# Анализ данных (Байесовские методы машинного обучения)

М. В. Лебедев

Московский авиационный институт

19 февраля 2022 г.

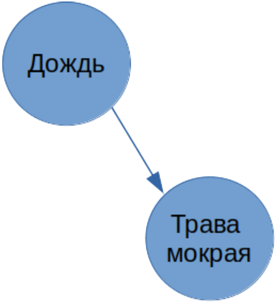
# Байесовские сети

# Байесовские сети

## Определение

**Байесовская сеть** — это ориентированный ациклический граф, каждой вершине которого соответствует случайная переменная, а дуги графа кодируют отношения условной независимости между этими переменными.

- **Узлы графа** — случайные переменные.
- **Дуги графа** — условия.

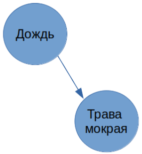


# Байесовские сети

## Определение

**Байесовская сеть** — это ориентированный ациклический граф, каждой вершине которого соответствует случайная переменная, а дуги графа кодируют отношения условной независимости между этими переменными.

- **Узлы графа** — случайные переменные.
- **Дуги графа** — условия.



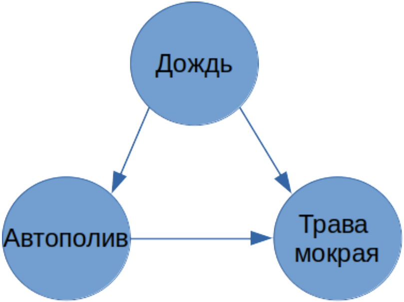

$$P(D, T) = P(T|D)\,P(D),\ \ D \text{ — дождь},\ \ T \text{ — Трава мокрая.}$$

# Байесовские сети

## Определение

**Байесовская сеть** — это ориентированный ациклический граф, каждой вершине которого соответствует случайная переменная, а дуги графа кодируют отношения условной независимости между этими переменными.

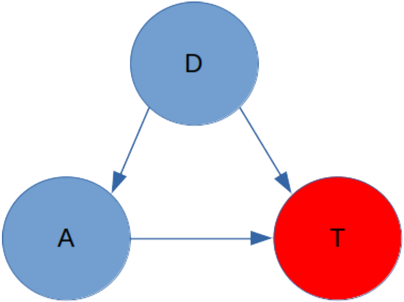
- **Узлы графа** — случайные переменные.
- **Дуги графа** — условия.



# Байесовские сети (Вероятностная модель)

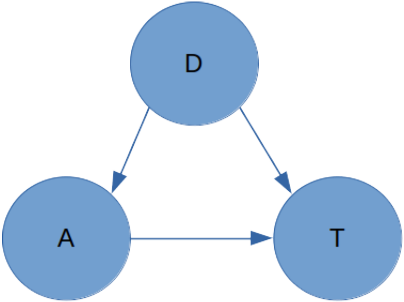
## Совместное распределение

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_k | \underbrace{Pa(X_k)}_{\text{Родители}}) \quad (1)$$



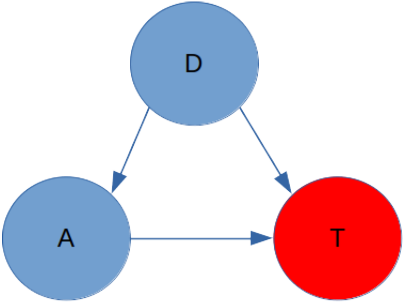

$$Pa(T) = \{D, A\}$$

# Вероятностная модель



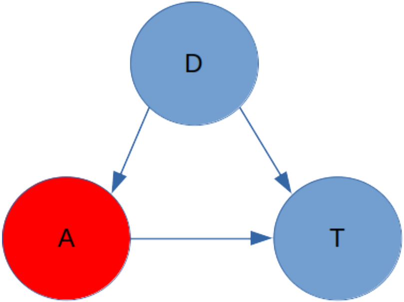

$$\mathrm{P}(A, D, T) = \ldots$$

# Вероятностная модель



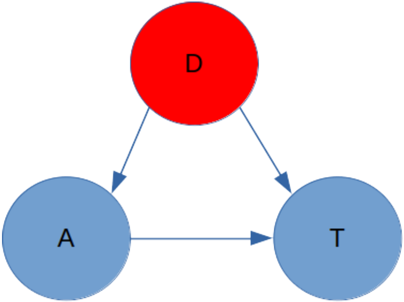

$$\mathrm{P}(A, D, T) = \color{red}{\mathrm{P}(T|D, A)} \cdots$$

# Вероятностная модель



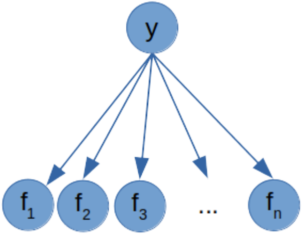

$$\mathrm{P}(A, D, T) = \mathrm{P}(T|D, A) \cdot \color{red}{\mathrm{P}(A|D)} \cdots$$

# Вероятностная модель




$$\mathrm{P}(A, D, T) = \mathrm{P}(T|D, A) \cdot \mathrm{P}(A|D) \cdot \mathrm{P}(D)$$

# Наивный Байесовский классификатор




$$p(y, f) = p(y, f_1, f_2, f_3, \ldots, f_n) = \mathrm{P}(y) \prod_{i=1}^{n} p(f_i|y)$$

# Байесовский классификатор

# Байесовский классификатор

- **Заданы**: $X$ — объекты, $Y$ — классы, $X \times Y$ — в.п. с плотностью $p(x, y)$. $X_n = (x_i, y_i)_{i=1}^n \sim p(x, y)$ — простая выборка.
- **Найти:** $a : X \to Y$ с минимальной вероятностью ошибки.

Известна совместная плотность

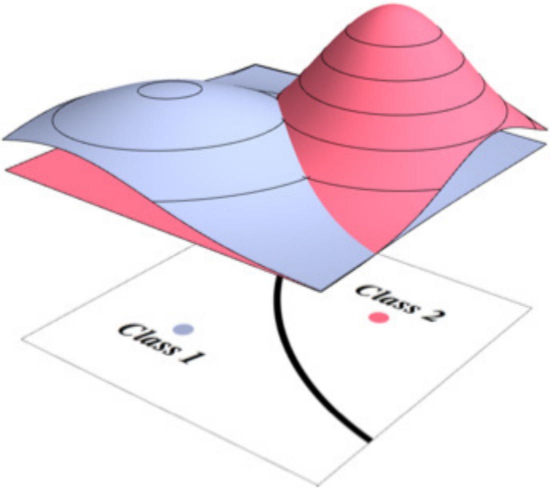
$$p(x, y) = p(x)P(y|x) = P(y)p(x|y).$$

$P(y)$ — априорная вероятность класса $y$.
$p(x|y)$ — функция правдоподобия класса $y$.
$P(y|x)$ — апостериорная вероятность класса $y$.
По формуле Байеса: $P(y|x) = \frac{P(y)p(x|y)}{p(x)}$
**Байеовский классификатор**:

$$a(x) = \arg\max_{y \in Y} P(y|x) = \arg\max_{y \in Y} P(y)p(x|y)$$

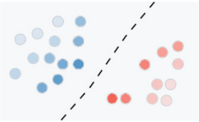
**Частный случай:**

$$a(x) = \arg\max p(x|y) \text{ при равных } P(y)$$

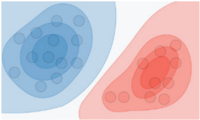


# Два подхода к обучению классификации

1. **Дискриминативный** (discriminative):
   $x$ — неслучайные векторы, $P(y|x, w)$ — модель классификациию
   **Примеры:** LR, GLM, SVM, RBF.



2. **Генеративный** (generative):
   $x \sim p(x|y)$ — случайные векторы. $p(x|y, \theta)$ — модель генерации данных.
   **Примеры:** GMM, FLD, RBF.



**Байесовские модели классификации — генеративные**

- меделируют форму классов не только вдоль границы, но и на всем пространстве, что избыточно для классификации;
- требуют больше данных для обучения;
- более устойчивы к шумовым выбросам.

# Оптимальный байесовский классификатор

**Теорема**

Пусть $P(y)$ и $p(x|y)$ известны, $\lambda_y > 0$ — потеря на ошибке класса $y \in Y$ . Тогда минимум среднего риска

$$R(a) = \sum_{y \in Y} \lambda_y \int [a(x) \neq y] p(x, y) dx$$

достигается оптимальным байесовским классификатором

$$a(x) = \arg\max_{y \in Y} \lambda_y P(y) p(x|y).$$

**Замечание 1:** после подстановки эмпирических оценок $\hat{P}(y)$, $\hat{p}(x|y)$ байесовский классифиатор уже не оптимален.
**Замечание 2:** задача оценки плотности распределения – более сложная, чем задача классификации.

# Наивный байесовский классификатор (Naive Bayes)

**Наивное предположение**

Признаки $f_j : X \to D_j$ — независимые случайные величины с плотностью распределения $p_j(x|y)$, $y \in Y$, $j = 1, \dots, n$.

Тогда функция правдоподобия классов представимы в виде произведения одномерных плотностей по признакам $f_j$:

$$p(f|y) = p(f_1|y) \cdot p(f_2|y) \cdots p(f_n|y), \quad f = (f_1, \dots, f_n), \quad y \in Y.$$

Прологарифмировав под arg max, получим классификатор:

$$a(x) = \arg\max_{y \in Y} \left( \log \lambda_y \hat{P}(y) + \sum_{j=1}^{n} \log \hat{p}_j(f_j|y) \right).$$

# Больше информации

Больше информации можно найти тут:

- Байесовская теория классификации
- Машинное обучение (курс лекций, К.В.Воронцов)

# Сопряжённое априорное распределение

# Апостериорное распределение

$$\underbrace{\mathsf{P}(\theta|X)}_{\text{Апостериорная}} = \frac{\overbrace{\mathsf{P}(X|\theta)}^{\text{Правдоподобие}} \cdot \overbrace{\mathsf{P}(\theta)}^{\text{Априорная}}}{\underbrace{\mathsf{P}(X)}_{\text{Маргинальная (Evidence)}}}$$

Что такое $\mathsf{P}(X)$?

# Апостериорное распределение

$$\underbrace{P(\theta|X)}_{\text{Апостериорная}} = \frac{\overbrace{P(X|\theta)}^{\text{Правдоподобие}} \cdot \overbrace{P(\theta)}^{\text{Априорная}}}{\underbrace{P(X)}_{\text{Маргинальная (Evidence)}}}$$

Что такое $P(X)$?

- Изображение, текст, сигнал и т.п.

$$\underbrace{\mathrm{P}(\theta|X)}_{\text{Апостериорная}} = \frac{\overbrace{\mathrm{P}(X|\theta)}^{\text{Правдоподобие}} \cdot \overbrace{\mathrm{P}(\theta)}^{\text{Априорная}}}{\underbrace{\mathrm{P}(X)}_{\text{Маргинальная (Evidence)}}}$$

Что такое $\mathrm{P}(X)$?

- Изображение, текст, сигнал и т.п.
- Обычно очень тяжело получить.

# Максимум апостериорной вероятности

Можно упростить задачу:

$$\hat{\theta}_{MA} \in \arg\max_{\theta} \mathsf{P}(\theta|X) \Rightarrow$$

$$\hat{\theta}_{MA} \in \arg\max_{\theta} \frac{\mathsf{P}(X|\theta)\,\mathsf{P}(\theta)}{\mathsf{P}(X)} \Rightarrow$$

$$\hat{\theta}_{MA} \in \arg\max_{\theta} \mathsf{P}(X|\theta)\,\mathsf{P}(\theta).$$

Таким образом, получена оптимизационная задача.

## Проблемы с MA

1. Чувствительность к изменению параметризации

$$\nu = f(\theta) \Rightarrow \theta_{MA} \neq \hat{\nu}_{MA}$$

# Проблемы с МА

1. Чувствительность к изменению параметризации

$$\nu = f(\theta) \Rightarrow \theta_{MA} \neq \hat{\nu}_{MA}$$

2. Проблем с объединением информации, т.е. вычислении априорной вероятности для следующего шага:

$$p_k(\theta) = \frac{p(x_k|\theta)p_{k-1}(\theta)}{p(x_k)},$$

$$p_k(\theta) = \frac{p(x_k|\theta)\delta(\theta - \hat{\theta}_{MA})}{p(x_k)} = \delta(\theta - \hat{\theta}_{MA})$$

# Проблемы с MA

1. Чувствительность к изменению параметризации

$$\nu = f(\theta) \Rightarrow \theta_{MA} \neq \hat{\nu}_{MA}$$

2. Проблем с объединением информации, т.е. вычислении априорной вероятности для следующего шага:

$$p_k(\theta) = \frac{p(x_k|\theta)p_{k-1}(\theta)}{p(x_k)},$$

$$p_k(\theta) = \frac{p(x_k|\theta)\delta(\theta - \hat{\theta}_{MA})}{p(x_k)} = \delta(\theta - \hat{\theta}_{MA})$$

3. Проблема с вычисление оптимального значения $\hat{\theta}_{MA}$.

# Проблемы с МА

1. Чувствительность к изменению параметризации

$$\nu = f(\theta) \Rightarrow \theta_{MA} \neq \hat{\nu}_{MA}$$

2. Проблем с объединением информации, т.е. вычислении априорной вероятности для следующего шага:

$$p_k(\theta) = \frac{p(x_k|\theta)p_{k-1}(\theta)}{p(x_k)},$$

$$p_k(\theta) = \frac{p(x_k|\theta)\delta(\theta - \hat{\theta}_{MA})}{p(x_k)} = \delta(\theta - \hat{\theta}_{MA})$$

3. Проблема с вычисление оптимального значения $\hat{\theta}_{MA}$.
4. Нет возможности построить доверительный интервал для $\hat{\theta}_{MA}$.

# Сопряжённое априорное распределение

$$P(\theta|X) = \frac{\overbrace{P(X|\theta)}^{\text{Задано моделью}} \cdot \overbrace{P(\theta)}^{\text{Наш выбор}}}{\underbrace{P(X)}_{\text{Задается данными}}}$$

# Сопряжённое априорное распределение

## Определение

Семейство распределений $\mathsf{P}(\theta)$ называется сопряжённым семейству функций правдоподобия $\mathsf{P}(X|\theta)$ если апостериорное распределение $\mathsf{P}(\theta|X)$ принадлежит тому же семейству вероятностных распределений, что и априорное распределение $\mathsf{P}(\theta)$:

$$\underbrace{\mathsf{P}(\theta|X)}_{\in\mathcal{A}(v')} = \frac{\mathsf{P}(X|\theta)\cdot\overbrace{\mathsf{P}(\theta)}^{\in\mathcal{A}(v)}}{\mathsf{P}(X)}.$$

# Пример (Гауссовский случай)

$$p(x|\theta) = \mathcal{N}(x|\theta, \sigma^2)$$

$$\mathcal{A}(v) = ?$$

$$\underbrace{p(\theta|X)}_{\in \mathcal{A}(v')} = \frac{\overbrace{p(x|\theta)}^{\mathcal{N}(x|\theta,\sigma^2)} \cdot \overbrace{p(\theta)}^{\in \mathcal{A}(v)}}{p(x)}.$$

# Пример (Гауссовский случай)

$$p(x|\theta) = \mathcal{N}(x|\theta, \sigma^2)$$

$$\mathcal{A}(v) = ?$$

$$\underbrace{p(\theta|x)}_{\in\mathcal{A}(v')} = \frac{\overbrace{p(x|\theta)}^{\mathcal{N}(X|\theta,\sigma^2)} \cdot \overbrace{p(\theta)}^{\mathcal{N}(\theta|m,s^2)}}{\mathrm{P}(x)}.$$

# Пример (Гауссовский случай)

$$p(x|\theta) = \mathcal{N}(x|\theta, \sigma^2)$$

$$\mathcal{A}(v) = \mathcal{N}(\theta|a, b^2)$$

$$\underbrace{p(\theta|x)}_{\mathcal{N}(\theta|a,b^2)} = \frac{\overbrace{p(X|\theta)}^{\mathcal{N}(X|\theta,\sigma^2)} \cdot \overbrace{p(\theta)}^{\mathcal{N}(\theta|m,s^2)}}{p(X)}.$$

# Пример (Гауссовский случай)

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{\mathcal{N}(x|\theta,1)\mathcal{N}(\theta|0,1)}{p(x)}$$

$$p(\theta|x) \propto e^{-\frac{1}{2}(x-\theta)^2} e^{-\frac{1}{2}\theta^2}$$

# Пример (Гауссовский случай)

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{\mathcal{N}(x|\theta,1)\mathcal{N}(\theta|0,1)}{p(x)}$$

$$p(\theta|x) \propto e^{-\frac{1}{2}(x-\theta)^2} e^{-\frac{1}{2}\theta^2} = e^{-\frac{x^2}{2}+\theta x-\theta^2} =$$

$$= e^{-\frac{x^2}{4}+\theta x-\theta^2-\frac{x^2}{4}} = e^{-(\theta-\frac{x}{2})^2-\frac{x^2}{4}} \Rightarrow$$

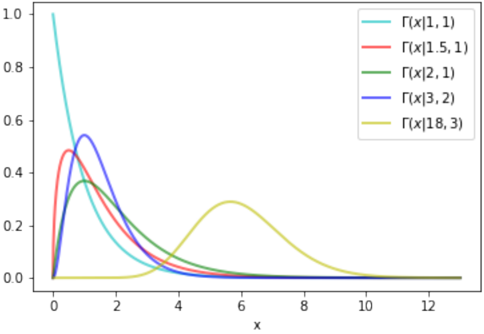
$$p(\theta|x) \propto e^{-(\theta-\frac{x}{2})^2} \Rightarrow p(\theta|x) = \mathcal{N}(\theta|\frac{x}{2},\frac{1}{2})$$

# Гамма распределение

# Гамма распределение

$$\Gamma(x|a,b) = \begin{cases} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

где $\gamma,\ a,\ b > 0, \Gamma(n) = (n-1)!$



# Гамма распределение

$$\Gamma(x|a,b) = \begin{cases} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, & x \geq 0 \\ 0, & x < 0. \end{cases}$$

$$\mathrm{M}\gamma = \frac{a}{b},$$

$$Mode(\gamma) = \frac{a-1}{b},$$

$$\mathrm{D}\gamma = \frac{a}{b^2}.$$

# Точность

**Точность** $\gamma = \frac{1}{\sigma^2}$.

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Rightarrow$$

$$\mathcal{N}(x|\mu,\gamma^{-1}) = \frac{\sqrt{\gamma}}{\sqrt{2\pi}} e^{-\gamma\frac{(x-\mu)^2}{2}} \Rightarrow$$

**Заметим**, что $\mathcal{N}(x|\mu,\gamma^{-1}) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}$.
**Пусть** априорное распределение для параметра аналогичное: $\gamma \sim p(\gamma) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}$, тогда

$$p(\gamma|x) = \frac{p(x|\gamma)p(\gamma)}{p(x)} \propto \gamma e^{-\gamma(b+\frac{(x-\mu)^2}{2})}$$

# Точность

**Точность** $\gamma = \frac{1}{\sigma^2}$.

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Rightarrow$$

$$\mathcal{N}(x|\mu, \gamma^{-1}) = \frac{\sqrt{\gamma}}{\sqrt{2\pi}} e^{-\gamma\frac{(x-\mu)^2}{2}} \Rightarrow$$

**Пусть** априорное распределение для параметра аналогичное:
$\gamma \sim p(\gamma) \propto \gamma^{\frac{1}{2}} e^{-b\gamma}$, тогда

$$\xcancel{p(\gamma|x) = \frac{p(x|\gamma)p(\gamma)}{p(x)} \propto \gamma e^{-\gamma(b + \frac{(x-\mu)^2}{2})}}$$

# Точность (Гамма распределение)

**Точность** $\gamma = \frac{1}{\sigma^2}$.

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Rightarrow$$

$$\mathcal{N}(x|\mu, \gamma^{-1}) = \frac{\sqrt{\gamma}}{\sqrt{2\pi}} e^{-\gamma\frac{(x-\mu)^2}{2}} \Rightarrow$$

**Пусть** теперь априорное распределение для параметра будет **Гамма распределение**: $\gamma \sim p(\gamma) \propto \gamma^{a-1} e^{-b\gamma}$, тогда

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

# Точность (Гамма распределение)

$$p(\gamma) = \Gamma(\gamma | a, b) \propto \gamma^{a-1} e^{-b\gamma}$$

$$p(\gamma | x) \propto p(x|\gamma) p(\gamma) \Rightarrow$$

$$p(\gamma | x) \propto \left(\gamma^{\frac{1}{2}} e^{-\gamma \frac{(x-\mu)^2}{2}}\right) \cdot \left(\gamma^{a-1} e^{-b\gamma}\right) \Rightarrow$$

$$p(\gamma | x) \propto \gamma^{\frac{1}{2}+a-1} e^{-\gamma (b + \frac{(x-\mu)^2}{2})} \Rightarrow$$

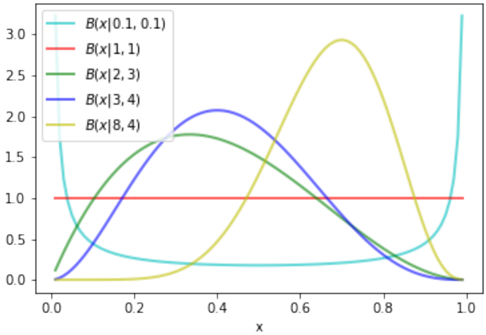
$$p(\gamma | x) \propto \Gamma(a + \frac{1}{2}, b + \frac{(x-\mu)^2}{2}).$$

# Бета распределение

# Бета распределение

$$B(x|a,b) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & , x \in [0,1], \\ 0 & , x \notin [0,1], \end{cases}$$

где $a, b > 0$, $B(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$.



# Бета распределение

$$B(x|a,b) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & , x \in [0,1], \\ 0 & , x \notin [0,1] \end{cases}$$

где $a, b > 0$.

$$\mathrm{M}\beta = \frac{a}{a+b},$$

$$Mode(\beta) = \frac{a-1}{a+b-2},$$

$$\mathrm{D}\beta = \frac{ab}{(a+b)^2(a+b-1)}.$$

# Сопряженное распределение

Для семейства априорных Бета распределений сопряженным является распределение Бернулли.

Функция правдоподобия: $p(X|theta) = \theta^{N_1}(1-\theta)^{N_0}$

$$p(\theta) = B(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$
$$p(\theta|X) \propto p(X|\theta)p(\theta) \Rightarrow$$
$$p(\theta|X) \propto \theta^{N_1}(1-\theta)^{N_0} \cdot \theta^{a-1}(1-\theta)^{b-1} \Rightarrow$$
$$p(\theta|X) \propto \theta^{N_1+a-1}(1-\theta)^{N_0+b-1} \Rightarrow$$
$$p(\theta|X) = B(N_1+a, N_0+b).$$

# Другие сопряженные распределения

Ссылка на таблицу:

- Сопряжённое априорное распределение (Wiki).

# Плюсы и минусы

- Плюсы:
  - Точное выражение для апостериорного распределения.
  - Можно использовать при объединение вероятностных моделей.
- Минусы:
  - Сопряженное априорное распределение может быть неадекватным.