# Две лекции про Байесовский подход

Винни-Пух

May 29, 2018

## Байесовский подход

В классическом методе максимального правдоподобия неизвестный параметр $\theta$ - это константа. А в байесовском подходе $\theta$ - это ненаблюдаемая случайная величина.

А наблюдения, $Y_1$, $Y_2$, ..., $Y_n$, и в классическом частотном подходе, и в байесовском считаются наблюдаемыми случайными величинами.

В байесовском подходе можно выделить два предположения:

1. Изначальное мнение о неизвестном параметре $\theta$, сформулированное в виде закона распределения. Этот закон называется априорным законом распределения.
2. Функция правпдоподобия - представление о законе распределения наблюдений $Y_1$, $Y_2$, ..., $Y_n$.

Если к этим двум предположениям добавить сами наблюдения $Y_1$, $Y_2$, ..., $Y_n$, то по формуле условной вероятности можно получить вывод:

1. Мнение о неизвестном параметре $\theta$ с учётом полученных наблюдений. Это апостериорный закон распределения.

### Первая задача

Пример. Подходит ко мне в поезде подозрительный тип и говорит, “а давай сыграем в орлянку моей монеткой”. Здесь явно возникает неизвестный параметр $p$ - вероятность выпадения орла.

Предположим, что:

1. Изначально я ничего не знаю про этого типа, может он любит делать монетки, чаще выпадающие орлом, а может и нет :) Поэтому буду считать, что изначально вероятность выпадения орла равномерна на отрезке $[0;1]$, $p \sim U[0;1]$.
2. Если бы $p$ было известно и фиксировано, то результаты подбрасывания монетки, $Y_i$, были независимы, и $P(Y_i = \text{Орёл}) = p$.

Пока подозрительный тип отвернулся, я успел три раза подбросить монетку и получил выборку $Y_1 = \text{Орёл}$, $Y_2 = \text{Решка}$, $Y_3 = \text{Орёл}$.

Каким должно быть моё мнение о монетке с учётом полученной информации?

Чему, например, равна вероятность $P(p > 0.5 \mid \text{Данные})$?

Чему равно $E(p \mid \text{Данные})$?

Перейдём к решению!

Маленькое обозначение :)

Для удобства мы введём значок “пропорционально” $\propto$.

Например, формула

$$f(x) \propto x^2$$

означает, что функция $f(x)$ может равняться $5x^2$ или $19x^2$, но никак не $\cos(x) + 4$.

....

В качестве основной формулы получаем:

$$f(\theta \mid \text{Данные}) \propto f(\theta) \cdot f(\text{Данные} \mid \theta)$$

Словами:

$$\text{Апостериорная плотность} \propto \text{Априорная плотность} \cdot \text{Функция правдоподобия}$$

...

Получаем апостериорную функцию плотности

$$f(p \mid \text{Данные }) = \begin{cases} 12p^2(1-p), \text{ если } p \in [0;1] \\ 0, \text{ иначе} \end{cases}$$

С помощью неё легко найти $P(p > 0.5 \mid \text{Данные})$ и $E(p \mid \text{Данные})$.

## Соответствия

Некоторые вопросы, на которые байесовский подход даёт ответы, например, чему равна вероятность $P(p > 0.5 \mid \text{Данные})$ бессмысленны в классическом подходе. Ведь в классическом подходе $p$ - неизвестная константа и потому любая вероятность связанная с истинным параметром либо равна 0, либо 1, а чему конкретно - узнать невозможно.

В классическом подходе находят точечные оценки и интервальные оценки неизвестных параметров. Что заменяет их в байесовском подходе?

1. Точечную оценку $\hat{\theta}_{ML}$ обычно заменяют на:

- апостериорное среднее $\hat{\theta}_{BA} = E(\theta \mid \text{Данные})$;
- апостериорную моду $\hat{\theta}_{MAP} = Mode(\theta \mid \text{Данные})$, MAP расшифровывается как maximum aposteriori estimator.

1. Интервальные оценки в байесовском подходе строят естественным образом: находят такие точки $a$ и $b$, что $P(\theta \in [a;b] \mid \text{Данные}) = 0.95$. При этом есть два популярных варианта:

- симметрично по вероятности интервалы: слева от $a$ оказывается такая же вероятность, как справа от $b$;
- HPD-интервал (highest probability density): интервал строится так, чтобы плотность на отрезке $[a;b]$ была выше, чем за отрезком $[a;b]$. Если апостериорное распределение не симметричное, то HPD интервал может оказаться существенно короче.

## Всё не так просто!

В реальности всё гораздо сложнее и посчитать апостериорную плотность явно практически никогда не получается. Вместо явного вывода громоздкой многомерной плотности мы будем хранить в памяти компьютера большую выборку из апостериорного распределения, $\theta^{(1)}$, $\theta^{(2)}$, ....

И, например, если мы захотим посчитать вероятность $P(p > 0.5 \mid \text{Данные})$, то мы просто посмотрим, какой процент $p^{(i)}$ из нашей апостериорной выборки оказался больше 0.5.

Отметим, что размер выборки из апостериорного распределения никак не связан с количеством наблюдений. У нас может быть всего три наблюдения и выборка размера 100 тысяч из апостериорного распределения. Чем больше будет размер выборки из апостериорного распределения, тем точнее мы опишем апостериорный закон распределения.

Чтобы построить выборку из апостериорного закона распределения используют разные алгоритмы. Например, алгоритм Гиббса, алгоритм Метрополиса-Гастингса или Гамильтоновское Монте-Карло.

## Алгоритм Гиббса

Допустим у нас три параметра, которые мы хотим оценить, $\theta = (\theta_1, \theta_2, \theta_3)$.

На выходе из алгоритма Гиббса мы хотим получить выборку из апостериорного закона распределения.

Шаг 1. Выбираем произвольные допустимые стартовые значения $\theta_1^{(1)}$, $\theta_1^{(2)}$, $\theta_1^{(3)}$. Пусть будет $\theta_1^{(1)} = 0$, $\theta_2^{(1)} = 0$, $\theta_3^{(1)} = 0$.

Шаг 2. По очереди обновляем каждый из параметров, используя самые свежии версии остальных параметров.

Шаг 2.1. Случайно генерируем $\theta_1^{(i+1)}$ из условного распределения $f(\theta_1|\theta_2 = \theta_2^{(i)}, \theta_3 = \theta_3^{(i)}, \text{Данные})$.

Шаг 2.2. Случайно генерируем $\theta_2^{(i+1)}$ из условного распределения $f(\theta_2|\theta_1 = \theta_1^{(i+1)}, \theta_3 = \theta_3^{(i)}, \text{Данные})$.

Шаг 2.3. Случайно генерируем $\theta_3^{(i+1)}$ из условного распределения $f(\theta_3|\theta_1 = \theta_1^{(i+1)}, \theta_2 = \theta_2^{(i+1)}, \text{Данные})$.

Шаг 3. Пока не получилась достаточно большая выборка переходим к шагу 2.

## Алгоритм Метрополиса-Гастингса