Лекция 5. Решение задачи выбора модели по Байесу. Обоснованность модели

Д. П. Ветров[1] Д. А. Кропотов[2]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

План лекции

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

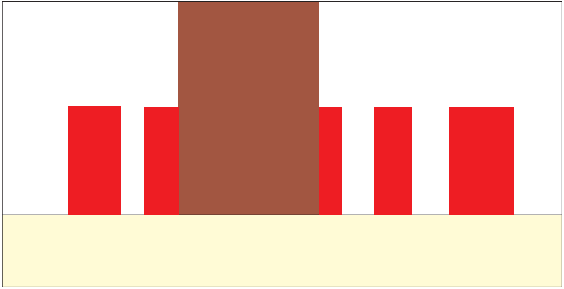
  Наполеон однажды спросил Лапласа (полушутя, полусерьёзно): «Что-то я не вижу в Вашей теории места для Бога». На что Лаплас, якобы, ответил: «Сир, у меня не было нужды в этой гипотезе».

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

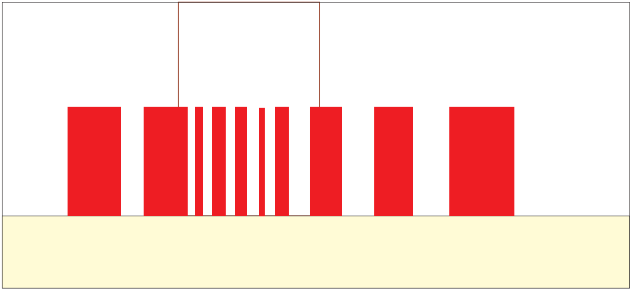
План лекции

Простой пример I



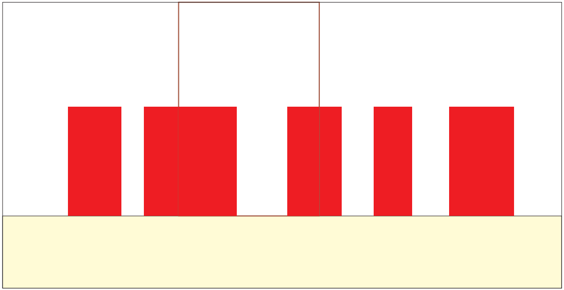
Сколько ящиков за деревом?

Простой пример II



С точки зрения максимума правдоподобия, любое количество ящиков одинаково приемлемо

Простой пример III



Наша же интуиция, а точнее априорные знания о характерной ширине ящика, базирующиеся на наблюдениях ящиков справа и слева, подсказывает иной ответ

План лекции

Получение апостериорного распределения

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

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

- 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

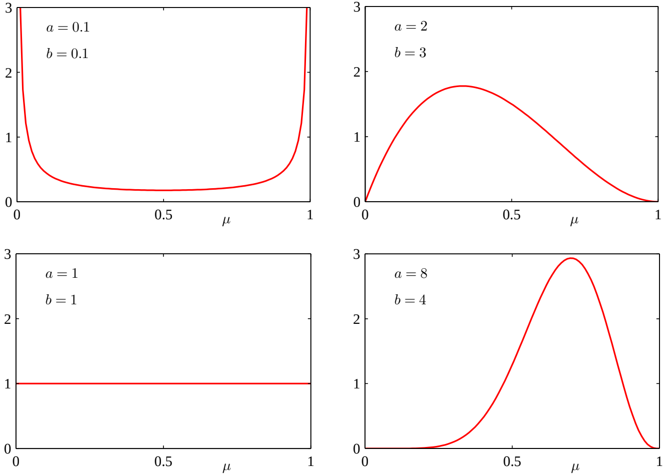
$$p(m|n, q) = C_n^m q^m (1-q)^{n-m} \sim \mathcal{B}(m|n, q)$$

- Сопряженным к распределению Бернулли является бета-распределение

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

- Легко показать, что интеграл от произведения распределения Бернулли и бета-распределения берется аналитически (Упр.)

Различные формы бета-распределения



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

Продолжение примера

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

План лекции

Байесовское обучение

- Рассмотрим дискриминативную модель $p(t, \theta|\mathbf{x}) = p(t|\mathbf{x}, \theta)p(\theta)$
- На этапе обучения нам задана обучающая выборка (X, T), которая играет роль наблюдаемых переменных, а настраиваемые параметры θ играют роль скрытых переменных
- В ходе обучения оценивается апостериорное распределение на параметры θ, т.е.

$$p(\theta|X, T) = \frac{p(T|X, \theta)p(\theta)}{p(T|X)} = \frac{\prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \theta)p(\theta)}{\int \prod_{i=1}^{n} p(t_i|\mathbf{x}_i, \theta)p(\theta)d\theta}$$

- Затем это распределение используется для подсчета величины

$$p(t_*|\mathbf{x}_*, X, T) = \int p(t_*|\mathbf{x}_*, \theta)p(\theta|X, T)d\theta$$

Последнее выражение представляет собой взвешенное голосование по всевозможным решающим правилам данной модели

Использование MAP-оценки

- Взвешенное голосование представляет собой интегрирование в многомерном пространстве параметров θ, которое не всегда удается провести аналитически
- Альтернативой является использование наиболее вероятного значения $\theta_{MP} = \arg\max p(\theta|X, T)$
- Это эквивалентно приближению апостериорного распределения дельта-функцией в точке θ_{MP}

$$p(t_*|\mathbf{x}_*, X, T) = \int p(t_*|\mathbf{x}_*, \theta)p(\theta|X, T)d\theta \approx$$

$$\approx \int p(t_*|\mathbf{x}_*, \theta)\delta_{\theta_{MP}}(\theta)d\theta = p(t_*|\mathbf{x}_*, \theta_{MP})$$

- Если апостериорное распределение унимодально и имеет острый пик, такое приближение осмысленно

План лекции

Выбор априорного распределения

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

$$p(\alpha|X, T) = \frac{p(T|X, \alpha)p(\alpha)}{p(T|X)} = \frac{p(T|X, \alpha)p(\alpha)}{\int p(T|X, \alpha)p(\alpha)d\alpha}$$

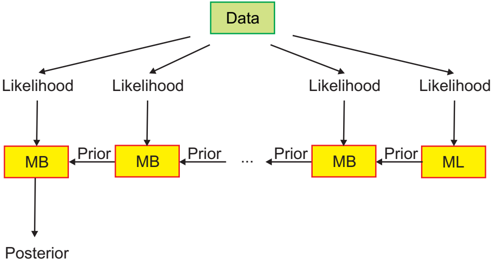
- В качестве правдоподобия относительно α выступает т.н. обоснованность $p(T|X, \alpha) = \int p(T|X, \theta)p(\theta|\alpha)d\theta$, полученная путем исключения (integrate out) переменной θ

Важное замечание

- Обратите внимание, что обоснованность является неполным правдоподобием, т.е. вероятностью наблюдаемых переменных, в качестве которых у нас выступает обучающая выборка $p(T|X,\alpha)$
- Мы не пытаемся максимизировать правдоподобие на все скрытые переменные $p(T|X,\theta,\alpha)$ сразу, т.к. это приведет к выбору параметров модели α, которые не ограничивают значения весов θ
- Таким образом, при автоматическом подборе априорного распределения на скрытые переменные, необходимо оптимизировать неполное правдоподобие наблюдаемых переменных относительно параметров априорного распределения
- Такой подход называется принципом максимальной обоснованности модели

Иерархическая схема байесовского вывода

- На параметры априорного распределения α может быть заданно свое априорное распределение, известное 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 или явно задать априорное распределение

План лекции

Обоснованность модели

- На практике обычно ограничиваются двумя уровнями вывода, применяя метод максимального правдоподобия для оценки параметров модели
- Параметры модели настраиваются путем максимизации неполного правдопдобия (обоснованности), поэтому применение ММП не приводит к переобучению
- Функция правдоподобия относительно параметров модели называется обоснованностью (evidence) модели

$$p(T|X,\alpha) = \int_{\Theta} p(T|X,\theta)p(\theta|\alpha)d\theta$$

- Гиперпараметры подбираются путем максимизации обоснованности

$$\alpha_{ME} = \arg\max p(T|X,\alpha)$$

- Далее находим апостериорное распределение $p(\theta|\alpha_{ME}, X, T)$ или его аргмаксимум θ_{MP}

Принцип наибольшей обоснованности с точки зрения байесовского подхода

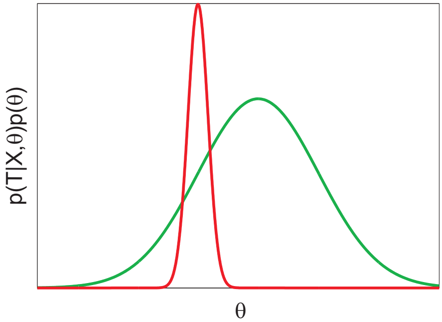
- Применение метода максимального правдоподобия на втором уровне байесовского вывода означает, что все модели для нас одинаково приемлемы, т.е. $p(\alpha) = Const$
- В этом случае легко показать, что $p(\alpha|X,T) \propto p(T|X,\alpha)$
- Строгое применение байесовского вывода предполагает взвешенное голосование по апостериорному распределению на (θ, α), т.е.

$$p(t_*|\boldsymbol{x}_*, X, T) = \int p(t_*|\boldsymbol{x}_*, \theta) p(\theta, \alpha|X, T) d\theta d\alpha$$

- С учетом $p(\theta, \alpha|X, T) = p(\theta|\alpha, X, T)p(\alpha|X, T) \approx p(\theta|\alpha_{ME}, X, T)\delta_{\alpha_{ME}}(\alpha)$ получаем

$$p(t_*|\boldsymbol{x}_*, X, T) \approx \int p(t_*|\boldsymbol{x}_*, \theta) p(\theta|\alpha_{ME}, X, T) d\theta$$

Иллюстрация понятия обоснованности



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

Разные варианты моделей

Пик	Высокий	Низкий
Узкий	Модель слишком сложная и помимо обучающей выборки может настроиться на много чего еще – велик риск переобучения. Обоснованность низкая	Модель неадекватна. Обоснованность очень низкая
Широкий	В рамках модели существует много решающих правил, хорошо объясняющих обучающую выборку, при этом способность к адаптации на произвольный набор входных данных ограничена. Обоснованность высокая	В рамках выбранной модели невозможно найти все закономерности между наблюдаемыми и скрытыми переменными. Как правило, это признак слишком простой модели. Обоснованность низкая.

План лекции

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

$$D(\boldsymbol{q}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha^0})} q_1^{\alpha_1^0 - 1} \dots q_N^{\alpha_N^0 - 1}, \quad \sum_{i=1}^N q_i = 1, \ q_i \geq 0$$

при $\alpha_1^0 = \dots = \alpha_N^0 = 1$

- Тогда применяя формулу Байеса, учитывая, что правдоподобие равно $p(x_1, x_2|\boldsymbol{q}) = q_8 q_6$, получаем

$$p(\boldsymbol{q}|x_1, x_2) = \frac{1}{Z} q_1^0 \dots q_5^0 q_6^1 q_7^0 q_8^1 q_9^0 \dots q_N^0 = D(\boldsymbol{q}|\boldsymbol{\alpha^1}),$$

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

$$q_6 = q_8 = \frac{2}{12} = \frac{1}{6} \approx 0.16 \quad q_i = \frac{1}{12} \approx 0.08 \quad \forall i \neq 6, 8$$

при $N = 100$ получаем

$$q_6 = q_8 = \frac{2}{102} = \frac{1}{51} \approx 0.02 \quad q_i = \frac{1}{102} \approx 0.01 \quad \forall i \neq 6, 8$$

- 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, $Ev(N = 10) > Ev(N = 100)$

Обоснованность модели при разных значениях N

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»