

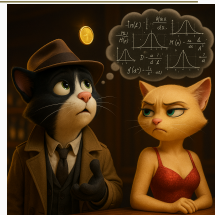
БАЙЕСОВСКИЙ ВЫВОД ДЛЯ МОНЕТКИ

Сергей Николенко

СПбГУ — Санкт-Петербург

11 сентября 2025 г.

Random facts:



- 11 сентября — не самая приятная дата: в 1847 г. 300 человек погибли в результате урагана на Ньюфаундленде, в 1927 г. произошло землетрясение в Крыму с большими разрушениями от Севастополя до Феодосии, в 2012 г. в пожарах на двух ткацких фабриках в Пакистане погибли 315 сотрудников, в 2015 г. башенный кран упал на мечеть Аль-Харам в Мекке и задавил насмерть 111 человек, а в 2001 году — сами понимаете...
- 11 сентября — Национальный день Каталонии, учреждённый в память о взятии Барселоны во время войны за испанское наследство 11 сентября 1713 года
- 11 сентября 1943 г. на стекольном заводе в Гусь-Хрустальном был выпущен первый советский гранёный стакан; Вера Мухина здесь была ни при чём
- 11 сентября 1962 г. The Beatles завершили запись своего первого сингла P.S. I Love You; на барабанах был Энди Уайт, сессионный ударник, которого пригласила студия, потому что Ринго очень волновался и никак не мог выдать нужный результат
- 11 сентября 1973 г. в результате профинансированного ЦРУ переворота в Чили президент Альенде совершил самоубийство, а к власти пришел Аугусто Пиночет; так прекратился проект Киберсин о централизованном компьютерном управлении плановой экономикой под руководством Стаффорда Бира

БАЙЕСОВСКИЙ ВЫВОД ДЛЯ МОНЕТКИ

- Мы остановились на том, что в статистике обычно ищут *гипотезу максимального правдоподобия* (maximum likelihood):

$$\theta_{ML} = \arg \max_{\theta} p(D | \theta).$$

- В байесовском подходе ищут *апостериорное распределение* (posterior)

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

и, возможно, *максимальную апостериорную гипотезу* (maximum a posteriori):

$$\theta_{MAP} = \arg \max_{\theta} p(\theta | D) = \arg \max_{\theta} p(D | \theta)p(\theta).$$

- Простая задача вывода: дана нечестная монетка, она подброшена N раз, имеется последовательность результатов падения монетки. Надо определить её «нечестность» и предсказать, чем она выпадет в следующий раз.

- Если у нас есть вероятность $p_h = \theta$ того, что монетка выпадет орлом (вероятность решки $p_t = 1 - \theta$), то вероятность того, что выпадет последовательность s , которая содержит n орлов и m решек, равна

$$p(s|p_h = \theta) = \theta^n (1 - \theta)^m.$$

- Сделаем предположение: будем считать, что вероятность монетки до экспериментов для нас распределена равномерно, т.е. у нас нет априорных предпочтений насчёт θ .
- Теперь нужно использовать теорему Байеса и вычислить скрытые параметры.

- Правдоподобие: $p(\theta|s) = \frac{p(s|\theta)p(\theta)}{p(s)}$.
- Здесь $p(\theta)$ следует понимать как непрерывную случайную величину, сосредоточенную на интервале $[0, 1]$, коей она и является. Наше предположение о равномерном распределении в данном случае значит, что априорная вероятность $p(\theta) = 1, \theta \in [0, 1]$ (т.е. априори мы не знаем, насколько нечестна монетка, и предполагаем это равновероятным). А $p(s|\theta)$ мы уже знаем.
- Итого получается:

$$p(\theta|s) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

- Итого получается:

$$p(\theta|s) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

- $p(s)$ можно подсчитать как

$$\begin{aligned} p(s) &= \int_0^1 \theta^n(1-\theta)^m d\theta = \\ &= \frac{\Gamma(n+1)\Gamma(m+1)}{\Gamma(n+m+2)} = \frac{n!m!}{(n+m+1)!}, \end{aligned}$$

но найти $\arg \max_{\theta} p(\theta | s) = \frac{n}{n+m}$ можно и без этого.

- Итого получается:

$$p(\theta|s) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

- Но это ещё не всё. Чтобы предсказать следующий исход, надо найти $p(\text{heads}|s)$:

$$\begin{aligned} p(\text{heads}|s) &= \int_0^1 p(\text{heads}|\theta)p(\theta|s)d\theta = \\ &= \int_0^1 \frac{\theta^{n+1}(1-\theta)^m}{p(s)} dp_h = \\ &= \frac{(n+1)!m!}{(n+m+2)!} \cdot \frac{(n+m+1)!}{n!m!} = \frac{n+1}{n+m+2}. \end{aligned}$$

- Получили правило Лапласа.

- Итого получается:

$$p(\theta|s) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

- Это была иллюстрация двух основных задач байесовского вывода:
 - найти апостериорное распределение (posterior distribution) на гипотезах/параметрах:

$$p(\theta | D) \propto p(D|\theta)p(\theta)$$

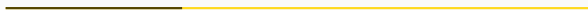
и/или найти максимальную апостериорную гипотезу (maximum a posteriori) $\theta_{\text{MAP}} = \arg \max_{\theta} p(\theta | D)$;

- найти предсказательное распределение (predictive distribution) исходов дальнейших экспериментов:

$$p(x | D) \propto \int_{\theta \in \Theta} p(x | \theta)p(D|\theta)p(\theta)d\theta.$$

СОПРЯЖЁННЫЕ
РАСПРЕДЕЛЕНИЯ

АПРИОРНЫЕ



- Напоминаю, что основная наша задача – как обучить параметры распределения и/или предсказать следующие его точки по имеющимся данным.
- В байесовском выводе участвуют:
 - $p(x | \theta)$ – правдоподобие данных;
 - $p(\theta)$ – априорное распределение;
 - $p(x) = \int_{\Theta} p(x | \theta)p(\theta)d\theta$ – маргинальное правдоподобие;
 - $p(\theta | x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ – апостериорное распределение;
 - $p(x' | x) = \int_{\Theta} p(x' | \theta)p(\theta | x)d\theta$ – предсказание нового x' .
- Задача обычно в том, чтобы найти $p(\theta | x)$ и/или $p(x' | x)$.

- Когда мы проводим байесовский вывод, у нас, кроме правдоподобия, должно быть ещё *априорное распределение* (prior distribution) по всем возможным значениям параметров.
- Мы раньше к ним специально не присматривались, но они очень важны.
- Задача байесовского вывода – как подсчитать $p(\theta | x)$ и/или $p(x' | x)$.
- Но чтобы это сделать, сначала надо выбрать $p(\theta)$. Как выбирать априорные распределения?

- Разумная цель: давайте будем выбирать распределения так, чтобы они оставались такими же и *a posteriori*.
- До начала вывода есть априорное распределение $p(\theta)$.
- После него есть какое-то новое апостериорное распределение $p(\theta | x)$.
- Я хочу, чтобы $p(\theta | x)$ тоже имело тот же вид, что и $p(\theta)$, просто с другими параметрами.

- Не слишком формальное определение: семейство распределений $p(\theta \mid \alpha)$ называется семейством *сопряжённых априорных распределений* для семейства правдоподобий $p(x \mid \theta)$, если после умножения на правдоподобие апостериорное распределение $p(\theta \mid x, \alpha)$ остаётся в том же семействе: $p(\theta \mid x, \alpha) = p(\theta \mid \alpha')$.
- α называются *гиперпараметрами* (hyperparameters), это «параметры распределения параметров».
- Тривиальный пример: семейство всех распределений будет сопряжённым чему угодно, но это не очень интересно.

- Разумеется, вид хорошего априорного распределения зависит от вида распределения собственно данных, $p(x \mid \theta)$.
- Сопряжённые априорные распределения подсчитаны для многих распределений, мы приведём несколько примеров.

- Каким будет сопряжённое априорное распределение для бросания нечестной монетки (испытаний Бернулли)?
- Ответ: это будет бета-распределение; плотность распределения нечестности монетки θ

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

- Плотность распределения нечестности монетки θ

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

- Тогда, если мы посэмплируем монетку, получив s орлов и f решек, получится

$$p(s, f \mid \theta) = \binom{s+f}{s} \theta^s (1-\theta)^f, \text{ и}$$

$$\begin{aligned} p(\theta \mid s, f) &= \frac{\binom{s+f}{s} \theta^{s+\alpha-1} (1-\theta)^{f+\beta-1} / B(\alpha, \beta)}{\int_0^1 \binom{s+f}{s} x^{s+\alpha-1} (1-x)^{f+\beta-1} / B(\alpha, \beta) dx} = \\ &= \frac{\theta^{s+\alpha-1} (1-\theta)^{f+\beta-1}}{B(s+\alpha, f+\beta)}. \end{aligned}$$

- Итого получается, что сопряжённое априорное распределение для параметра нечестной монетки θ – это

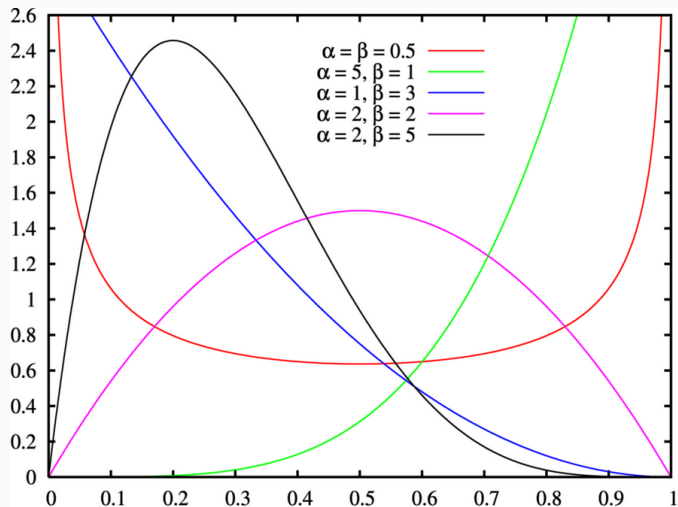
$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- После получения новых данных с s орлами и f решками гиперпараметры меняются на

$$p(\theta \mid s + \alpha, f + \beta) \propto \theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}.$$

- На этом этапе можно забыть про сложные формулы и выводы, получилось очень простое правило обучения (под обучением теперь понимается изменение гиперпараметров).

БЕТА--РАСПРЕДЕЛЕНИЕ



- Простое обобщение: рассмотрим мультиномиальное распределение с n испытаниями, k категориями и по x_i экспериментов дали категорию i .
- Параметры θ_i показывают вероятность попасть в категорию i :

$$p(x \mid \theta) = \binom{n}{x_1, \dots, x_n} \theta_1^{x_1} \theta_2^{x_2} \dots \theta_k^{x_k}.$$

- Сопряжённым априорным распределением будет распределение Дирихле:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \dots \theta_k^{\alpha_k-1}.$$

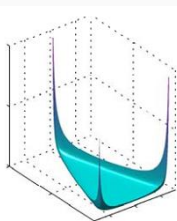
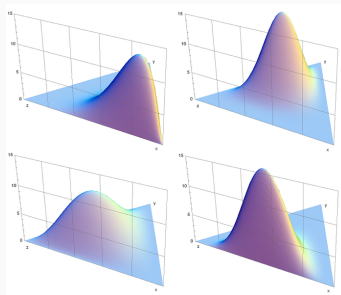
- Сопряжённым априорным распределением будет распределение Дирихле:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \dots \theta_k^{\alpha_k-1}.$$

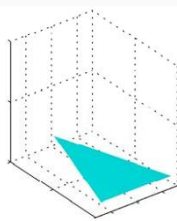
Упражнение. Докажите, что при получении данных $x_1, \dots, x_k$ гиперпараметры изменятся на

$$p(\theta \mid x, \alpha) = p(\theta \mid x + \alpha) \propto \theta_1^{x_1+\alpha_1-1} \theta_2^{x_2+\alpha_2-1} \dots \theta_k^{x_k+\alpha_k-1}.$$

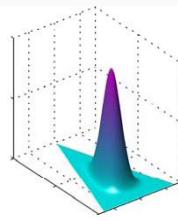
РАСПРЕДЕЛЕНИЕ ДИРИХЛЕ



$$\{\alpha_k\} = 0.1$$



$$\{\alpha_k\} = 1$$



$$\{\alpha_k\} = 10$$

ЕЩЁ НЕМНОГО О ВЕРОЯТНОСТЯХ

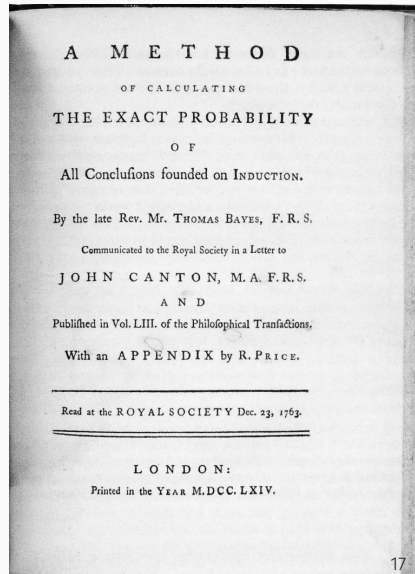
- Приведём классический пример из классической области применения статистики — медицины.
- Пусть некий тест на какую-нибудь болезнь имеет вероятность успеха 95% (т.е. 5% — вероятность как позитивной, так и негативной ошибки).
- Всего болезнь имеется у 1% респондентов (отложим на время то, что они разного возраста и профессий).
- Пусть некий человек получил позитивный результат теста (тест говорит, что он болен). С какой вероятностью он действительно болен?

- Приведём классический пример из классической области применения статистики — медицины.
- Пусть некий тест на какую-нибудь болезнь имеет вероятность успеха 95% (т.е. 5% — вероятность как позитивной, так и негативной ошибки).
- Всего болезнь имеется у 1% респондентов (отложим на время то, что они разного возраста и профессий).
- Пусть некий человек получил позитивный результат теста (тест говорит, что он болен). С какой вероятностью он действительно болен?
- Ответ: 16%.

- Обозначим через t результат теста, через d — наличие болезни.
- $p(t = 1) = p(t = 1|d = 1)p(d = 1) + p(t = 1|d = 0)p(d = 0)$.
- Используем теорему Байеса:

$$\begin{aligned} p(d = 1|t = 1) &= \\ &= \frac{p(t = 1|d = 1)p(d = 1)}{p(t = 1|d = 1)p(d = 1) + p(t = 1|d = 0)p(d = 0)} = \\ &= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} = 0.16. \end{aligned}$$

- Вот такие задачи составляют суть вероятностного вывода (probabilistic inference).
- Поскольку они обычно основаны на теореме Байеса, вывод часто называют байесовским (Bayesian inference).
- Но не только поэтому.

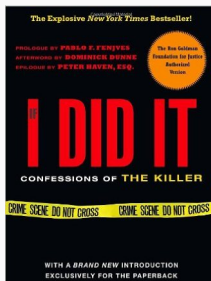


1. У моего знакомого два ребёнка. Будем предполагать, что пол ребёнка выбирается независимо и равновероятно, с вероятностью $\frac{1}{2}$. Две постановки вопроса:
 - (1) я спросил, есть ли у него мальчики, и он ответил «да»; какова вероятность того, что один из детей – девочка?
 - (2) я встретил одного из его детей, и это мальчик; какова вероятность того, что второй ребёнок – девочка?

2. Произошло убийство. На месте убийства найдена кровь, которая явно принадлежит убийце. Кровь принадлежит редкой группе, которая присутствует у 1% населения, в том числе у подсудимого.
- (1) Прокурор говорит: «Шанс, что у подсудимого была бы именно такая группа крови, если бы он был невиновен – всего 1%; значит, с вероятностью 99% он виновен». В чём не прав прокурор?
- (2) Адвокат говорит: «В городе живёт миллион человек, то есть у 10000 из них такая группа крови. Значит, всё, что говорит нам эта кровь – это что подсудимый совершил убийство с вероятностью 0.01%; никакое это не доказательство». В чём не прав адвокат?

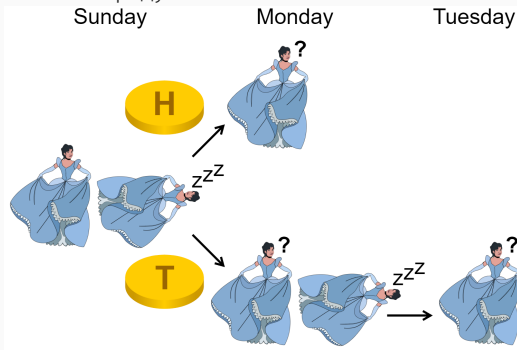
3. Реальные случаи.

- (1) Прокурор указал, что O.J. Simpson уже бил жену в прошлом. Адвокат ответил: «Убивают только одну из 2500 женщин, подвергавшихся семейному насилию, так что это вообще нерелевантно». Суд согласился с адвокатом; верно ли это рассуждение?
- (2) У Sally Clark погибли два младенца; прокурор указал, что вероятность двух случаев SIDS в одной семье, которую он получил из статистики одиночных случаев, — около 1 из 73 миллионов; в чём он не прав?



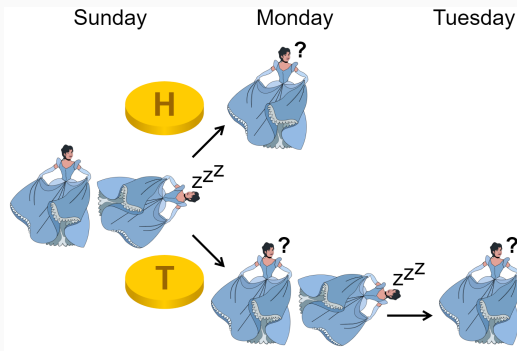
ЗАДАЧА О СПЯЩЕЙ КРАСАВИЦЕ

- А вот парадокс, который до сих пор вызывает споры (Zuboff, 1990; Elga, 2000). В воскресенье Спящую красавицу усыпляют и бросают монетку:
 - если выпадает орёл, её будят в понедельник и опять усыпляют до среды
 - если решка, то будят в понедельник, делают укол, от которого она забывает о пробуждении, и будят ещё раз во вторник, а потом так же в среду



ЗАДАЧА О СПЯЩЕЙ КРАСАВИЦЕ

- Спящая красавица знает всю процедуру эксперимента, но не может различить разные пробуждения и отличить понедельник от вторника, когда просыпается
- Представьте себя на месте Спящей красавицы; вас разбудили; какова вероятность того, что монета упала орлом?



СПАСИБО!

Спасибо за внимание!

