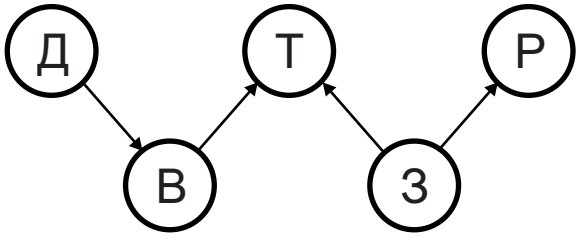


Рис. 12.1. Графическая модель, соответствующая примеру про Джона и колокольчик для воров (см. главу 6)

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

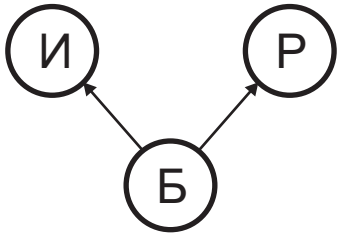


Рис. 12.2. Графическая модель «Варвары и Рим времен заката»

- 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

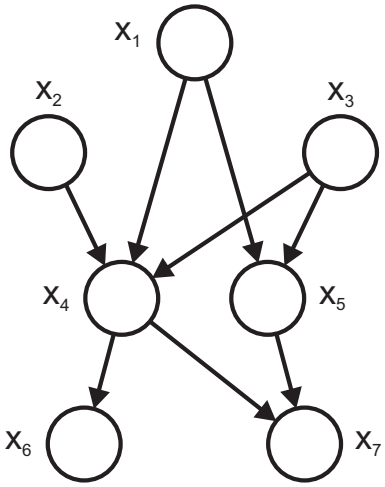


Рис. 12.3.

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)

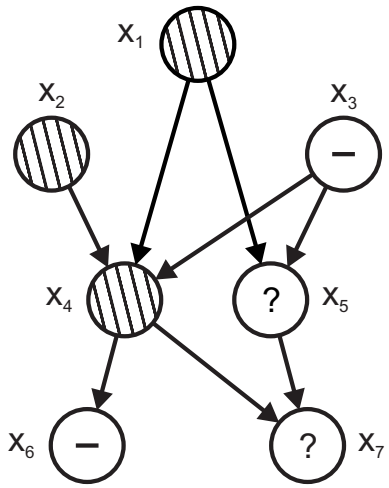


Рис. 12.4.

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

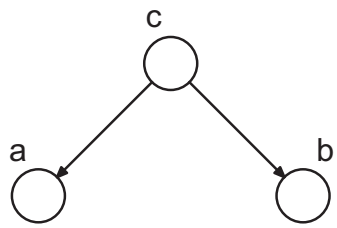


Рис. 12.5.

- 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

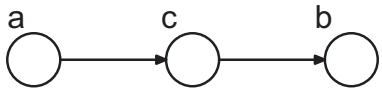


Рис. 12.6.

- 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)

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

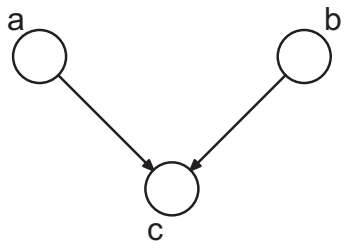


Рис. 12.7.

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)

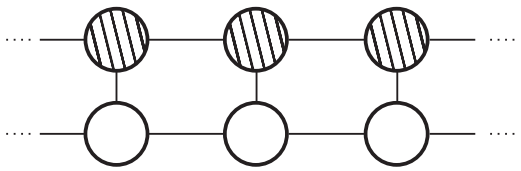
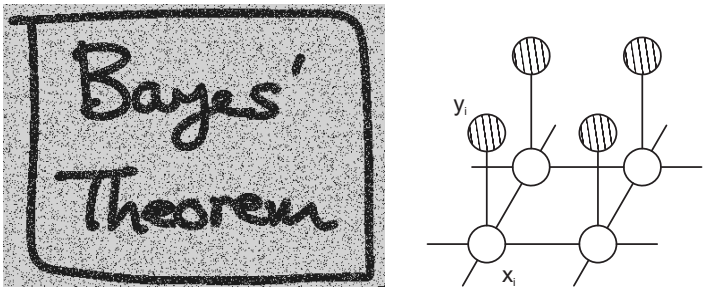


Рис. 12.8.

- 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 $\boldsymbol{t}$

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

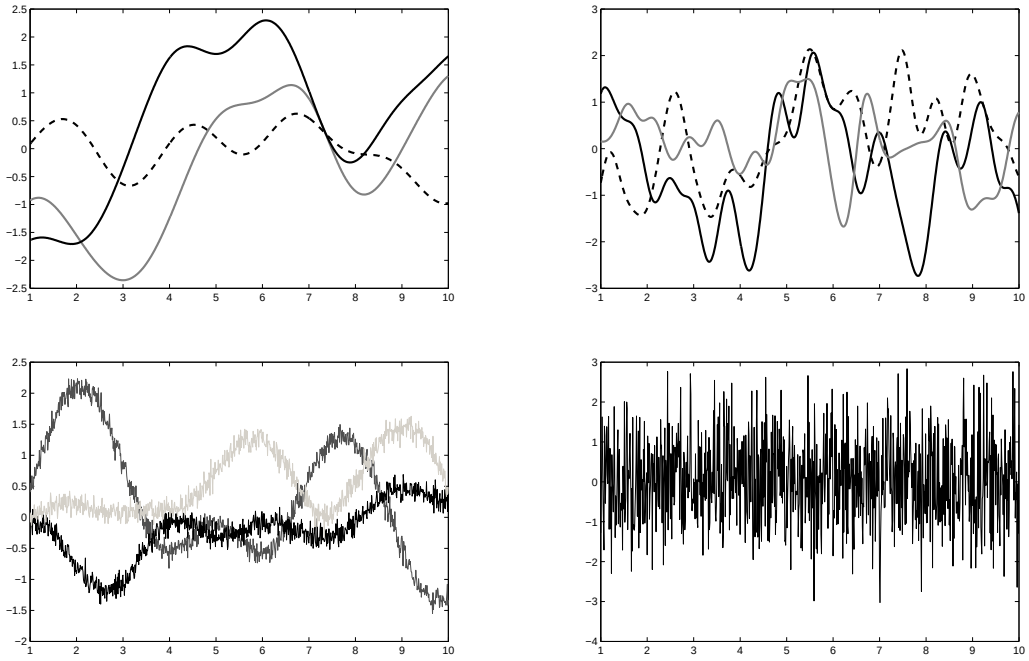


Рис. 12.10. Примеры реализаций стационарных гауссовских случайных процессов с различными ковариационными функциями

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

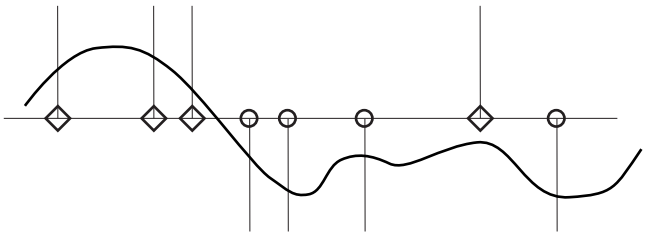


Рис. 12.11. При решении задачи классификации пользователю известен лишь знак реализации процесса в конечном числе точек

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