

Прикладная статистика в машинном обучении

Лекция 8

Байесовские методы

И. К. Козлов
(Мехмат МГУ)

2022

Теорема Байеса

1 Теорема Байеса

2 Правила суммы и произведения

3 Байесовская вероятность

- Субъективная неопределённость
- Байесовский вывод
- Выбор априорного распределения

4 Сопряжённые распределения

5 Трамвай

6 Применимость байесовского подхода

Be afraid, be very afraid

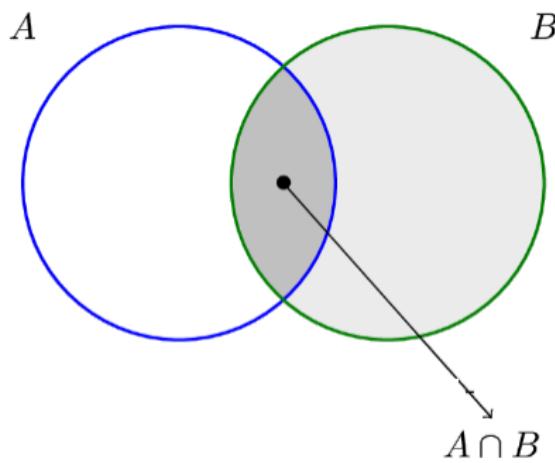
Мы закончили с частотным подходом к статистике.

Перейдём к [байесовским методам и генеративным моделям](#).



Условная вероятность

Условная вероятность:



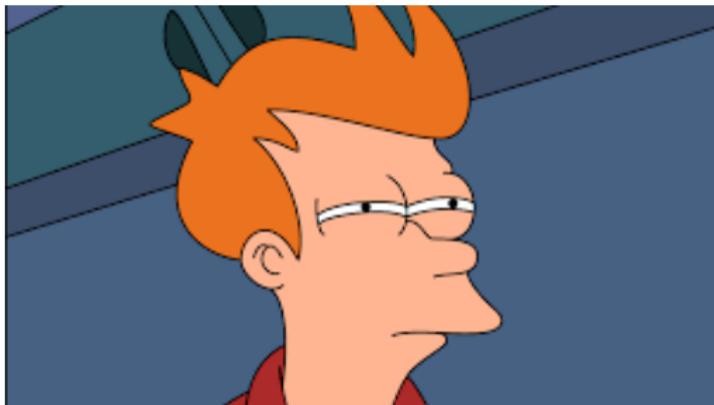
$$P(A|B) = \frac{\text{Probability of } A \text{ and } B}{\text{Probability of } B}$$

Something wicked this way comes

Выразим совместную вероятность событий двумя способами:

$$P(A \cap B) = P(A | B)P(B) = P(B | A)P(A)$$

Вглядитесь — сейчас мы узрим одну из важнейших формул курса!



Теорема Байеса

Теорема Байеса:

$$P(A|B) = P(A) \times \frac{P(B|A)}{P(B)}$$

posterior prior likelihood marginal



Формула полной вероятности

Чтобы применить теорему Баайеса, вспомним важный факт.

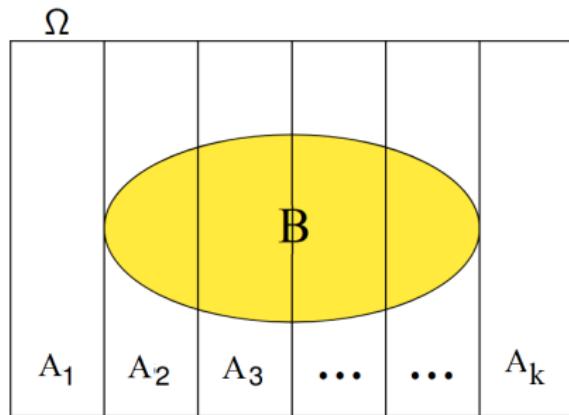
Формула полной вероятности. Пусть A_1, \dots, A_k — разбиение пространства элементарных событий Ω , т.е. A_i измеримы и

$$\Omega = \bigcup_{i=1}^k A_i, \quad A_i \cap A_j = \emptyset, \quad i \neq j.$$

Тогда для любого события B выполнено

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

Формула полной вероятности



Доказательство. События $B \cap A_i$ попарно не пересекаются и вместе образуют B , поэтому

$$\mathbb{P}(B) = \sum_{i=1}^k \mathbb{P}(B \cap A_i) = \sum_{i=1}^k \mathbb{P}(B | A_i) \mathbb{P}(A_i).$$

Формула Байеса

Применим формулу полной вероятности к формуле Байеса:

$$\mathbb{P}(A_i | B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B | A_i)}{\sum_{j=1}^k \mathbb{P}(A_j)\mathbb{P}(B | A_j)}$$

ML пример

Пример. Применение теоремы Байеса в ML.

У нас есть данные Data и мы хотим отличить котика от пёсика.

$$f(\text{dog}) \rightarrow \text{dog}$$

$$f(\text{cat}) \rightarrow \text{cat}$$

Q: Какую вероятность мы хотим посчитать?

ML пример

Пример. Применение теоремы Байеса в ML.

У нас есть данные Data и мы хотим отличить котика от пёсика.

$$f(\text{dog}) \rightarrow \text{dog}$$

$$f(\text{cat}) \rightarrow \text{cat}$$

Q: Какую вероятность мы хотим посчитать?

A: Вероятность класса при наблюдаемых данных:

$$\mathbb{P}(\text{Cat} | \text{Data}).$$

ML пример

- В жизни проверить — подходит ли описание котику

$$\mathbb{P}(\text{Data} \mid \text{Cat})$$

гораздо проще, чем угадать — какое животное описано

$$\mathbb{P}(\text{Cat} \mid \text{Data}).$$

- Также несложно оценить доли котиков и пёсиков:

$$\mathbb{P}(\text{Cat}), \quad \mathbb{P}(\text{Dog})$$

ML пример

Остаётся применить **теорему Байеса**,

$$\mathbb{P}(\text{Cat} | \text{Data}) = \frac{\mathbb{P}(\text{Data} | \text{Cat}) \mathbb{P}(\text{Cat})}{\mathbb{P}(\text{Data} | \text{Cat}) \mathbb{P}(\text{Cat}) + \mathbb{P}(\text{Data} | \text{Dog}) \mathbb{P}(\text{Dog})}$$

Voilà, мы нашли вероятность котика ☺.

ML пример

Остаётся применить **теорему Байеса**,

$$\mathbb{P}(\text{Cat} | \text{Data}) = \frac{\mathbb{P}(\text{Data} | \text{Cat}) \mathbb{P}(\text{Cat})}{\mathbb{P}(\text{Data} | \text{Cat}) \mathbb{P}(\text{Cat}) + \mathbb{P}(\text{Data} | \text{Dog}) \mathbb{P}(\text{Dog})}$$

Voilà, мы нашли вероятность котика ☺.

Замечание. Первое слагаемое в знаменателе — всегда числитель.

Правила суммы и произведения

1 Теорема Байеса

2 Правила суммы и произведения

3 Байесовская вероятность

- Субъективная неопределённость
- Байесовский вывод
- Выбор априорного распределения

4 Сопряжённые распределения

5 Трамвай

6 Применимость байесовского подхода

Учимся складывать и умножать

Вспомним 2 важных факта — как складывать и умножать в Теории Вероятностей.

Замечание. В лекциях по Байесовским методам будем обозначать плотность вероятности через $p(x)$, а не $f(x)$.



Правило суммы

Правило суммы. Если A_1, \dots, A_k — разбиение Ω , то

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

В интегральном виде это записывается в виде

$$p(b) = \int p(b, a) da = \int p(b | a) p(a) da.$$

Правило произведения

Правило произведения. Переписываем формулу условной вероятности:

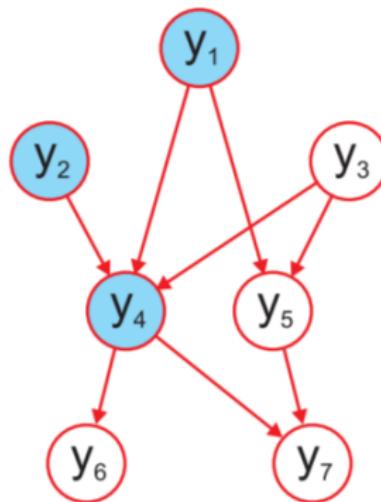
$$p(a, b) = p(a | b)p(b).$$

В общем случае

$$p(x_1, \dots, x_n) = p(x_1 | x_2, \dots, x_n)p(x_2 | x_3, \dots, x_n) \dots p(x_{n-1} | x_n)p(x_n).$$

Условные вероятности

Зависимости между переменными удобно выражать диаграммой следующего вида:



Все условные вероятности вычисляются при помощи **правил суммы и произведения**.

На семинаре вычислим $p(y_5, y_7)$ при известных y_1, y_2, y_4 и неизвестных y_3, y_6 .

Маргинализация и обуславливание

Общее правило. Пусть мы хотим вычислить $p(x)$ по $p(x, y)$.

- Если y **неизвестно**, то мы **маргинализуем** по ним:

$$p(x) = \int p(x, y) dy$$

- Если y **известно**, то мы **обуславливаем** плотность по нему:

$$p(x | y) = \frac{p(x, y)}{p(y)}.$$

Байесовская вероятность

1 Теорема Байеса

2 Правила суммы и произведения

3 Байесовская вероятность

- Субъективная неопределённость
- Байесовский вывод
- Выбор априорного распределения

4 Сопряжённые распределения

5 Трамвай

6 Применимость байесовского подхода

Субъективная неопределённость

1 Теорема Байеса

2 Правила суммы и произведения

3 Байесовская вероятность

- Субъективная неопределённость
- Байесовский вывод
- Выбор априорного распределения

4 Сопряжённые распределения

5 Трамвай

6 Применимость байесовского подхода

Опишем Байесовский подход к Статистике.
Объясним — чем он отличается от Частотного подхода.

%

Probability of the
events observed
given a theory

FREQUENTIST
STATISTICS



%

Probability of the
multiple theories
given the observed events

BAYESIAN
STATISTICS



2 подхода к Статистике:

Частотный

θ - постоянные, неизвестные параметры

Вероятности - частоты в экспериментах

Строим точечные оценки или
доверительные интервалы для θ

ОБЪЕКТИВНАЯ РЕАЛЬНОСТЬ

Байесовский

$p(\theta)$ - распределение на параметры

Вероятности - степени уверенности

Применяем теорему Байеса, находим
апостериорное распределение $p(\theta | x)$

СУБЪЕКТИВНАЯ НЕОПРЕДЕЛЁННОСТЬ

Частотная и байесовская статистики

На самом деле люди рассуждают скорее “по-байесовски”.

Мы не можем проводить бесконечную серию экспериментов.

Пример. Подбросили монетку. **Q:** С какой вероятностью выпадет орёл?



Частотная и байесовская статистики

- **Частотный ответ:** Если подбрасывать бесконечное число раз, то в среднем будет 50% орлов.
- **Байесовский ответ:** Чем больше данных (направление и время полёта, скорость и т.д.) — тем яснее, что выпадет.



Байесовский вывод

- 1 Теорема Байеса
- 2 Правила суммы и произведения
- 3 Байесовская вероятность
 - Субъективная неопределённость
 - Байесовский вывод
 - Выбор априорного распределения
- 4 Сопряжённые распределения
- 5 Трамвай
- 6 Применимость байесовского подхода

Байесовский вывод

Байесовский вывод

- ① Фиксируем **априорное распределение** на параметры $p(\theta)$.

Закладываем в него наши знания и представления о данных.

Байесовский вывод

Байесовский вывод

- ① Фиксируем **априорное распределение** на параметры $p(\theta)$.

Закладываем в него наши знания и представления о данных.

- ② Выбираем статистическую модель $p(x | \theta)$ распределения данных.

Байесовский вывод

Байесовский вывод

- ① Фиксируем **априорное распределение** на параметры $p(\theta)$.

Закладываем в него наши знания и представления о данных.

- ② Выбираем статистическую модель $p(x | \theta)$ распределения данных.
- ③ Наблюдаем данные $x^n = (x_1, \dots, x_n)$ и вычисляем **апостериорное распределение по формуле Байеса**

$$p(\theta | x^n) = \frac{p(x^n | \theta) p(\theta)}{\int p(x^n | \theta) p(\theta) d\theta}.$$

Байесовский вывод

Для одного наблюдения формула Байеса имеет вид

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

Байесовский вывод

Для одного наблюдения формула Байеса имеет вид

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

Для независимых наблюдений **функция правдоподобия** распадётся в произведение:

$$p(x^n | \theta) = \prod_{i=1}^n p(x_i | \theta).$$

Итоговая формула:

$$p(\theta | x^n) = \frac{\prod_{i=1}^n p(x_i | \theta) p(\theta)}{\int \prod_{i=1}^n p(x_i | \theta) p(\theta) d\theta}.$$

Нормировочная константа

Посмотрим на формулу Байеса:

$$p(\theta | x^n) = \frac{p(x^n | \theta) p(\theta)}{\int p(x^n | \theta) p(\theta) d\theta}.$$

Q: Зависит ли знаменатель от θ ?

Нормировочная константа

Посмотрим на формулу Байеса:

$$p(\theta | x^n) = \frac{p(x^n | \theta) p(\theta)}{\int p(x^n | \theta) p(\theta) d\theta}.$$

Q: Зависит ли знаменатель от θ ?

A: Нет, это нормировочная константа, чтобы получалась вероятность:

$$p(\theta | x^n) = \frac{1}{Z} p(x^n | \theta) p(\theta).$$

Пропорциональность

Равенство с точностью до умножения на константу обозначают знаком “ \propto ”.

$$\text{Апостериорное} \propto \text{Правдоподобие} \cdot \text{Априорное}$$

Перерыв

Перерыв

Выбор априорного распределения

1 Теорема Байеса

2 Правила суммы и произведения

3 Байесовская вероятность

- Субъективная неопределённость
- Байесовский вывод
- Выбор априорного распределения

4 Сопряжённые распределения

5 Трамвай

6 Применимость байесовского подхода

Оставь надежду всяк сюда входящий

Q: Как выбрать априорное распределение $p(\theta)$?

A: В него нужно “вложить наши *субъективные представления и знания*” о данных.

Многих учёных ожидаемо “триггерит” от подобного.



Рис.: Добро пожаловать в ад Субъективизма

Несобственный прайор

Один из “универсальных” выходов (не лучший).

Взять **несобственный прайор** — пропорциональный константе

$$f(\theta) \propto c.$$

Несобственный прайор

Такое априорное может быть НЕ распределением (!)

$$\int f(\theta) d\theta = \infty.$$

А апостериорное — уже возможно распределение

$$\int f(\theta | x^n) \propto \mathcal{L}_n(\theta).$$



Нормировочная константа

Q: Что технически самое сложное в теореме Байеса?

$$p(\theta | x^n) = \frac{p(x^n | \theta) p(\theta)}{\int p(x^n | \theta) p(\theta) d\theta}.$$

A: Посчитать интеграл:



Пайплайн

При этом “в идеале” мы хотели бы “поточное” обновление параметров по мере поступления новых данных:

$$\text{Apriori}_1 \xrightarrow{\text{Data}_1} \text{Aposteriori}_1 = \text{Apriori}_2 \xrightarrow{\text{Data}_2} \text{Aposteriori}_2 = \dots$$

Рассмотрим случай, когда всё хорошо считается аналитически.

Сопряжённые распределения

1 Теорема Байеса

2 Правила суммы и произведения

3 Байесовская вероятность

- Субъективная неопределённость
- Байесовский вывод
- Выбор априорного распределения

4 Сопряжённые распределения

5 Трамвай

6 Применимость байесовского подхода

Сопряжённые распределения

Пусть

- Априорное распределение $p(\theta)$ — из семейства \mathcal{A} ,
- Функция правдоподобия $p(x^n | \theta)$ — из семейства \mathcal{B} .

Семейства \mathcal{A} и \mathcal{B} — **сопряжённые**, если апостериорное распределение $p(\theta | x^n)$ тоже из семейства \mathcal{A} .

Сопряжённые распределения

Пример. Рассмотрим бросание n монеток, ему соответствует $\text{Binomial}(n, \theta)$.

- Априорное распределение

$$\text{Beta}(a, b) = C_1 \theta^{a-1} (1 - \theta)^{b-1}.$$

Константы не важны — для проверки сопряжения достаточно смотреть на вид выражения относительно θ .

Сопряжённые распределения

Пример. Рассмотрим бросание n монеток, ему соответствует $\text{Binomial}(n, \theta)$.

- Априорное распределение

$$\text{Beta}(a, b) = C_1 \theta^{a-1} (1 - \theta)^{b-1}.$$

Константы не важны — для проверки сопряжения достаточно смотреть на вид выражения относительно θ .

- Функция правдоподобия, если “выпало k орлов”:

$$\binom{n}{k} \theta^k (1 - \theta)^{n-k}.$$

Q: Какой вид у апостериорного?

Сопряжённые распределения

Пример. Рассмотрим бросание n монеток, ему соответствует $\text{Binomial}(n, \theta)$.

- Априорное распределение

$$\text{Beta}(a, b) = C_1 \theta^{a-1} (1 - \theta)^{b-1}.$$

Константы не важны — для проверки сопряжения достаточно смотреть на вид выражения относительно θ .

- Функция правдоподобия, если “выпало k орлов”:

$$\binom{n}{k} \theta^k (1 - \theta)^{n-k}.$$

Q: Какой вид у апостериорного?

- Апостериорное распределение

$$\text{Beta}(a + k, b + n - k) = C_2 \theta^{a+k-1} (1 - \theta)^{b+n-k-1}.$$

Сопряжённые распределения

Список сопряжённых распределений можно найти в Wikipedia:

https://en.wikipedia.org/wiki/Conjugate_prior



Трамвай

1 Теорема Байеса

2 Правила суммы и произведения

3 Байесовская вероятность

- Субъективная неопределённость
- Байесовский вывод
- Выбор априорного распределения

4 Сопряжённые распределения

5 Трамвай

6 Применимость байесовского подхода

Трамвай!

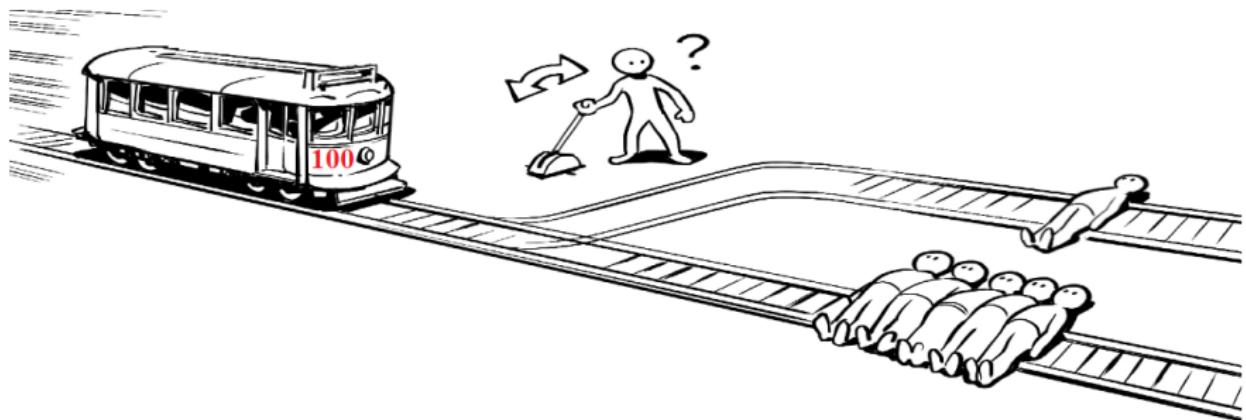
Давайте поговорим про трамвай



Рис.: Да, про трамвай

Tramcar problem

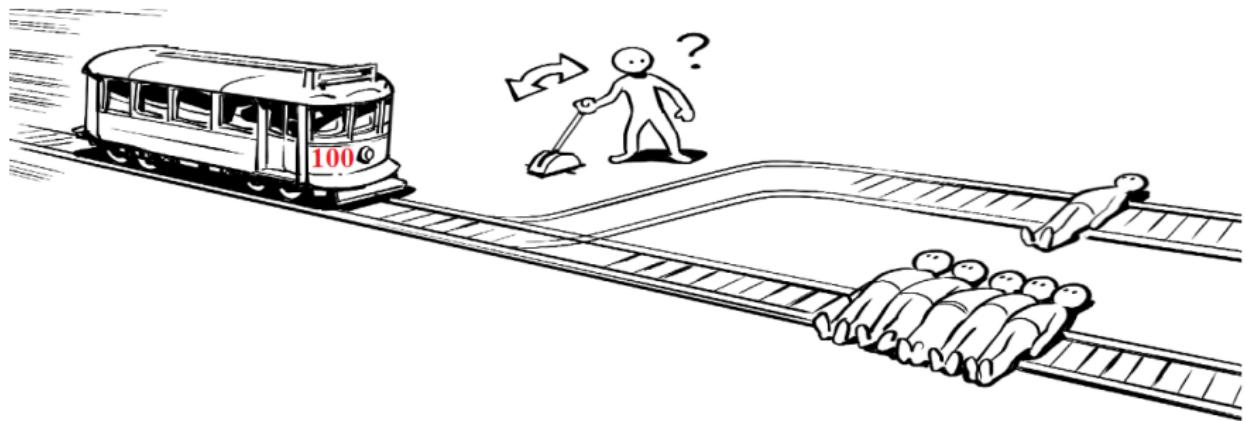
Приходите в город и видите трамвай под **номером 100**.



Естественно, у любого математика тут же возникает вопрос:

Tramcar problem

Приходите в город и видите трамвай под **номером 100**.



Естественно, у любого математика тут же возникает вопрос:

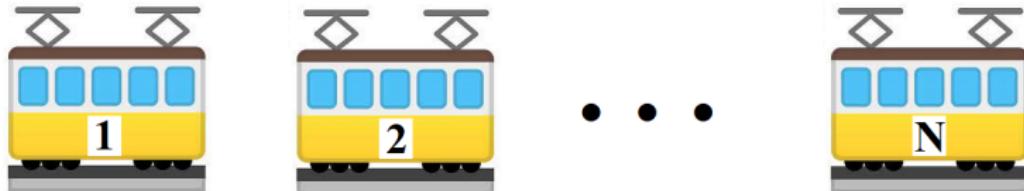
“Сколько трамвайных маршрутов в городе?”

Tramcar problem

Нужно выбрать (статистическую) модель для описания реальности.

Предположение. В городе есть трамваи под номерами $1, \dots, N$.

Трамваи независимы. Номера, которые мы видим, равновероятны.



Частотный подход

Частотный подход.

Наблюдаем за номерами трамваев X_1, \dots, X_n .

Оцениваем N как статистику от них.

Скажем, оценка максимального правдоподобия:

$$N_{ML} = X_{(n)} = \max(X_1, \dots, X_n).$$

Логичный вывод: количество маршрутов — самый большой номер N , что мы увидим за долгое время.

Байесовский подход

Q: С чем сопрягается дискретное равномерное распределение?

Смотрим Wikipedia — беда!

When likelihood function is a discrete distribution [edit]
 This section needs additional citations for verification. Please help improve it.
2020) (Learn how and when to remove this template message)

| Likelihood | Model parameters | Conjugate prior distribution | Prior hyperparameters |
|--|--|------------------------------|--------------------------------|
| Bernoulli | p (probability) | Beta | $\alpha, \beta \in \mathbb{R}$ |
| Binomial | p (probability) | Beta | $\alpha, \beta \in \mathbb{R}$ |
| Negative binomial with known failure number, r | p (probability) | Beta | $\alpha, \beta \in \mathbb{R}$ |
| Poisson | λ (rate) | Gamma | $k, \theta \in \mathbb{R}$ |
| | | | $\alpha, \beta^{[note 4]}$ |
| Categorical | p (probability vector), k (number of categories, i.e., size of p) | Dirichlet | $\alpha \in \mathbb{R}^k$ |
| Multinomial | p (probability vector), k (number of categories, i.e., size of p) | Dirichlet | $\alpha \in \mathbb{R}^k$ |
| Hypergeometric with known total population size, N | M (number of target members) | Beta-binomial ^[3] | $n = N, \alpha, \beta$ |
| Geometric | p_0 (probability) | Beta | $\alpha, \beta \in \mathbb{R}$ |

Нет дискретного равномерного!

Байесовский подход

Ладно, меняем — упрощаем модель.

Трамваи — непрерывно равномерно распределены от 0 до N : $U[0, N]$.

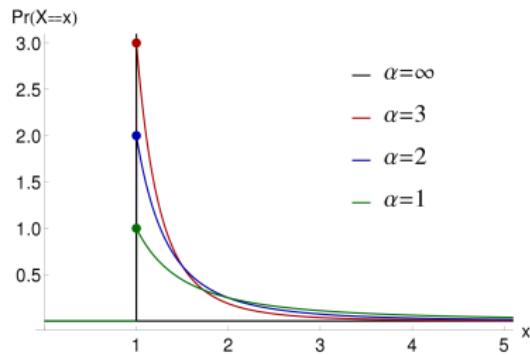


Распределение Парето

Смотрим в Wiki — что сопрягается с $U[0, N]$?

Распределение Парето:

$$\text{Pareto}(\theta|a, b) = \begin{cases} \frac{ba^b}{\theta^{b+1}}, & \theta \geq a, \\ 0, & \text{иначе.} \end{cases}$$



Распределение Парето

Распределение Парето часто используется в социологии и экономике.

Изначально создавалось, чтобы описать распределение богатства по принципу

“20% самых богатых владеет 80% богатств”

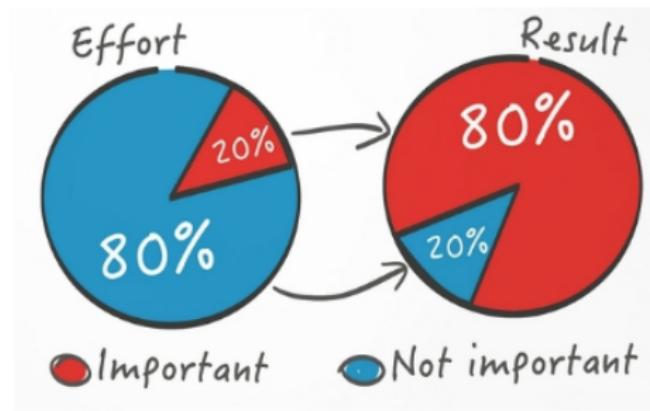


Рис.: Закон Парето — принцип 80/20

Функция правдоподобия

Трамваи равномерны:

$$p(x) = \frac{1}{\theta}, \quad x \in [0, \theta]$$

Находим функцию правдоподобия, если наблюдали M трамваев x_1, \dots, x_M :

$$L_\theta(x) = \prod_{i=1}^M p(x_i) = \begin{cases} \frac{1}{\theta^M}, & 0 \leq x_i \leq \theta, \quad i = 1, \dots, M, \\ 0, & \text{иначе.} \end{cases}$$

Апостериорное распределение

- Априорное:

$$p(\theta) = \frac{ba^b}{\theta^{b+1}} I_{\theta \geq a}$$

- Правдоподобие:

$$p(x_1, \dots, x_M | \theta) = \frac{1}{\theta^M} I_{\{0 \leq x_{(1)} \leq x_{(M)} \leq \theta\}}.$$

Апостериорное распределение

- Априорное:

$$p(\theta) = \frac{ba^b}{\theta^{b+1}} I_{\theta \geq a}$$

- Правдоподобие:

$$p(x_1, \dots, x_M | \theta) = \frac{1}{\theta^M} I_{\{0 \leq x_{(1)} \leq x_{(M)} \leq \theta\}}.$$

- По Теореме Байеса

$$p(\theta | x_1, \dots, x_M) \propto p(x_1, \dots, x_M | \theta) p(\theta)$$

апостериорное

$$p(\theta | x_1, \dots, x_M) \propto \frac{ba^b}{\theta^{M+b+1}} I_{\{\theta \geq \max(a, x_{(M)})\}}.$$

Физический смысл параметров

Параметры апостериорного распределения:

$$a' = \max(a, x_{(M)}), \quad b' = b + M.$$

Присмотревшись — видим смысл параметров!

Физический смысл параметров

Параметры апостериорного распределения:

$$a' = \max(a, x_{(M)}), \quad b' = b + M.$$

Присмотревшись — видим смысл параметров!

- a — наибольший номер, что мы видели;
- b — общее число увиденных трамваев.

'IF YOU GAZE FOR LONG INTO AN ABYSS,
THE ABYSS GAZES ALSO INTO YOU.'

— NIETZSCHE



Физический смысл параметров

Упражнение. Для распределения Pareto($\theta|a, b$)

- Мода равна a
- Матожидание равно $\frac{ba}{b - 1}$.

Как ни смотри: количество маршрутов — самый большой номер N , что мы увидим за долгое время.

Применимость байесовского подхода

1 Теорема Байеса

2 Правила суммы и произведения

3 Байесовская вероятность

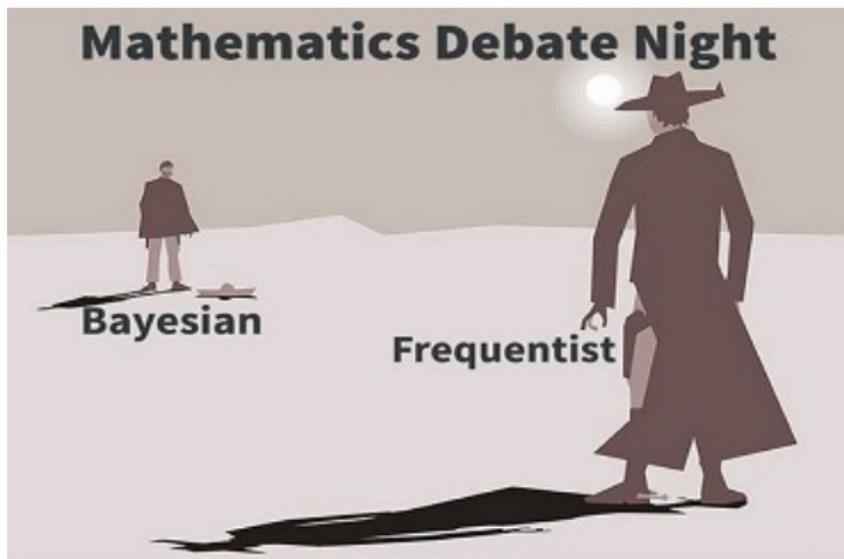
- Субъективная неопределённость
- Байесовский вывод
- Выбор априорного распределения

4 Сопряжённые распределения

5 Трамвай

6 Применимость байесовского подхода

Обсудим плюсы и минусы байесовского подхода.



WARNING! Это холиварная тема!

Преимущества и недостатки

Частотные оценки верны
для больших выборок.

Байесовские оценки -
для любой выборки.

Даже размера 0 (!)

Тяжело считать.

Какой прайер брать?

Получаем какую-то
субъективную (!)
"степень уверенности".



Ещё плюсы и минусы Байесовского подхода:



- 1. Допускает пропуски в данных**
- 2. Естественное обновление параметров**



- Плохо работает при:**
- 1. Больших выборках**
 - 2. Плохой функции правдоподобия**

В пределе подходы совпадают

На больших выборках байесовский подход согласуется с классическим частотным.

Мода апостериорного θ_{MP} (MAP-оценка) стремится к оценке максимального правдоподобия:

$$\lim_{n \rightarrow \infty} \theta_{MP} = \theta_{ML}$$

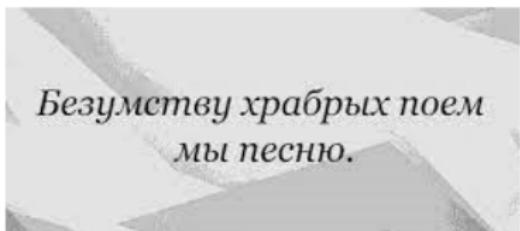
Подробнее — см. Теорему 11.5 в



Wasserman L.
All of Statistics.

Резюме

Байесовский подход — зачастую не самый эффективный, но по-своему интригующий и завораживающий взгляд на ML.

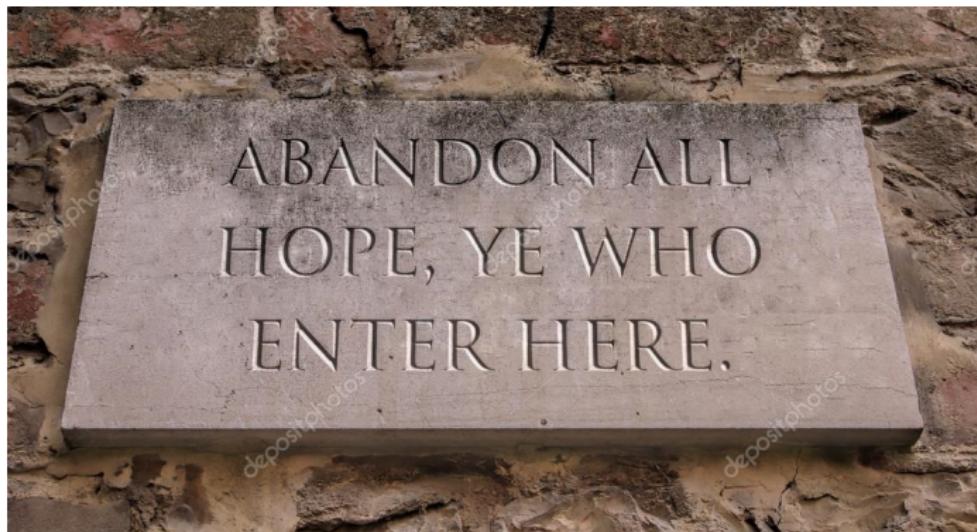


Часто VS **Субъективно**

Abandon All Hope

Мы начали наше приключение в “Байесовском мире”.

Что нас ждёт в конце?



To Be Continued  