# Байесовский подход и графические модели

К. В. Воронцов, А.В. Зухба
vokov@forecsys.ru
a__l@mail.ru

сентябрь 2016

## Содержание

1. Два подхода к теории вероятностей
2. Формула Байеса и оценивание значений
3. Факторизация
4. Сопряженные распределения

# Sum- и Product- rule

**Sum-rule**

- Пусть $A_1, \dots, A_k$ взаимоисключающие события, одно из которых **всегда происходит**.

$$P(A_i \cup A_j) = P(A_i) + P(A_j) \qquad \sum_{i=1}^{k} P(A_i) = 1$$

- Очевидное следствие (формула полной вероятности): $\forall B$ верно $\sum_{i=1}^{k} P(A_i|B) = 1$, откуда

$$\sum_{i=1}^{k} \frac{P(B|A_i)P(A_i)}{P(B)} = 1 \qquad P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

**Product-rule**

- Правило произведения (product rule) гласит, что любую совместную плотность всегда можно разбить на множители

$$p(a, b) = p(a|b)p(b) \qquad P(A, B) = P(A|B)P(B)$$

- Можно показать (Jaynes, 1995), что Sum- и Product- rule являются единственными возможными операциями, позволяющими рассматривать вероятности как промежуточную ступень между истиной и ложью.

## Два подхода к теории вероятностей

**Частотный подход:** случайность есть *объективная неопределенность*.

- Величины четко делятся на случайные и детерминированные
- Теоретические результаты работают на больших выборках
- Точечные и интервальные оценки неизвестных параметров
- Основной метод - максимального правдоподобия

**Байесовский подход:** случайность есть *мера нашего незнания*.

- Все величины и параметры считаются случайными
- Методы работают даже при нулевом объеме выборки
- Оценки неизвестных параметров - апостериорные распределения
- Основным инструментом является формула Байеса

## Точечные оценки

- Математическое ожидание по апостериорному распределению. Весьма трудоемкая процедура

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

- Максимум апостериорной плотности.

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

- **Это регуляризация метода максимального правдоподобия!**

**Напоминание:** метод максимального правдоподобия

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

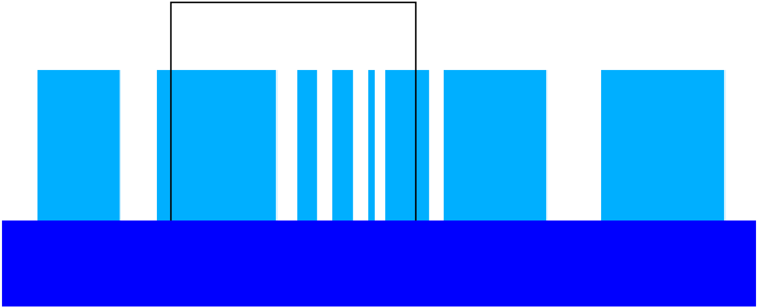
## Априорные и апостериорные суждения

- 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 $\theta$ посредством наблюдений некоторых ее косвенных характеристик $x|\theta$.

## Использование априорных знаний

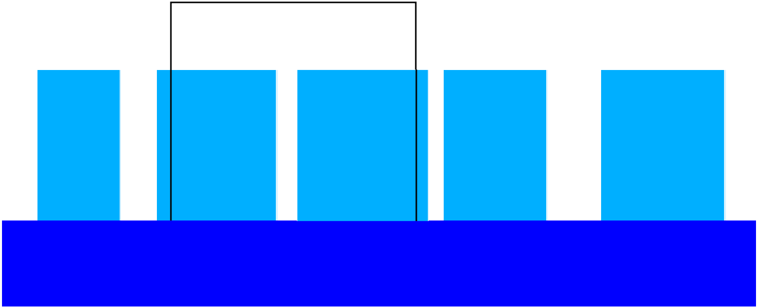
Сколько ящиков за препятствием?

## Использование априорных знаний



С точки зрения максимума правдоподобия, любое количество ящиков одинаково приемлемо

# Использование априорных знаний



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

## Формула Байеса

**Формула Байеса** устанавливает правила, по которым происходит преобразование знаний в процессе наблюдений:

Обозначим априорные знания о величине $\theta$ за $p(\theta)$.

Серия значений наблюдений: $\mathbf{x} = (x_1, \ldots, x_n)$.

Представления о значении $\theta$ меняются по 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 Баи?еса

## Пример: больной у врача

**Внешние симптомы:**
*Наблюдаемые переменные X*

1. Насморк
2. Температура
3. Головная боль
4. Количество эритроцитов

**Болезни:**
*Скрытые переменные Y*

1. Грипп
2. Аллергия
3. ...

**Генетические факторы:**
*Латентные переменные*

1. Предрасположенность к аллергиям
2. Предрасположенность к простудам

Ключевое наблюдение (из "житейской"логики):
P(грипп | насморк) » P(грипп | аллергия, насморк)

## Совместные вероятностные распределения:

Универсальный способ для дискретных переменных: таблица

| | 1 | 2 | 3 |
|---|---|---|---|
| 1 | $P(y_1 = 1, y_2 = 1)$ | $P(y_1 = 1, y_2 = 2)$ | $P(y_1 = 1, y_2 = 3)$ |
| 2 | $P(y_1 = 2, y_2 = 1)$ | $P(y_1 = 2, y_2 = 2)$ | $P(y_1 = 2, y_2 = 3)$ |
| 3 | $P(y_1 = 3, y_2 = 1)$ | $P(y_1 = 3, y_2 = 2)$ | $P(y_1 = 3, y_2 = 3)$ |
| 4 | $P(y_1 = 4, y_2 = 1)$ | $P(y_1 = 4, y_2 = 2)$ | $P(y_1 = 4, y_2 = 3)$ |
| 5 | $P(y_1 = 5, y_2 = 1)$ | $P(y_1 = 5, y_2 = 2)$ | $P(y_1 = 5, y_2 = 3)$ |

**Проблемы:**

- Экспоненциальный рост размера таблицы
- Экспоненциальный рост данных, необходимых для эмпирической оценки

**Вывод:** необходимы априорные предположения о структуре таблицы (т.е. вероятностного распределения)

# Факторизация

Ключевая идея: факторизация

$$P(y_1, y_2, y_3, y_4, y_5, y_6) = \frac{1}{Z}\Phi_1(y_1, y_2)\Phi_2(y_2, y_3)\Phi_3(y_3, y_4, y_5)\Phi_4(y_4, y_5, y_6)$$

Пусть $y_i \in \{1, 2, 3, 4\}$
Посчитаем экономию (количество параметров):
Было $4^6 - 1 = 4096 - 1$
Стало $16 + 16 + 64 + 64 - 4 = 160 - 4$ (оценка сверху)

**Общий вид:**

$$P(y) = \frac{1}{Z}\prod_{C_i}\Phi_i(y_{C_i})$$

## Аналитическое интегрирование

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

$$p(\theta|x) \sim A(\alpha_1)$$

- 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

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

- Сопряженным к распределению Бернулли является бета-распределение

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

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

# Пример

- Применяя формулу Байеса, получаем

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

- 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.

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

## 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

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

- 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.

## Структурное обучение

Реальное распределение данных $d(x, y)$ неизвестно. С точностью до параметра w задана модель предиктора

$$f_w(x) = \arg\max_y g(x, y, w)$$

**Дано:**

- однородная независимая выборка $\{(x^1, y^1), \dots, (x^n, y^n)\}$
- $\Delta : Y \times Y \to \mathbb{R}_+$ — функция потерь.

**Требуется найти:** $w^*$ т.ч. Байесовский риск насколько возможно мал.

$$\mathbb{E}_{(x,y)\sim d(x,y)} \Delta(y, f_{w^*}(x))$$

# Обычный SVM

Линейный классификатор:

$$a(x, w) = \mathrm{sign}\big(\langle w, x\rangle - w_0\big), \quad w, x \in \mathbb{R}^n, \; w_0 \in \mathbb{R}.$$

Пусть выборка $X^\ell = (x_i, y_i)_{i=1}^\ell$ линейно разделима:

$$\exists w, w_0 : \quad M_i(w, w_0) = y_i\big(\langle w, x_i\rangle - w_0\big) > 0, \quad i = 1, \dots, \ell$$

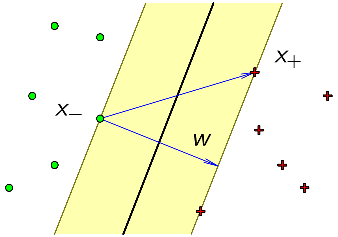
Нормировка: $\min\limits_{i=1,\dots,\ell} M_i(w, w_0) = 1.$

Разделяющая полоса:

$$\big\{x\colon \; -1 \leqslant \langle w, x\rangle - w_0 \leqslant 1\big\}.$$

Ширина полосы:

$$\frac{\langle x_+ - x_-, w\rangle}{\|w\|} = \frac{2}{\|w\|} \to \max.$$



## Обычный SVM

Линейно разделимая выборка

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min\limits_{w,w_0}; \\ M_i(w,w_0) \geqslant 1, \quad i = 1,\ldots,\ell. \end{cases}$$

Переход к линейно неразделимой выборке (эвристика)

$$\begin{cases} \dfrac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{\ell}\xi_i \to \min\limits_{w,w_0,\xi}; \\ M_i(w,w_0) \geqslant 1 - \xi_i, \quad i = 1,\ldots,\ell; \\ \xi_i \geqslant 0, \quad i = 1,\ldots,\ell. \end{cases}$$

Эквивалентная задача безусловной минимизации:

$$C\sum_{i=1}^{\ell}\big(1 - M_i(w,w_0)\big)_+ + \frac{1}{2}\|w\|^2 \to \min_{w,w_0}.$$

# Structured SVM

Обучение методом максимизации зазора:

$$\begin{cases} \gamma \to \max\limits_{w,\gamma}; \\ \|w\| \leqslant 1 \\ g(x^t, y^y, w) - g(x^t, y, w) \geqslant \gamma \\ \qquad t = 1, \dots n, \ y \in Y, \ y \neq y^t \end{cases}$$

$\gamma$ — величина зазора

$g(x^t, y^y, w) - g(x^t, y, w)$ — отступ

Физический смысл максимизации $\gamma$: помешать тому, чтобы значения функции совместности $g$ были одинаково большими для правильного и неправильного ответов.