

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

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

Д. П. Ветров¹ Д. А. Кропотов²

¹МГУ, ВМиК, каф. ММП

²ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

План лекции

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

Ликбез

Полный байесовский вывод

Пример использования априорных знаний
Сопряженные распределения
Иерархическая схема Байеса

Принцип наибольшей обоснованности

Обоснованность модели
Примеры использования

Бритва Оккама

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

- 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

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

- 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

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения
Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности

Ликбез

Полный байесовский вывод

Пример использования априорных знаний

Сопряженные распределения

Иерархическая схема Байеса

Принцип наибольшей обоснованности

Обоснованность модели

Примеры использования

Простой пример I

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

Ветров,
Кропотов

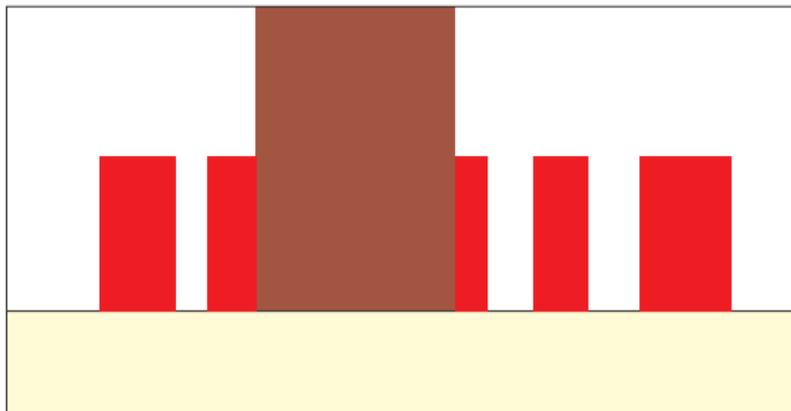
Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения
Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности



Сколько ящиков за деревом?

Простой пример II

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

Ветров,
Кропотов

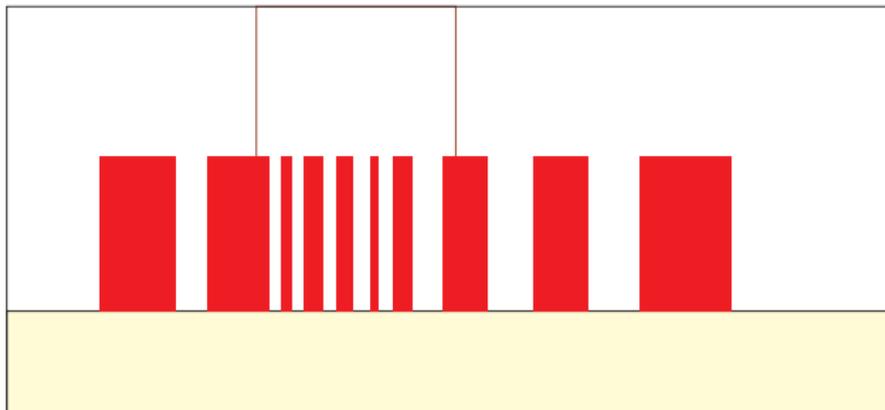
Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения
Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности



С точки зрения максимума правдоподобия, любое количество ящиков одинаково приемлемо

Простой пример III

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

Ветров,
Кропотов

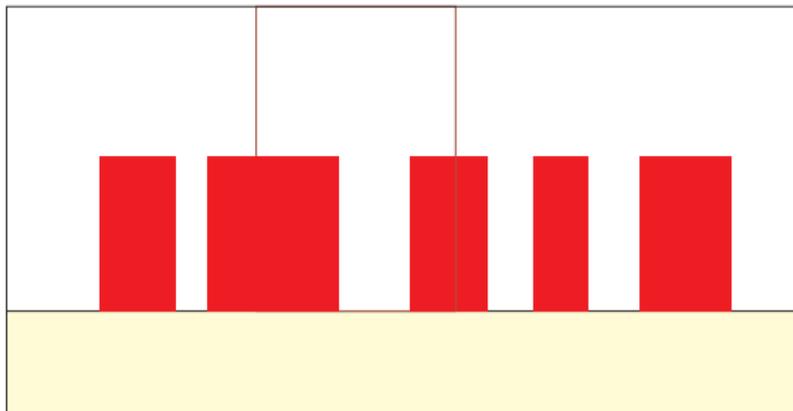
Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения
Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности



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

План лекции

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения

Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности

Ликбез

Полный байесовский вывод

Пример использования априорных знаний

Сопряженные распределения

Иерархическая схема Байеса

Принцип наибольшей обоснованности

Обоснованность модели

Примеры использования

Получение апостериорного распределения

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения

Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности

- 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

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения

Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности

- 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 $p(\mathbf{x}|\theta) \sim \mathcal{B}(\beta)$ являются сопряженными, если

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

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

Пример

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения

Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности

- 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.)

Различные формы бета-распределения

Лекция 5.

Решение задачи
выбора модели
по Байесу.
Обоснованность
модели

Ветров,
Кропотов

Ликбез

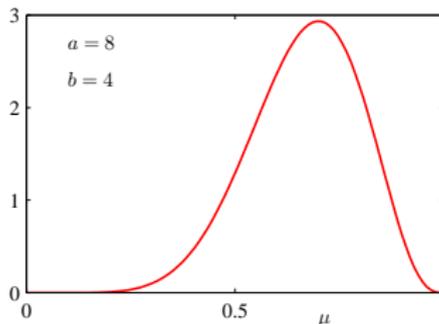
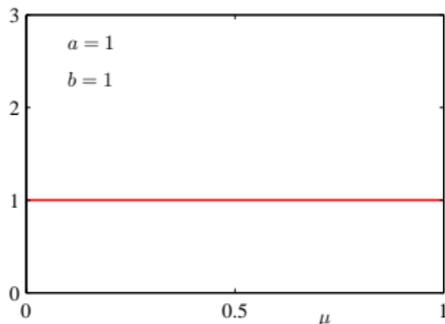
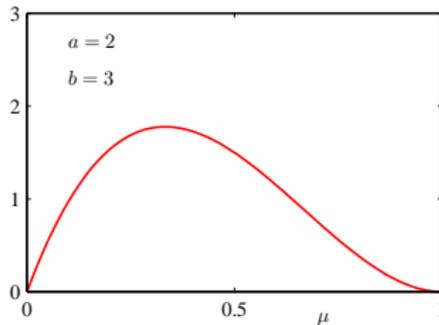
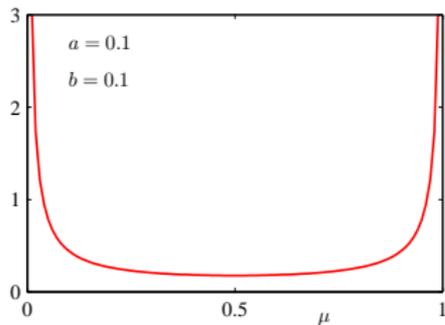
Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения

Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности



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

Продолжение примера

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения

Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности

- Применяя формулу Байеса, получаем

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

- 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 распределению на q приводит к характерной регуляризованной точечной оценке на вероятность выпадения орла

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

Примеры сопряженных распределений

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения
Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности

- 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

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

- 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

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения

Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности

Ликбез

Полный байесовский вывод

Пример использования априорных знаний

Сопряженные распределения

Иерархическая схема Байеса

Принцип наибольшей обоснованности

Обоснованность модели

Примеры использования

Выбор априорного распределения

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения

Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности

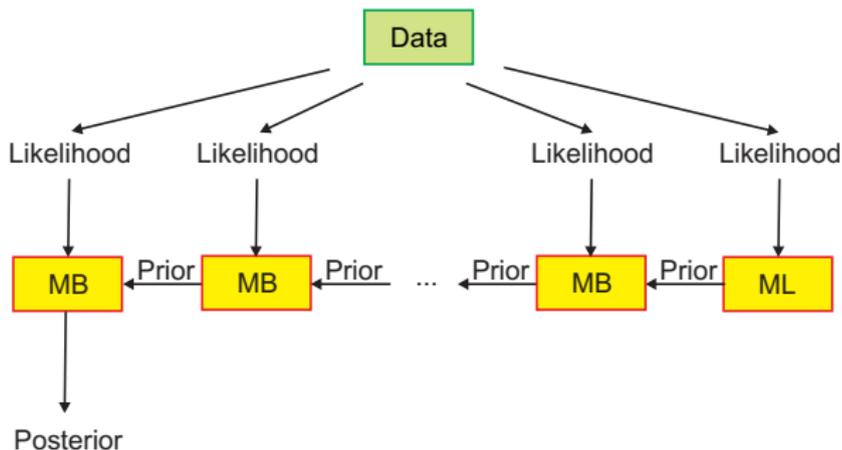
- Априорное распределение также может быть задано в параметрической форме $p(\theta) = p(\theta|\alpha)$
- Для того, чтобы применить формулу Байеса, необходимо сначала определить значение α
- Для оценки α можно вновь воспользоваться формулой Байеса, введя на α априорное распределение $p(\alpha)$. Тогда

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

- В качестве правдоподобия относительно α выступает т.н. обоснованность $p(\mathbf{x}|\alpha) = \int p(\mathbf{x}|\theta)p(\theta|\alpha)d\theta$, 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

Лекция 5.

Решение задачи
выбора модели
по Байесу.
Обоснованность
модели

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Пример
использования
априорных
знаний

Сопряженные
распределения
Иерархическая
схема Байеса

Принцип
наибольшей
обоснованности

План лекции

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

Обоснованность
модели

Примеры
использования

Ликбез

Полный байесовский вывод

Пример использования априорных знаний
Сопряженные распределения
Иерархическая схема Байеса

Принцип наибольшей обоснованности

Обоснованность модели

Примеры использования

Обоснованность модели

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

Обоснованность
модели

Примеры
использования

- 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
- Функция правдоподобия гиперпараметров называется обоснованностью (evidence) модели

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

- Гиперпараметры подбираются путем максимизации обоснованности

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

- Далее можно решать стандартную задачу на поиск максимума регуляризованного правдоподобия

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

Принцип наибольшей обоснованности с точки зрения байесовского подхода

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

Обоснованность
модели

Примеры
использования

- 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 $p(\alpha|\mathbf{x}) \propto p(\mathbf{x}|\alpha)$
- Вообще-то, это не совсем байесовский вывод...

ФОРМАЛЬНЫЙ ВЫВОД

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

$$\begin{aligned} p(x_{new}|\mathbf{x}) &= \int \int p(x|\theta, \alpha)p(\theta, \alpha|\mathbf{x})d\theta d\alpha = \left\{ p(\mathbf{x}|\theta, \alpha) = p(\mathbf{x}|\theta) \right\} = \\ &= \int \int p(x|\theta)p(\theta|\alpha, \mathbf{x})p(\alpha|\mathbf{x})d\theta d\alpha = \\ &= \left\{ p(\alpha|\mathbf{x}) \propto p(\mathbf{x}|\alpha) \right\} = \frac{1}{Z} \int \int p(x|\theta)p(\theta|\alpha, \mathbf{x})p(\mathbf{x}|\alpha)d\theta d\alpha = \\ &= \left\{ p(\mathbf{x}|\alpha) \approx \delta(\alpha - \alpha_{ME}) \right\} = \frac{1}{Z} \int p(x|\theta)p(\theta|\alpha_{ME}, \mathbf{x})d\theta = \\ &= \frac{1}{Zp(\mathbf{x}|\alpha_{ME})} \int p(x|\theta)p(\mathbf{x}|\theta, \alpha_{ME})p(\theta|\alpha_{ME})d\theta = \\ &= \left\{ p(\mathbf{x}|\theta, \alpha_{ME})p(\theta|\alpha_{ME}) \approx \delta(\theta - \theta_{MP}) \right\} \propto p(x|\theta_{MP}) \end{aligned}$$

Лекция 5.

Решение задачи
выбора модели
по Байесу.
Обоснованность
модели

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

Обоснованность
модели

Примеры
использования

Иллюстрация понятия обоснованности

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

Ветров,
Кропотов

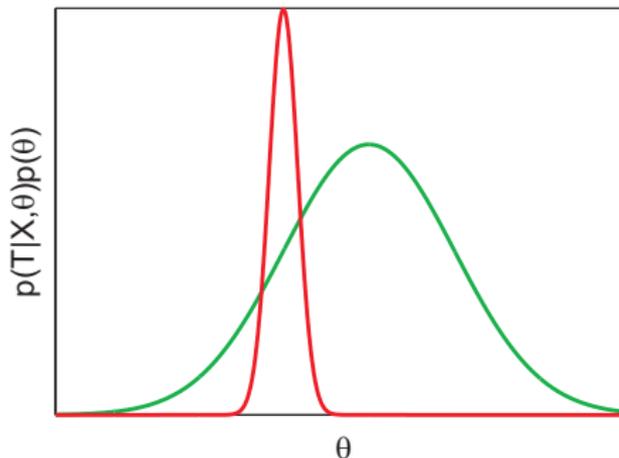
Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

Обоснованность
модели

Примеры
использования



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

План лекции

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

Обоснованность
модели

Примеры
использования

Ликбез

Полный байесовский вывод

Пример использования априорных знаний
Сопряженные распределения
Иерархическая схема Байеса

Принцип наибольшей обоснованности

Обоснованность модели
Примеры использования

Генератор случайных чисел I

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности
модели

Примеры
использования

- 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

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности
Обоснованность
модели

Примеры
использования

- Пусть вероятности выпадения чисел $1, 2, \dots, N$ равны, соответственно q_1, q_2, \dots, q_N
- Тогда правдоподобие выборки $\mathbf{x} = (x_1, \dots, x_n)$ равно

$$p(\mathbf{x}|\mathbf{q}) = \prod_{i=1}^n q_{x_i}$$

- Подставляя в формулу наши наблюдения получаем

$$p(x_1, x_2|\mathbf{q}) = q_8 q_6 \rightarrow \max_{\mathbf{q}}$$

- 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

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

Обоснованность
модели

Примеры
использования

- В отсутствие априорной информации о датчике случайных чисел, наиболее естественным является предположение о равномерности распределения вероятностей выпадения каждого числа $p(\mathbf{q}) = \text{Const}$
- Это частный случай распределения Дирихле

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

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

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

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

где $\alpha_6^1 = \alpha_8^1 = 2$, а все остальные $\alpha_i^1 = 1$

Байесовская оценка мат. ожидания

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

Обоснованность
модели

Примеры
использования

- Чтобы получить точечные оценки вероятностей q_1, \dots, 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

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

Обоснованность
модели

Примеры
использования

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

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

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

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

Ветров,
Кропотов

Ликбез

Полный
байесовский
вывод

Принцип
наибольшей
обоснованности

Обоснованность
модели

Примеры
использования

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»

