

Прикладная статистика в машинном обучении
Семинар 8
Байесовские методы

И. К. Козлов
(Мехмат МГУ)

2022

Ошибка прокурора

1 Ошибка прокурора

2 Парадокс Монти-Холла

3 Графические модели

4 Медицинский тест

5 QDA и LDA

6 Наивный Байес

Prosecutor's fallacy

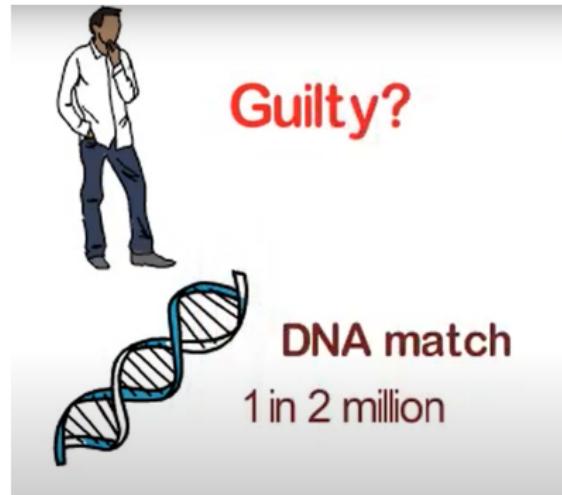
Ошибка прокурора



Ошибка прокурора

- ДНК подсудимого совпал с ДНК убийцы.
- Вероятность этого 1 на 2 млн.
- Эрго: подсудимый виновен.

Вероятность подобного события при его невиновности астрономически мала.



Ошибка прокурора

Q: Где обман?



Ошибка прокурора

Пусть $E = \text{Evidence}$, $I = \text{Innocence}$.

Ошибка прокурора:

$$\mathbb{P}(I | E) \neq \mathbb{P}(E | I)$$

Прокурор не выучил формулу Байеса:

$$\mathbb{P}(I | E) = \mathbb{P}(E | I) \frac{\mathbb{P}(I)}{\mathbb{P}(E)}$$

Ошибка прокурора

Наглядный пример. Пусть вероятность выиграть в лотерею: 1 на 2 млн.

Отсюда \Rightarrow победитель — жулик.

С ДНК тестом в 10 млн городе в среднем 5 человек с тем же ДНК.

Среди них вероятность вины не $\frac{1}{2\text{млн}}$, а 20%.

Ошибка прокурора

Ошибка прокурора — реальная ошибка:

- Салли Кларк — осуждена (1999) из-за смерти двух сыновей в младенчестве.
- Люсия де Берк — медсестра, была осуждена (2003) из-за 13 подозрительных смертей во время её дежурства.

Парадокс Монти-Холла

1 Ошибка прокурора

2 Парадокс Монти-Холла

3 Графические модели

4 Медицинский тест

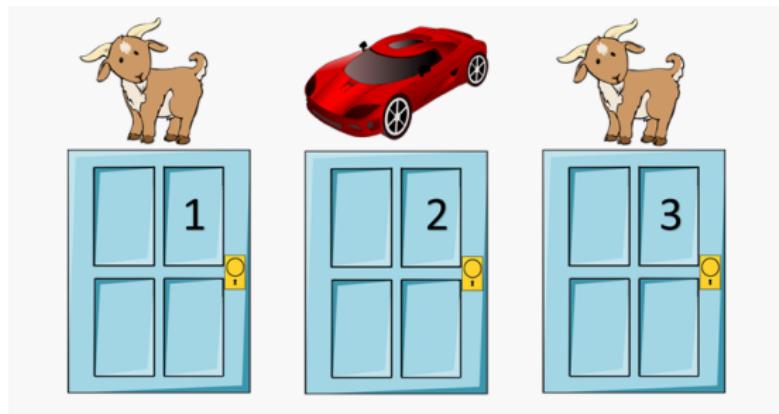
5 QDA и LDA

6 Наивный Байес

Парадокс Монти Холла

Парадокс Монти Холла

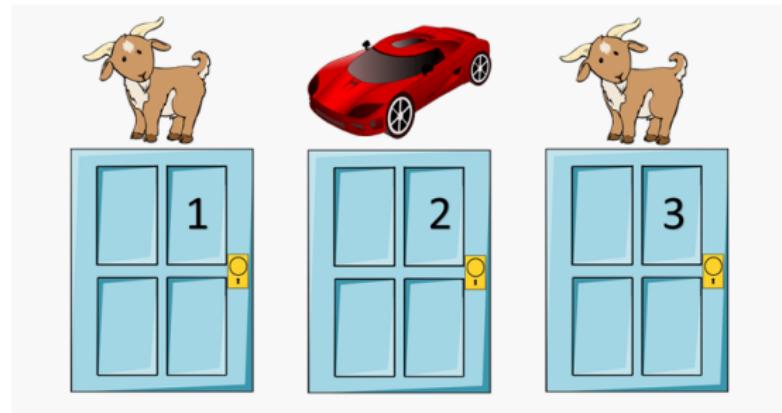
- Игра: нужно выбрать 1 из 3 дверей. За одной дверью - автомобиль, за 2 другими - козы.



Парадокс Монти Холла

Парадокс Монти Холла

- Игра: нужно выбрать 1 из 3 дверей. За одной дверью - автомобиль, за двумя - козы.



- Выбрали дверь. Затем **ведущий открывает одну из оставшихся дверей, за которой находится коза.**
- Ведущий знает, где автомобиль.**
- Q:** Увеличится ли вероятность выиграть, если изменить теперь свой выбор?

Парадокс Монти Холла

Главное — правильно посчитать вероятности и зависимость между событиями:

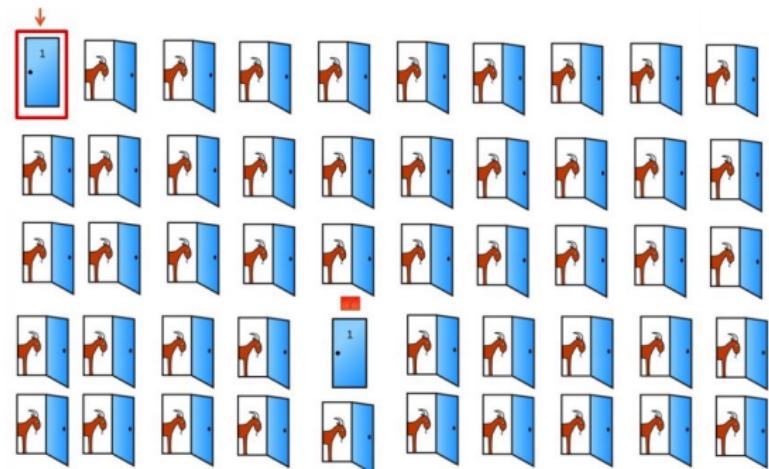
- автомобиль равновероятно размещён за любой из трёх дверей;
- ведущий всегда открывает дверь, за которой коза;
- **ведущий знает, где находится автомобиль;**
- если игрок сразу выбрал правильную дверь, ведущий открывает любую из оставшихся с одинаковой вероятностью.

Парadox Монти Холла

A: Лучше изменить выбор.

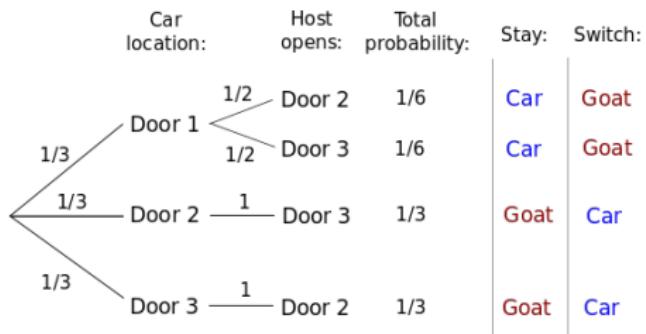
Наглядно — пусть дверей 1000. Открывают 998 из оставшихся 999.

Выбор очевиден?



Парadox Монти Холла

Решение “в лоб” — посчитать все вероятности:



Парадокс Монти Холла

Применим теорему Байеса!

А какие в задаче априорные и апостериорные вероятности?

- Априорные вероятности

$$1 : 1 : 1$$

- Пусть выбираем 1ую дверь. Вероятности не изменились.

Парадокс Монти Холла

- Считаем, что ведущий открыл Зью дверь.
- Q: Правдоподобие? (Вероятности этого события)

Парадокс Монти Холла

- Считаем, что ведущий открыл Зью дверь.
- **Q:** Правдоподобие? (Вероятности этого события)
- Обозначим события:
 - ▶ $A_i = \text{Автомобиль за дверью } i$
 - ▶ $O = \text{открыли дверь } 3.$
- **Правдоподобие:**

$$\mathbb{P}(O | A_i) = (50\%, \quad 100\%, \quad 0\%)$$

Парадокс Монти Холла

- Чему пропорциональна апостериорная вероятность по теореме Байеса:

$$\mathbb{P}(A_i | O) \propto \mathbb{P}(O | A_i) \mathbb{P}(A_i).$$

Парадокс Монти Холла

- Чему пропорциональна апостериорная вероятность по теореме Байеса:

$$\mathbb{P}(A_i | O) \propto \mathbb{P}(O | A_i) \mathbb{P}(A_i).$$

- Апостериорная вероятность:**

1 : 2 : 0

В 2 раза больше вероятность выиграть с другой дверью.

Парадокс Монти Холла

Если Вы выбрали **другую дверь**, то ...



Графические модели

1 Ошибка прокурора

2 Парадокс Монти-Холла

3 Графические модели

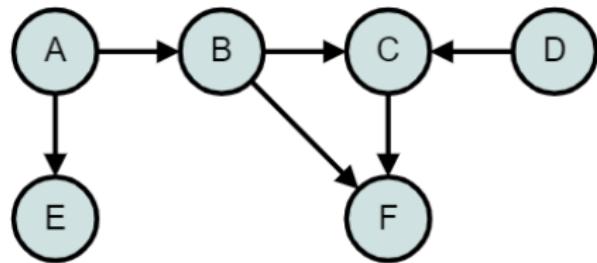
4 Медицинский тест

5 QDA и LDA

6 Наивный Байес

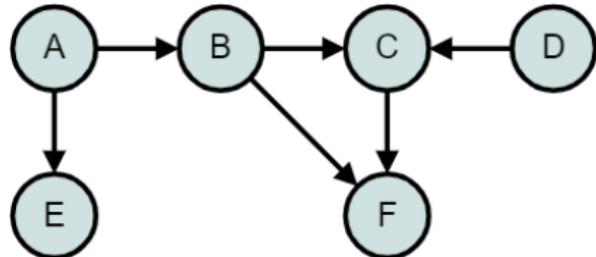
Графические модели

Зависимость между переменными удобно изображать **ориентированным ациклическим графом** (directed acyclic graph = DAG) $G = (V, D)$.



Графические модели

Зависимость между переменными удобно изображать **ориентированным ациклическим графом** (directed acyclic graph = DAG) $G = (V, D)$.



- Вершины $v \in V$ соответствуют случайным величинам.
- Рёбра $e \in D$ обозначают статистическую зависимость.

В графе **не должно быть циклов**.

- Для каждой вершины задаётся условная вероятность

$$p(x_i | x_{\text{parents}(i)}).$$

Оказывается, мы говорили прозой...

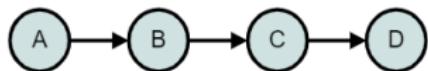
DAG — настолько естественное понятие, что
мы уже использовали его на прошлом занятии:



Оказывается, мы говорили прозой...

Полная вероятность **факторизуется** — распадается в произведения локальных распределений:

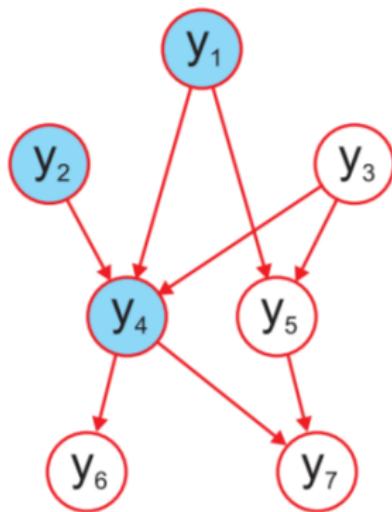
$$p(x_1, \dots, x_n) = \prod_{i \in V} p(x_i | x_{\text{parents}(i)})$$



Например, для графа на рис.

$$p(x_A, x_B, x_C, x_D) = p(x_A)p(x_B|x_A)p(x_C|x_B)p(x_D|x_C).$$

- Пусть нам необходимо найти распределение (y_5, y_7) при заданных значениях 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:

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

Маргинализация и обуславливание

- По определению условной вероятности

$$p(y_5, y_7 | y_1, y_2, y_4) = \frac{p(y_1, y_2, y_4, y_5, y_7)}{p(y_1, y_2, y_4)}$$

- Расписываем знаменатель

$$p(y_1, y_2, y_4) = p(y_1)p(y_2)p(y_4 | y_1, y_2) = \{\text{Sum rule}\}$$

$$p(y_1)p(y_2) \int p(y_4 | y_1, y_2, y_3)p(y_3)dy_3$$

- Аналогично числитель

$$\begin{aligned} p(y_1, y_2, y_4, y_5, y_7) &= p(y_1)p(y_2)p(y_4 | y_1, y_2)p(y_5 | y_1)p(y_7 | y_5, y_4) = p(y_1) \times \\ &p(y_2) \left(\int p(y_4 | y_1, y_2, y_3)p(y_3)dy_3 \right) \left(\int p(y_5 | y_1, y_3)p(y_3)dy_3 \right) p(y_7 | y_5, y_4) \end{aligned}$$

Медицинский тест

1 Ошибка прокурора

2 Парадокс Монти-Холла

3 Графические модели

4 Медицинский тест

5 QDA и LDA

6 Наивный Байес

Медицинский тест

- Болезнью болеет 10% населения

$$\mathbb{P}(D) = 10\%, \quad \mathbb{P}(\neg D) = 90\%$$

- Тест правильно определяет болен/не болен в 90% случаев:

$$\mathbb{P}(+ | D) = 90\%, \quad \mathbb{P}(- | \neg D) = 90\%$$

Медицинский тест

- Болезнью болеет 10% населения

$$\mathbb{P}(D) = 10\%, \quad \mathbb{P}(\neg D) = 90\%$$

- Тест правильно определяет болен/не болен в 90% случаев:

$$\mathbb{P}(+ | D) = 90\%, \quad \mathbb{P}(- | \neg D) = 90\%$$

- Тест положительный! Q: Какова вероятность болезни?

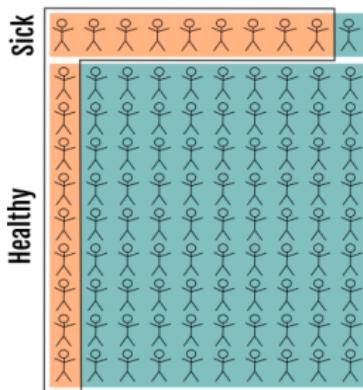
Медицинский тест

A: Теорема Байеса показывает, что не такая большая:

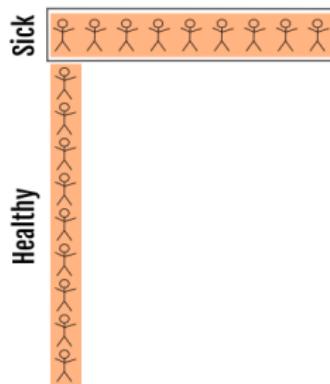
$$\begin{aligned}\mathbb{P}[D|+] &= \frac{\mathbb{P}(+ | D)\mathbb{P}(D)}{\mathbb{P}(+ | D)\mathbb{P}(D) + \mathbb{P}(+ | \neg D)\mathbb{P}(\neg D)} = \\ &= \frac{0.09}{0.09 + 0.09} = 50\%\end{aligned}$$

Медицинский тест

Как “графически” выглядит теорема Байеса:

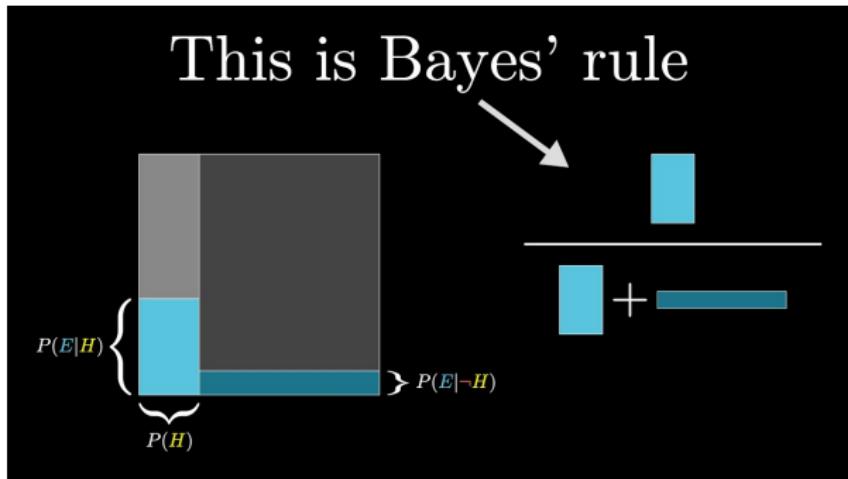


$$\#(\text{Positive}) = 18$$



$$p(\text{Sick} \mid \text{Positive}) = 9/18$$

Теорема Байеса графически



Перерыв

Перерыв

QDA и LDA

1 Ошибка прокурора

2 Парадокс Монти-Холла

3 Графические модели

4 Медицинский тест

5 QDA и LDA

6 Наивный Байес

ML пример

Сегодня мы уже применяли **теорему Байеса** для классификации:

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

Аналогично в общем виде:

$$P(y = k|x) = \frac{P(x|y = k)P(y = k)}{P(x)} = \frac{P(x|y = k)P(y = k)}{\sum_l P(x|y = l) \cdot P(y = l)}$$

Q: Какое самое правдоподобие $P(x | y)$ взять?

QDA

- Quadratic Discriminant Analysis (QDA):

$$P(x|y = k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^\top \Sigma_k^{-1}(x - \mu_k)\right)$$

QDA

- Quadratic Discriminant Analysis (QDA):

$$P(x|y = k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right)$$

- QDA очень прост в вычислении:

$$\begin{aligned}\log P(y = k|x) &= \log P(x|y = k) + \log P(y = k) + \text{const} = \\ &= -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \log P(y = k) + \text{const},\end{aligned}$$

LDA

- Linear Discriminant Analysis (LDA) — предполагаем, что все матрицы ковариации совпадают $\Sigma_k = \Sigma$.

Формулы становятся линейными:

$$\log P(y = k|x) = -\frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k) + \log P(y = k) + \text{const.}$$

LDA

- Linear Discriminant Analysis (LDA) — предполагаем, что все матрицы ковариации совпадают $\Sigma_k = \Sigma$.

Формулы становятся линейными:

$$\log P(y = k|x) = -\frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k) + \log P(y = k) + \text{const.}$$

Логарифм апостериорного линеен по x :

$$\log P(y = k|x) = \omega_k^T x + \omega_{k0} + \text{const.}$$

$$\omega_k = \Sigma^{-1} \mu_k, \quad \omega_{k0} = -\frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log P(y = k).$$

QDA

Реализация QDA и LDA:

https://scikit-learn.org/stable/modules/lda_qda.html

QDA

Логистическая регрессия и LDA оба приводят к линейному классификационному правилу

$$\log \left(\frac{\mathbb{P}(Y = 1 | X = x)}{\mathbb{P}(Y = 0 | X = x)} \right) = \beta_0 + \beta_1 x$$

Разница: LogReg — дискриминативная модель, а LDA — генеративная.

$$\prod_i f(x_i, y_i) = \underbrace{\prod_i f(x_i | y_i)}_{\text{Gaussian}} \underbrace{\prod_i f(y_i)}_{\text{Bernoulli}}.$$

$$\prod_i f(x_i, y_i) = \underbrace{\prod_i f(y_i | x_i)}_{\text{logistic}} \underbrace{\prod_i f(x_i)}_{\text{ignored}}.$$

Наивный Байес

1 Ошибка прокурора

2 Парадокс Монти-Холла

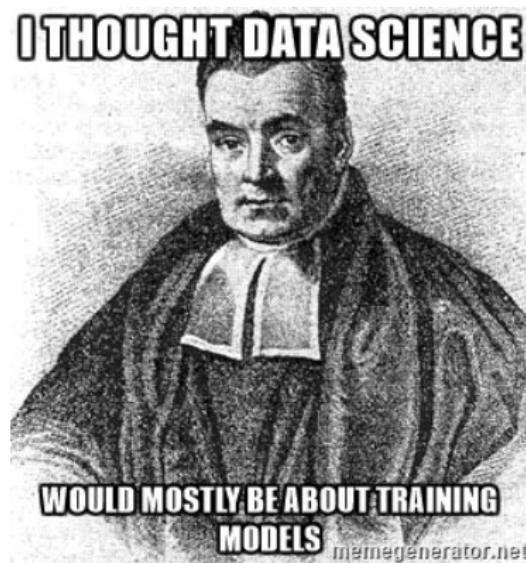
3 Графические модели

4 Медицинский тест

5 QDA и LDA

6 Наивный Байес

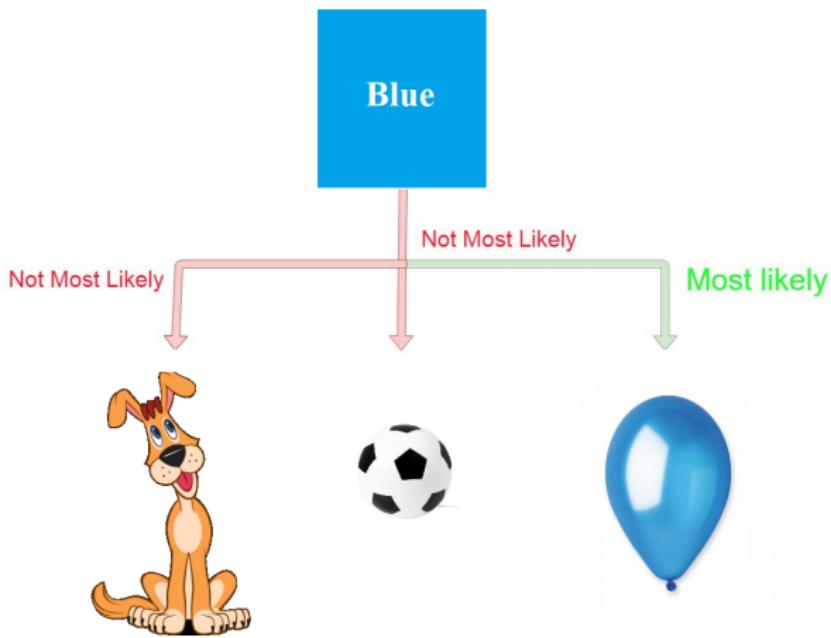
Naive Bayes classifier



Наивный Байес

Наивный пример.

- У нас есть 3 класса — пёсик, мячик и воздушный шарик.
- Мы знаем, что объект — синий. Q: Какой класс наиболее вероятен?



Наивный Байес

- Итак, знаем фичу \Rightarrow классы более или менее вероятны.
- Q: А что делать, если фичей несколько?

Наивный Байес

- Итак, знаем фичу \Rightarrow классы более или менее вероятны.
- **Q:** А что делать, если фичей несколько?
- **A:** Самое простое — перемножить вероятности, как будто фичи независимы.

Naive Bayes

- Даны n фичей — вектор $\mathbf{x} = (x_1, \dots, x_n)$. Мы хотим вычислить вероятность класса:

$$P(y | x_1, \dots, x_n) = \frac{P(y)P(x_1, \dots, x_n | y)}{P(x_1, \dots, x_n)}$$

- Предположение **Наивного Байеса** — фичи условно независимы

$$P(x_i | y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y).$$

Naive Bayes

Получаем классификационное правило:

$$\begin{aligned} P(y | x_1, \dots, x_n) &\propto P(y) \prod_{i=1}^n P(x_i | y) \\ &\Downarrow \\ \hat{y} &= \arg \max_y P(y) \prod_{i=1}^n P(x_i | y), \end{aligned} \tag{1}$$

Naive Bayes

Реализация Naive Bayes:

https://scikit-learn.org/stable/modules/naive_bayes.html

Naive Bayes

- Пусть X_1, \dots, X_n — независимые нормальные распределения $X_k \sim \mathcal{N}(\mu_i, \sigma_i^2)$.
- **Q:** Какое распределение у вектора (X_1, \dots, X_n) ?
- **A:** Многомерное нормальное распределение $\mathcal{N}(\mu, \Sigma)$, где

$$\mu = (\mu_1, \dots, \mu_n), \quad \Sigma = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{pmatrix}.$$

Gaussian Naive Bayes

- Непрерывные фичи можно приближать гауссовским распределением.
- **Q:** Какое распределение у классов?

Gaussian Naive Bayes

- Непрерывные фичи можно приближать гауссовским распределением.
- **Q:** Какое распределение у классов?
- **A:** Нормальное с одной и той же **диагональной** матрицей Σ .

Gaussian Naive Bayes =
= LDA with diagonal Σ

Наивный Байес — неплохой бейзлайн

При всей своей “наивности”, Naive Bayes обладает рядом **существенных преимуществ**:

- Очень простой. Обучается даже на малых выборках. **Неплохой бейзлайн.**
- Очень быстрый. Даже на очень больших данных.

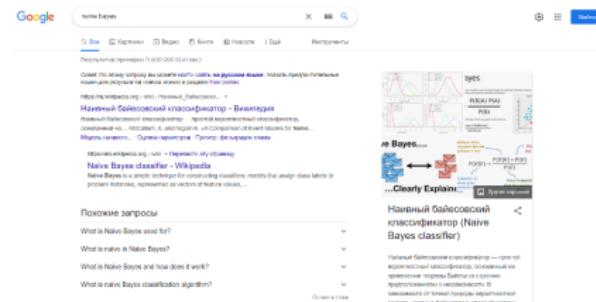


Когда скорость решает

Q: Что в ИТ нужно классифицировать очень быстро, “влёт”?



Spam filtering



A screenshot of a Google search results page for the query "朴素贝叶斯分类器". The top result is a link to a page titled "朴素贝叶斯分类器 - Википедия", which is described as "Начальный байесовский классификатор — простой вероятностный классификатор, основанный на законе Байеса, использующий условные概率 probabilities for making predictions." Below the link is a snippet of text explaining the algorithm. To the right of the search results, there is a sidebar titled "朴素贝叶斯分类器 (朴素 Bayes classifier)" with a "Clearly Explain" button. The sidebar contains a diagram illustrating the Naive Bayes classification process, showing how it takes input features and assigns them to a class based on prior probabilities and likelihood ratios.

Ranking / Recommendations