Вероятностный подход к машинному обучению

Лекция 8

Теория вероятностей

- Мир содержит много **неопределенности** $\rightarrow$
  модели, как правило, не могут быть абсолютно точными $\rightarrow$
  теория вероятностей (probability theory)
  - Пример: бросок монетки

Теория вероятностей: основные понятия

- **Дискретная случайная величина** – это случайная величина, множество значений которой конечно или счётно
  - Вероятности исходов в сумме дают единицу
  - Пример: бросок кубика
- **Функция вероятности** – функция, возвращающая вероятность того, что дискретная случайная величина примет определённое значение
  - Для кубика: $p(k) = 1/6\,, k = \{1, 2, 3, 4, 5, 6\}$

$$\sum_{i=0}^{\infty} p(x_i) = 1$$

Теория вероятностей: основные понятия

- **Непрерывная (одномерная) случайная величина** – набор исходов представляет собой вещественную прямую $\mathbb{R}$
- Вероятности отдельных исходов: **функция распределения**

$$F(a) = p(x < a) \text{ – неубывающая}$$

- **Плотность распределения**:

$$f(x) = \frac{dF}{dx}$$

$$\int_{-\infty}^{\infty} f(x)dx = F(\infty) - F(-\infty) = 1$$

Примеры

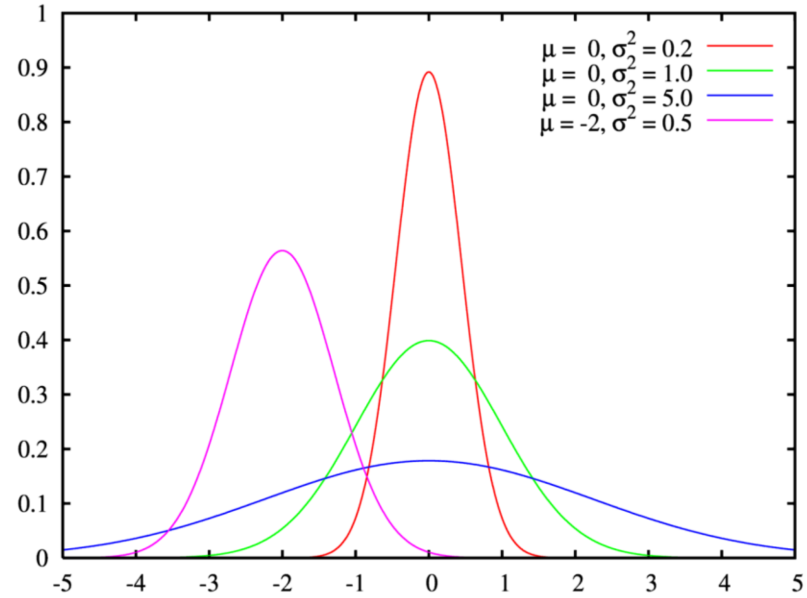
- Нормальное распределение $\mathcal{N}(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

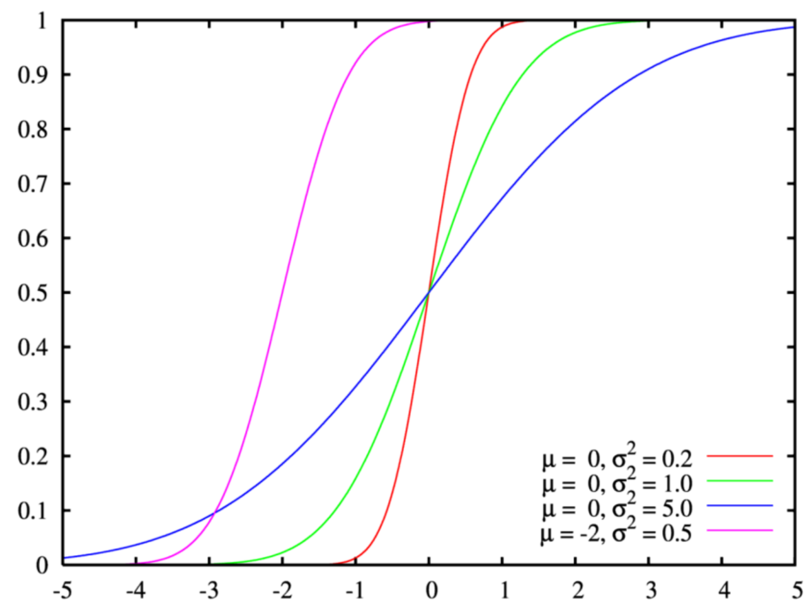
  - μ – математическое ожидание
  - σ – среднеквадратическое (стандартное) отклонение
  - σ^2 – дисперсия

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Плотность распределения



Функция распределения



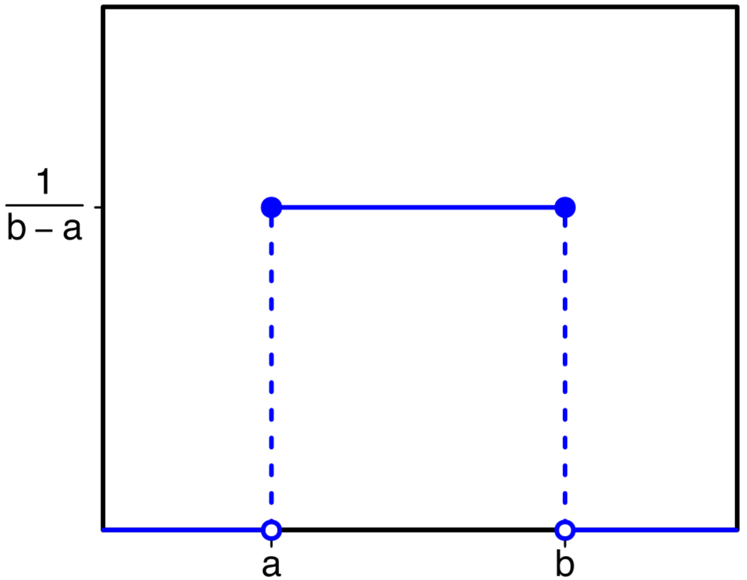
Примеры

- Непрерывное равномерное распределение $\mathcal{U}(a,b)$:

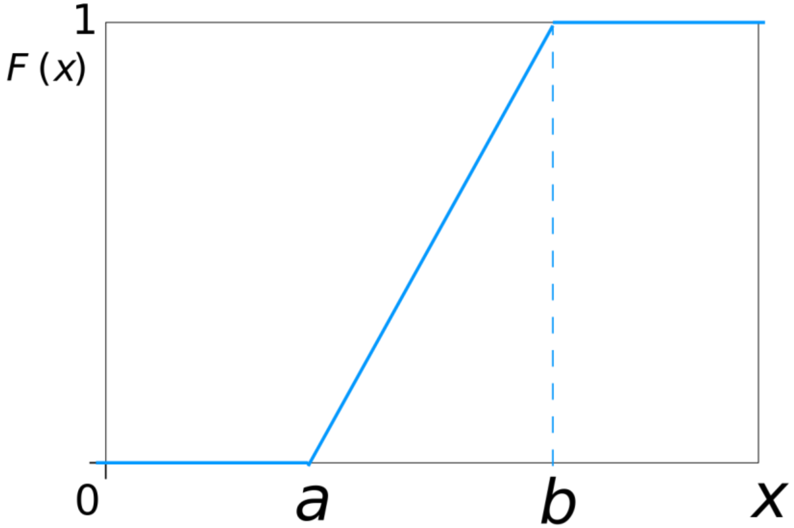
$$f(x) = \begin{cases} \dfrac{1}{b-a}, x \in [a,b] \\ 0, x \notin [a,b] \end{cases}$$

$$F(x) = \begin{cases} 0, x < a \\ \dfrac{x-a}{b-a}, a \le x < b \\ 1, x \ge b \end{cases}$$

Плотность распределения



Функция распределения



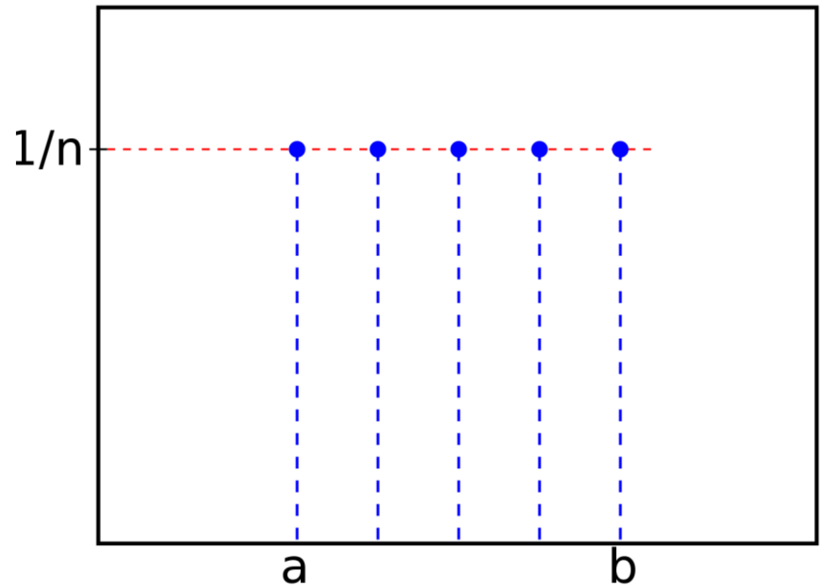
Примеры

- Дискретное равномерное распределение:

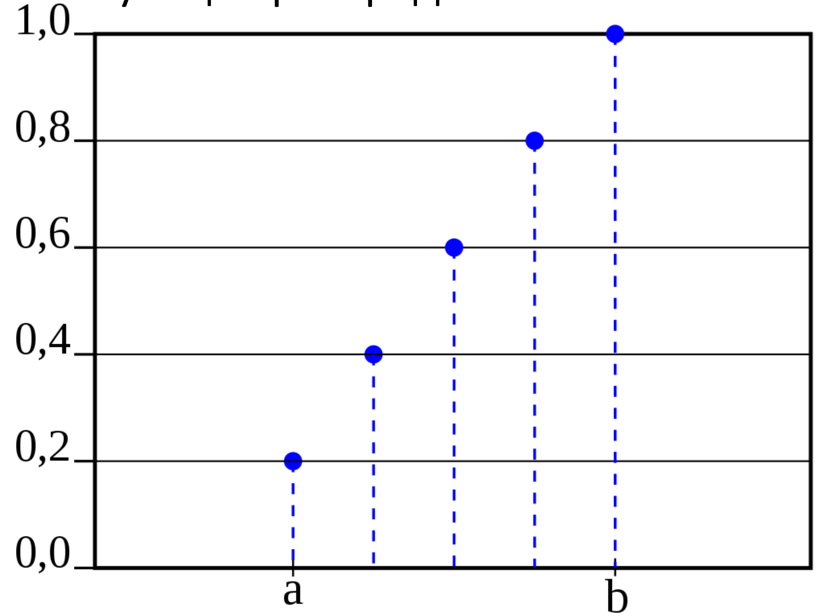
$$p(k) = \begin{cases} \frac{1}{n}, a \leq k \leq b \\ 0, \text{иначе} \end{cases}$$

$$P(k) = \begin{cases} 0, k < a \\ \frac{k - a + 1}{n}, a \leq k \leq b \\ 1, k > b \end{cases}$$

Функция вероятности



Функция распределения



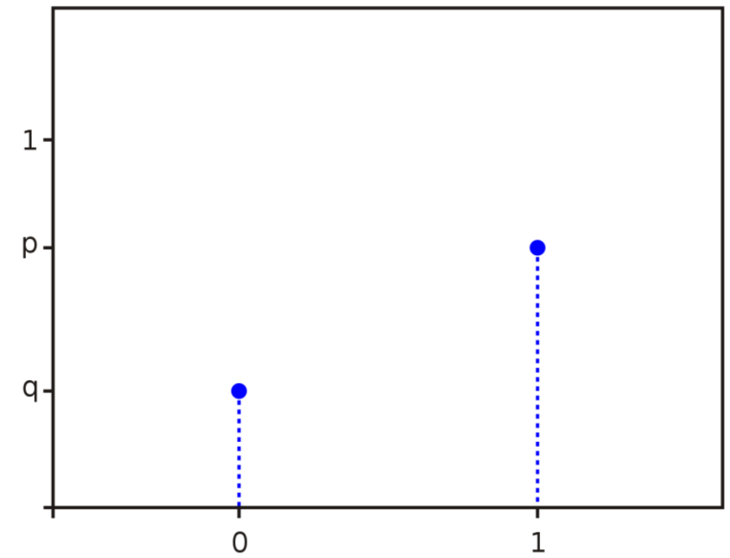
Примеры

- Распределение Бернулли:

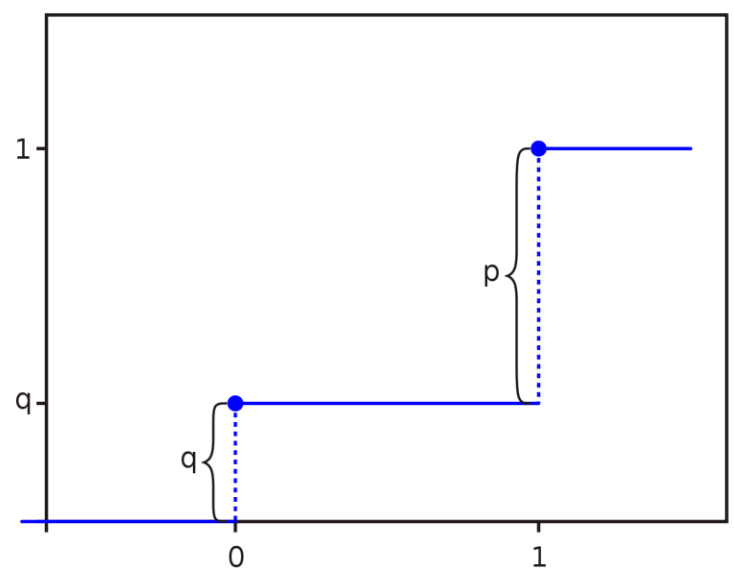
$$p(k) = \begin{cases} q, k = 0 \\ p, k = 1 \end{cases}$$

$$P(k) = \begin{cases} 0, k < 0 \\ q, 0 \leq k < 1 \\ 1, k \geq 1 \end{cases}$$

Функция вероятности



Функция распределения



Примеры

- Биномиальное распределение $\mathcal{B}(n,p)$:

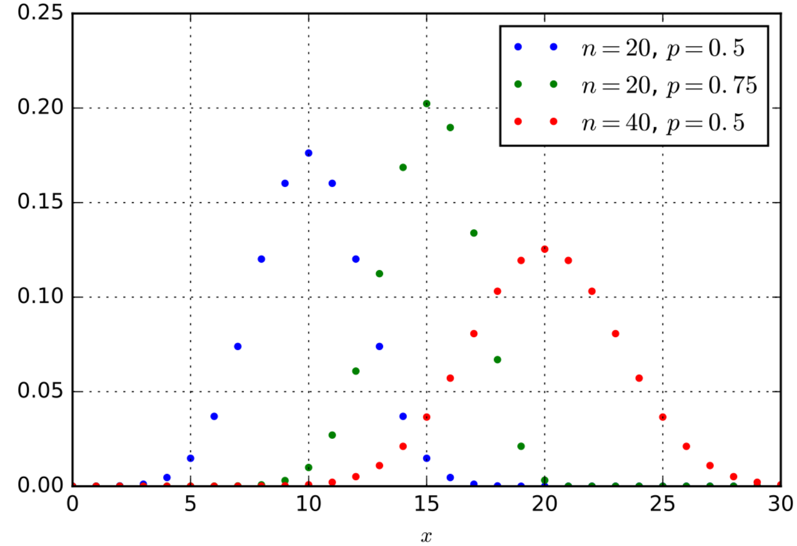
$$f(k) = C_n^k p^k q^{n-k}, k = 0, \ldots, n$$

$$C_n^k = \frac{n!}{(n-k)!\,k!}$$

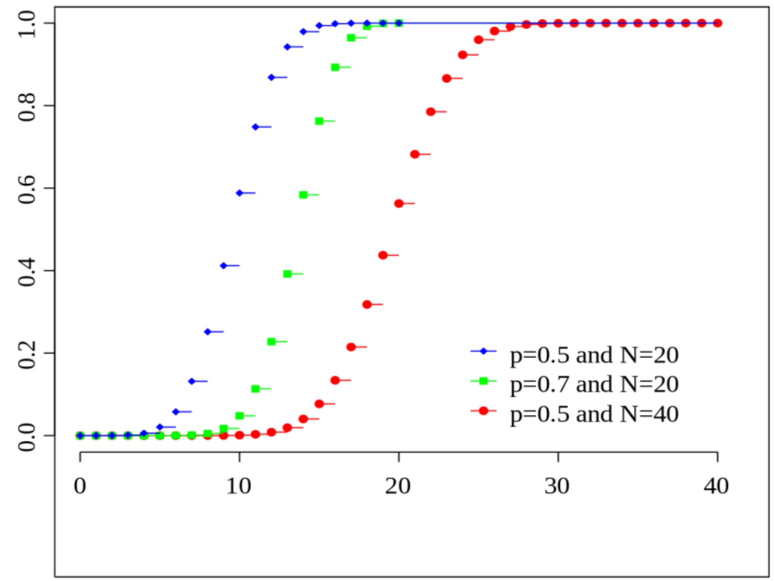
– биномиальный коэффициент

$$P(y) = \sum_{k=0}^{\lfloor y \rfloor} C_n^k p^k q^{n-k}, y \in \mathbb{R}$$

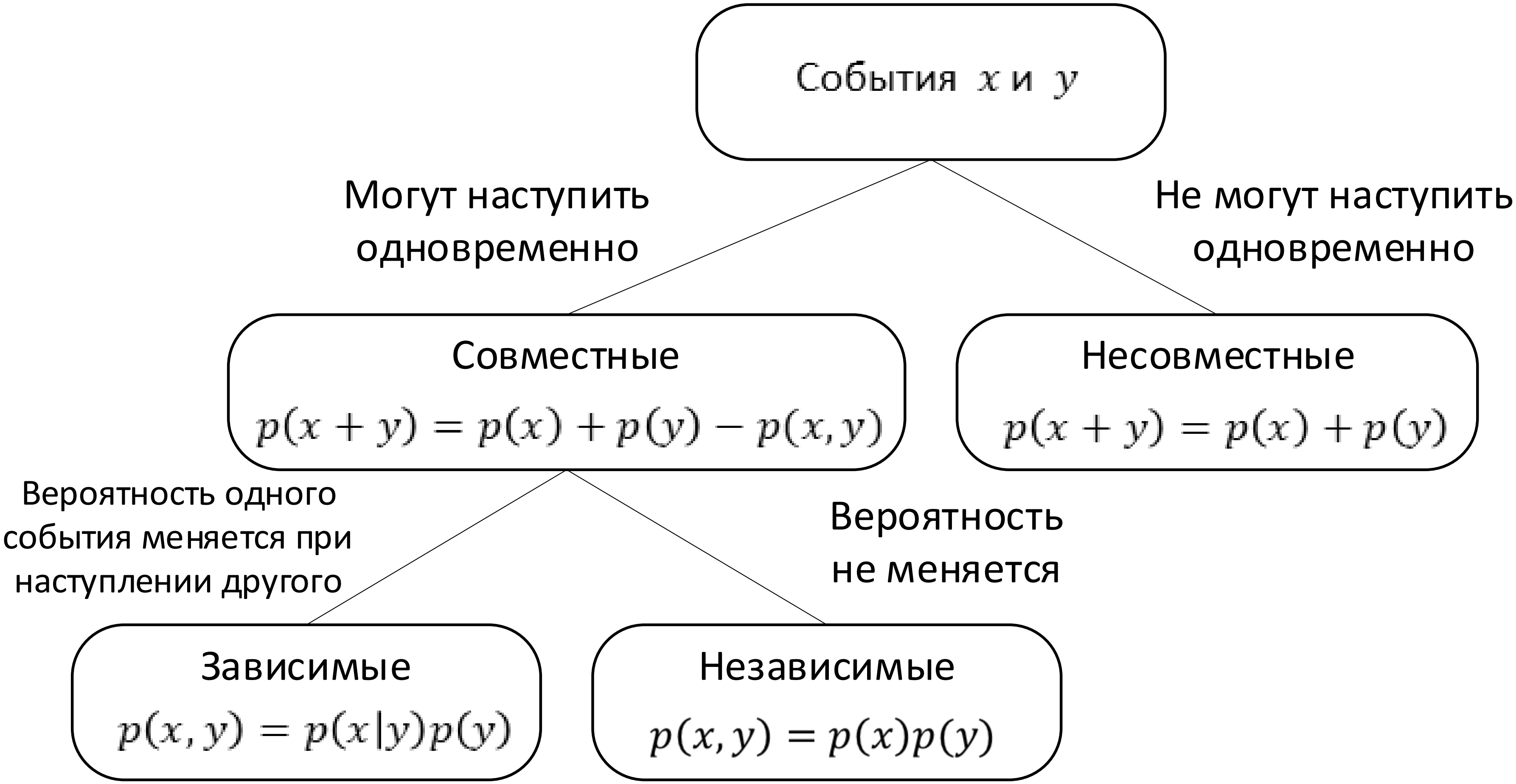
Функция вероятности



Функция распределения



Теория вероятностей: основные понятия



Теория вероятностей: основные понятия

$p(x)$

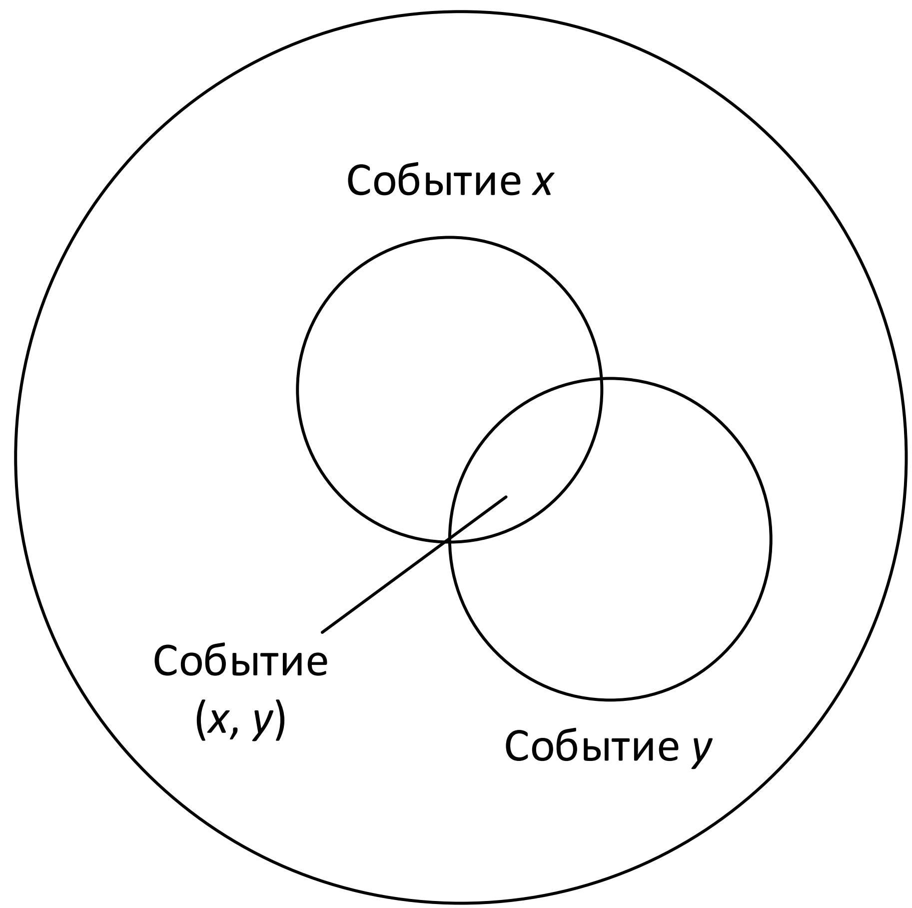
$p(y)$

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

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

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

Пространство элементарных исходов



Теория вероятностей: основные понятия

- **Совместная вероятность** (вероятность произведения событий) – вероятность одновременного наступления двух событий:

$$p(x, y)$$

  - Пример: два кубика – 36 исходов

- Две случайные величины называются **независимыми**, если:

$$p(x, y) = p(x)p(y)$$

  - Зависимость ≠ каузальность (причина-следствие)
  - Зависимость ≠ корреляция (линейная часть зависимости)

Теория вероятностей: основные понятия

- **Условная вероятность** – вероятность наступления одного события, если известно, что произошло другое:

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

- **Условная независимость** относительно третьего события:

$$p(x,y|z) = p(x|z)p(y|z)$$

- **Формула полной вероятности** (*маргинализация*):

$$p(x) = \sum_y p(x,y) = \sum_y p(x|y)\,p(y)$$

Теория вероятностей: основные понятия

- Задача: есть три одинаковые урны. В первой урне находятся 4 белых и 7 черных шаров, во второй – только белые и в третьей – только черные шары. Наудачу выбирается одна урна и из неё наугад извлекается шар. Какова вероятность того, что этот шар чёрный?
- A – из урны извлечен черный шар, B_i – выбрана i-я урна

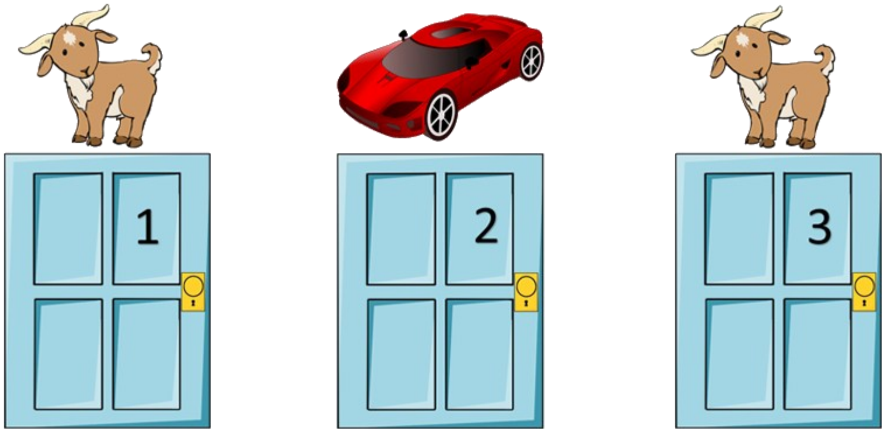
$$p(B_1) = p(B_2) = p(B_3) = 1/3$$

$$p(A|B_1) = \qquad p(A|B_2) = \qquad p(A|B_3) =$$

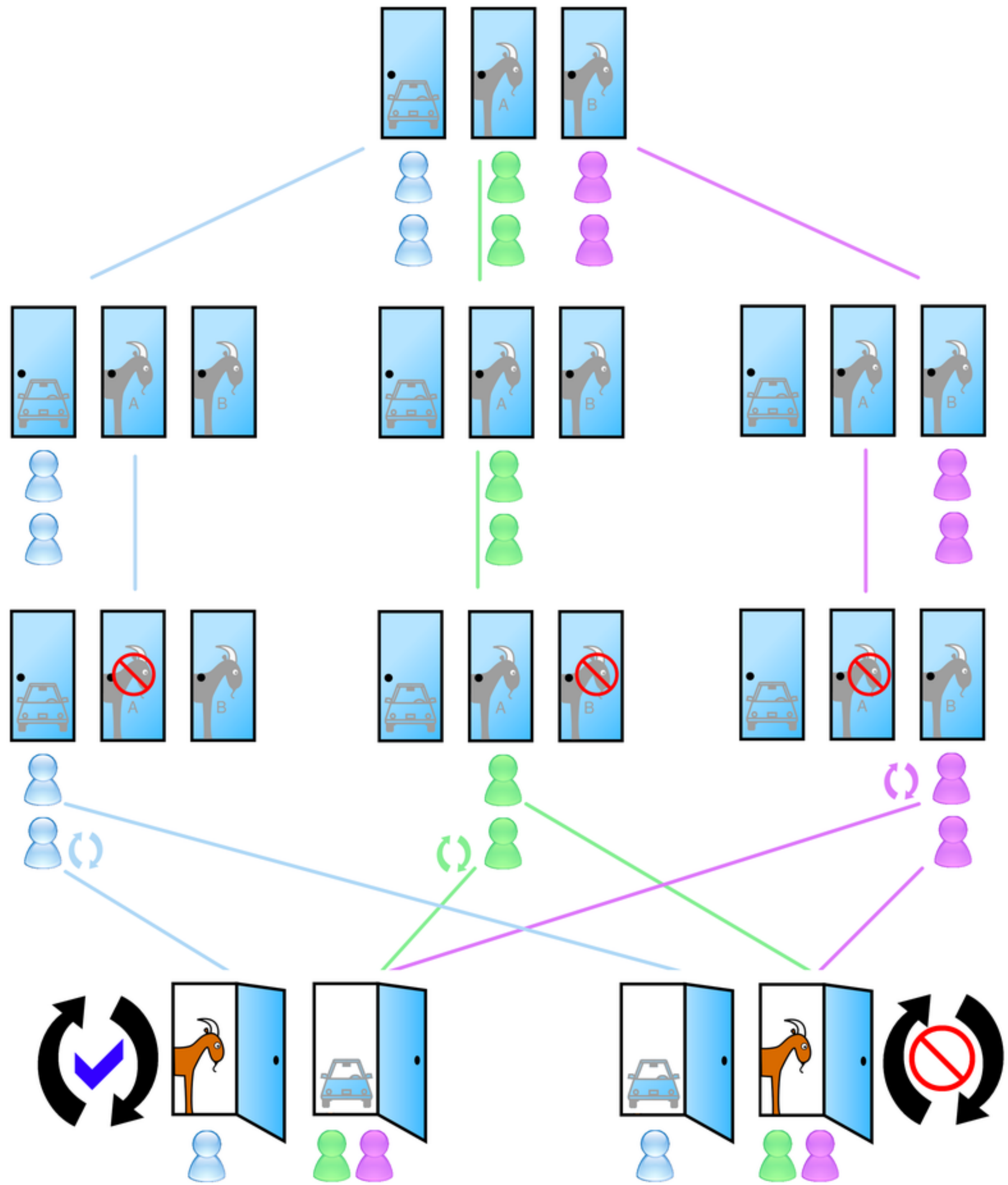
$$p(A) = p(A|B_1)p(B_1) + p(A|B_2)p(B_2) + p(A|B_3)p(B_3)$$

$$p(A) = \frac{7}{11} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{18}{33} = \frac{6}{11}$$

Парадокс Монти Холла



Парадокс Монти Холла



Теорема (формула) Байеса

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

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

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

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_{y'\in Y} p(x|y')p(y')}$$

Thomas Bayes (1702–1761)

[Ballesteros-Perez et al. – Explaining the Bayes' Theorem Graphically (2018)](#)

Теорема (формула) Байеса

Пример:

- тест t на болезнь d: точность 95%
- распространенность болезни d: 1%
- какова вероятность наличия болезни $(d = 1)$
  в случае положительного теста $(t = 1)$?

Теорема (формула) Байеса

$$p(d = 1) =$$

$$p(t = 1|d = 1) =$$

$$p(t = 1|d = 0) =$$

$$p(d = 1|t = 1) =$$

$$= \frac{p(t = 1|d = 1)p(d = 1)}{p(t = 1|d = 1)p(d = 1) + p(t = 1|d = 0)p(d = 0)}$$

$$p(d = 1|t = 1) = \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} = \frac{\overset{p(t = 1, d = 1)}{0.0095}}{\underset{p(t = 1)}{0.059}} = 0.16$$

Парадокс Монти Холла

Предположим, мы выбираем первую дверь.

$p(x_1 = 1)$ – вероятность того, что автомобиль за первой дверью

$p(x_1 = 0)$ – вероятность того, что автомобиля нет за первой дверью

$p(y_2 = 1)$ – вероятность того, что ведущий откроет вторую дверь

$p(x_1 = 1|y_2 = 1)$ – вероятность того, что автомобиль окажется за первой дверью при условии того, что ведущий откроет вторую дверь

$p(x_1 = 0|y_2 = 1)$ – вероятность того, что автомобиля не окажется за первой дверью при условии того, что ведущий откроет вторую дверь

Парадокс Монти Холла

$$p(x_1 = 1|y_2 = 1) =$$

$$= \frac{\overset{1/2}{p(y_2 = 1|x_1 = 1)}\overset{1/3}{p(x_1 = 1)}}{\underset{1/2}{p(y_2 = 1|x_1 = 1)}\underset{1/3}{p(x_1 = 1)} + \underset{1/2}{p(y_2 = 1|x_1 = 0)}\underset{2/3}{p(x_1 = 0)}} =$$

$$= \frac{1/6}{1/6 + 2/6} = \frac{1}{3}$$

$$p(x_1 = 0|y_2 = 1) =$$

$$== \frac{\overset{1/2}{p(y_2 = 1|x_1 = 0)}\overset{2/3}{p(x_1 = 0)}}{\underset{1/2}{p(y_2 = 1|x_1 = 1)}\underset{1/3}{p(x_1 = 1)} + \underset{1/2}{p(y_2 = 1|x_1 = 0)}\underset{2/3}{p(x_1 = 0)}} =$$

$$= \frac{2/6}{1/6 + 2/6} = \frac{2}{3}$$

Теорема Байеса в машинном обучении

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)} = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}\in\Theta} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}$$

- $\boldsymbol{\theta}$ – параметры модели
- D – данные
- $p(D|\boldsymbol{\theta})$ – правдоподобие (likelihood)
- $p(\boldsymbol{\theta})$ – априорная вероятность (prior probability)
- $p(\boldsymbol{\theta}|D)$ – апостериорная вероятность (posterior probability)
- $p(D)$ – вероятность данных (evidence)

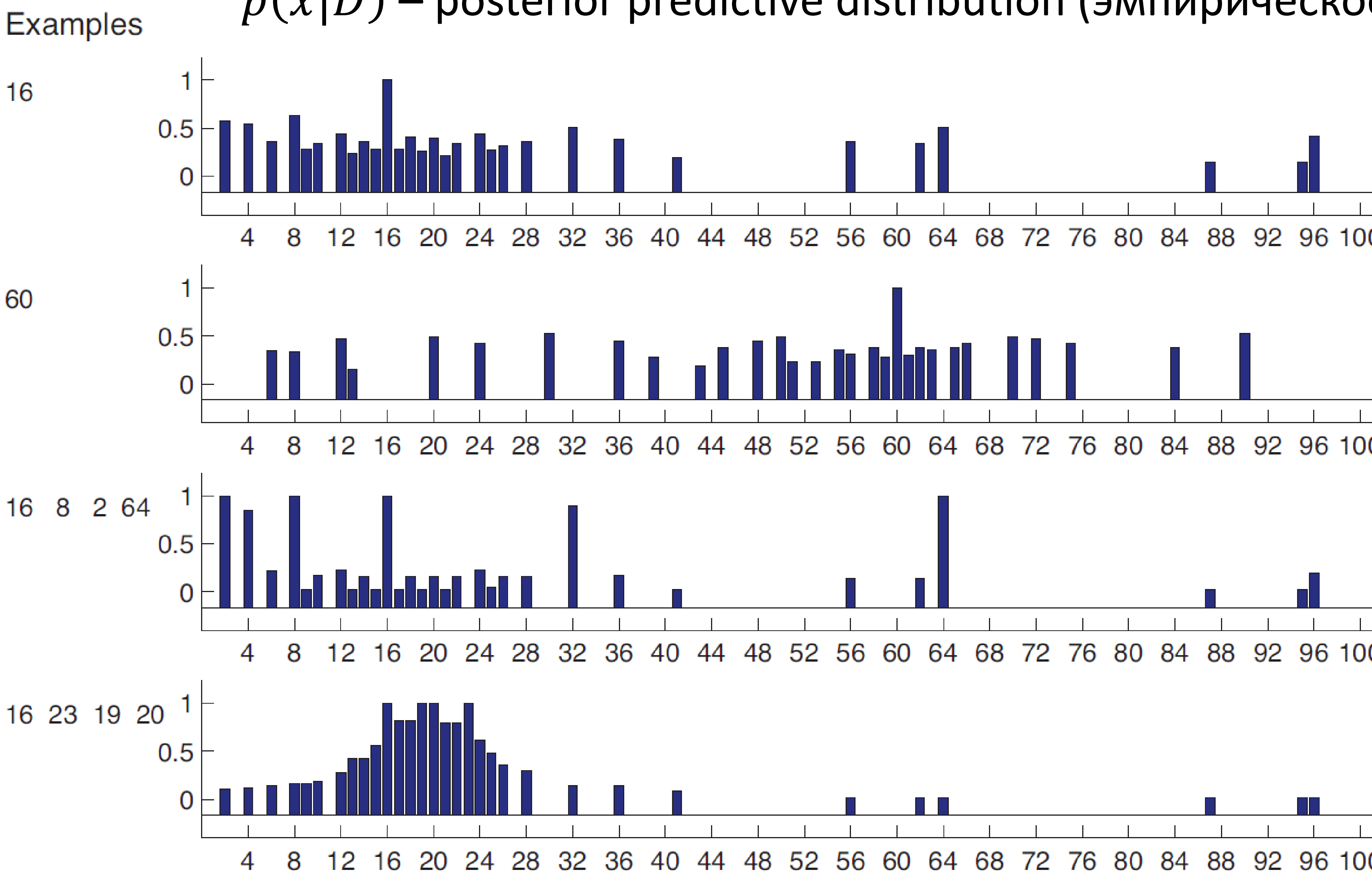
Теорема Байеса: пример

Игра в числа [Murphy, 2012, p. 65]

- Выбрано арифметическое понятие C, например:
  - простые числа
  - четные числа
  - степени двойки
- Из C случайно выбираются N чисел $D = \{x_1, \dots, x_N\}$ и предоставляются игроку
- Его задача – предсказать, принадлежит ли C новый пример $\tilde{x}$

Теорема Байеса: пример

$p(\tilde{x}|D)$ – posterior predictive distribution (эмпирическое)



Теорема Байеса: пример

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)}$$

- **Правдоподобие (likelihood)**:

$$p(D|h) = \prod_{n=1}^{N} p(x_n|h) = \left(\frac{1}{|h|}\right)^N \text{ (бритва Оккама)}$$

- Пусть $D = \{16\}$:

$$p(D|h_{two}) = \frac{1}{6}, \qquad p(D|h_{even}) = \frac{1}{50}$$

- Пусть $D = \{16, 8, 2, 64\}$:

$$p(D|h_{two}) = \left(\frac{1}{6}\right)^4 = 7.7 \cdot 10^{-4}, \qquad p(D|h_{even}) = \left(\frac{1}{50}\right)^4 = 1.6 \cdot 10^{-7}$$

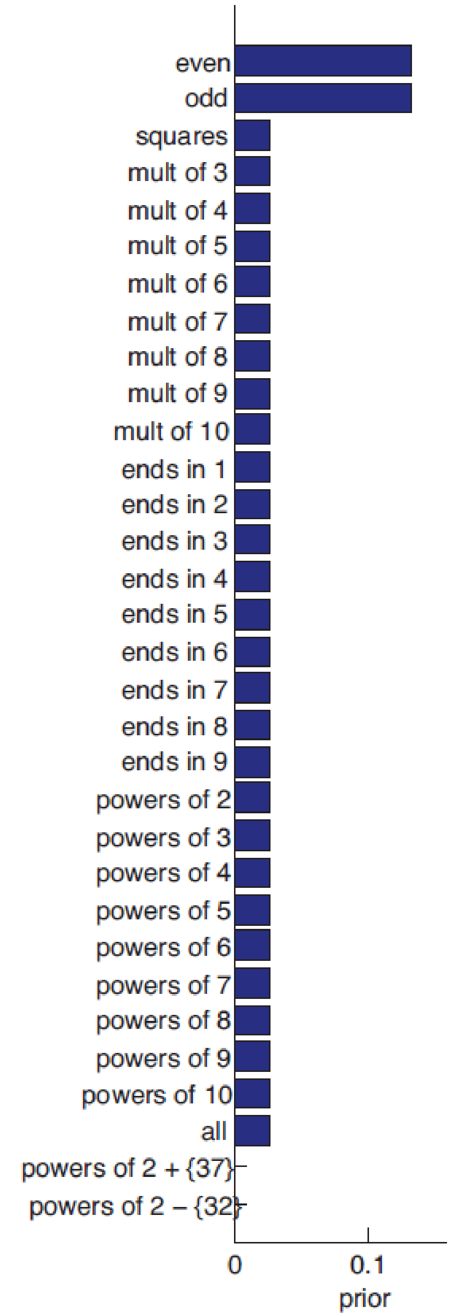
Теорема Байеса: пример

$$\boxed{p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)}}$$

Априорная вероятность (prior probability)

- Для $D = \{16, 8, 2, 64\}$ гипотеза h =«степени двойки» более правдоподобна, чем h' =«степени двойки за исключением 32»
- Субъективно, но позволяет учитывать «фоновое знание» (background knowledge)
  - Без этого обучение на небольшом количестве примеров становится невозможно

$D = \{16\}$



Теорема Байеса: пример

$$\color{red}{p(\boldsymbol{\theta}|D)} = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)}$$

- **Апостериорная вероятность (posterior probability)**:

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)} = \frac{p(D|h)p(h)}{\sum_{h' \in H} p(D, h')}$$

$$p(h|D) = \frac{p(h)\frac{\mathbb{I}(D \in h)}{|h|^N}}{\sum_{h' \in H} p(h')\frac{\mathbb{I}(D \in h')}{|h'|^N}},$$

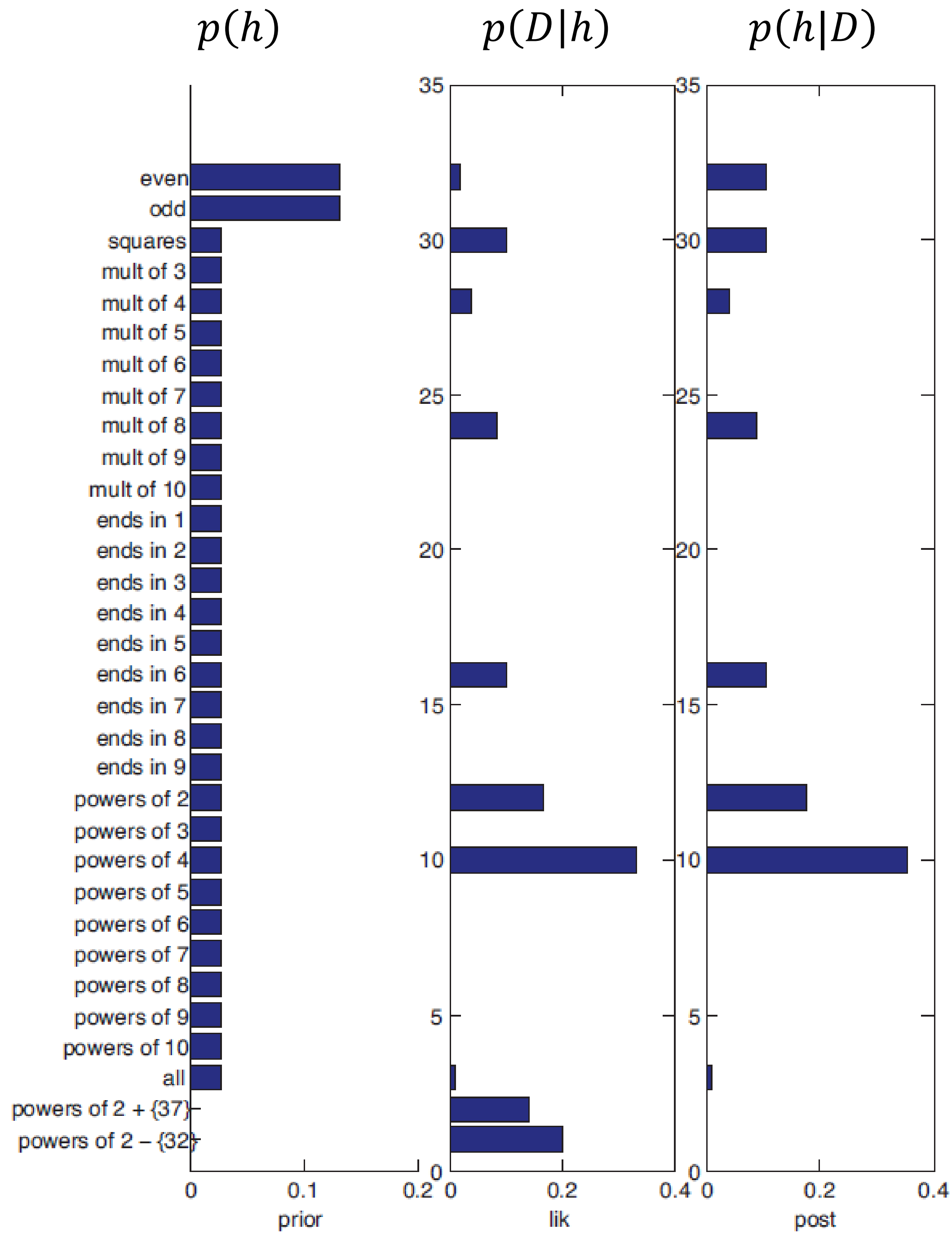
где $\mathbb{I}(D \in h) = 1$ если данные удовлетворяют гипотезе

Теорема Байеса: пример

$$D = \{16\}$$

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)} =$$

$$= \frac{p(h)\frac{\mathbb{I}(D \in h)}{|h|^N}}{\sum_{h' \in H} p(h')\frac{\mathbb{I}(D \in h')}{|h'|^N}}$$



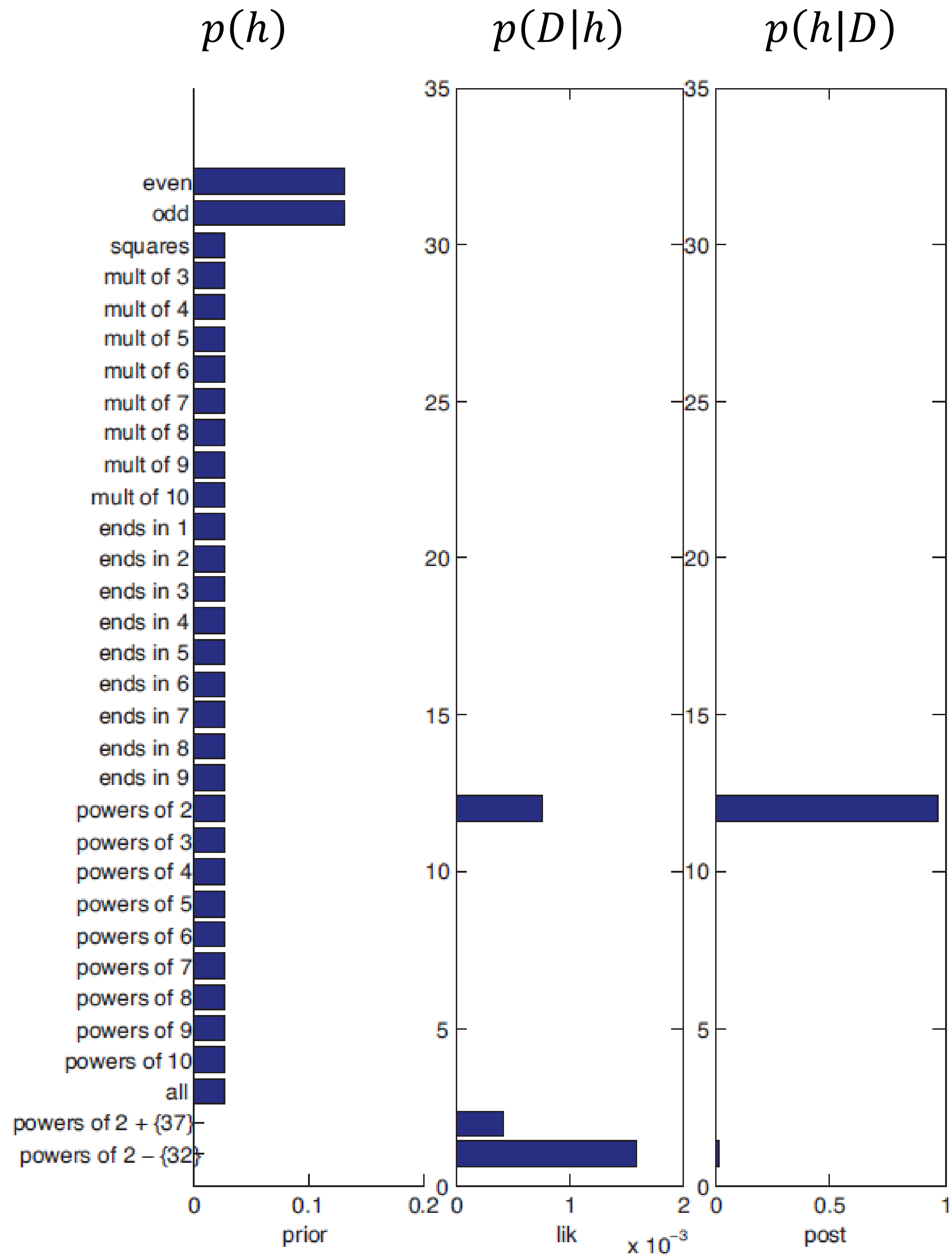
Теорема Байеса: пример

$$D = \{16, 8, 2, 64\}$$

$$p(h|D) = \frac{p(D|h)p(h)}{p(D)} =$$

$$= \frac{p(h)\frac{\mathbb{I}(D \in h)}{|h|^N}}{\sum_{h' \in H} p(h')\frac{\mathbb{I}(D \in h')}{|h'|^N}}$$

Необходимо снижать априорную вероятность неправдоподобных гипотез, иначе «переобучение»!



Теорема Байеса в машинном обучении

$$p(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(D)} = \frac{p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}\in\Theta} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})}$$

- В классической статистике ищут гипотезу максимального правдоподобия (maximum likelihood estimate, MLE):

$$\boldsymbol{\theta}_{MLE} = \arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})$$

- В байесовском подходе ищут апостериорное распределение:

$$p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

и максимальную апостериорную гипотезу (maximum a posteriori hypothesis, MAP):

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|D) = \arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Теорема Байеса в машинном обучении

- При увеличении количества данных MAP-гипотеза сходится к MLE-гипотезе (априорная вероятность становится менее информативной)
- Если истинная гипотеза содержится в пространстве гипотез, то MAP-гипотеза и MLE-гипотеза будут сходиться к ней

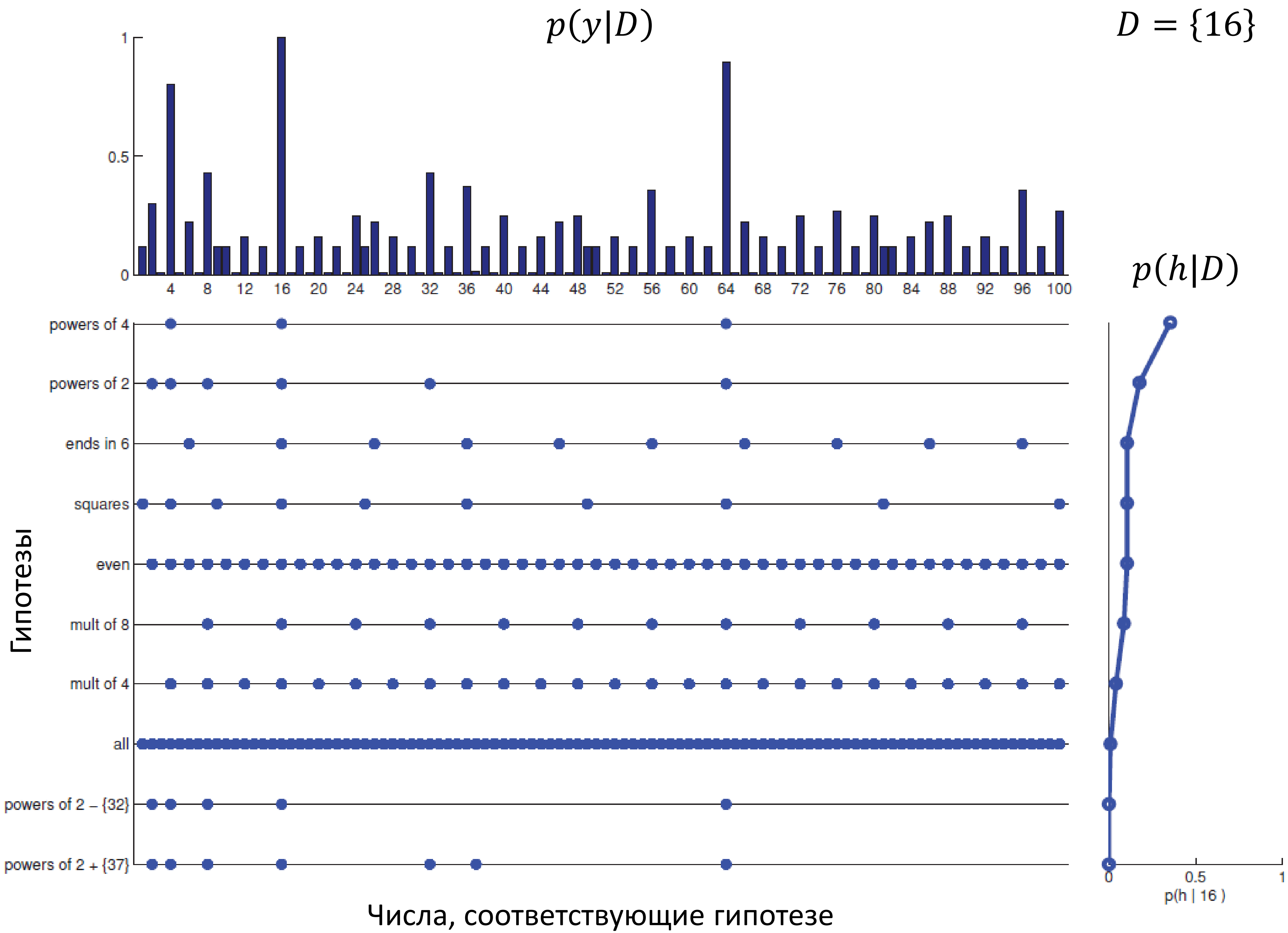
Теорема Байеса в машинном обучении

- **Апостериорное предсказательное распределение** (posterior predictive distribution):

$$p(y|\boldsymbol{x}, D) = \sum_h p(y|\boldsymbol{x}, h)p(h|D)$$

  – взвешенное среднее предсказаний всех гипотез

- Учитываются все гипотезы, а не только наилучшая, чтобы принять во внимание неопределенность относительно гипотез
  - Пример: пусть есть 4 гипотезы с апостериорными вероятностями $\{0.2, 0.2, 0.2, 0.4\}$. MAP-гипотеза – четвертая. Но если новый пример классифицируется положительно тремя первыми гипотезами, а четвертой – отрицательно, то общая вероятность положительной классификации = 0.6


$p(y|D)$
$D = \{16\}$
$p(h|D)$
powers of 4
powers of 2
ends in 6
squares
even
mult of 8
mult of 4
all
powers of 2 – {32}
powers of 2 + {37}
Гипотезы
p(h | 16)
Числа, соответствующие гипотезе

Теорема Байеса в машинном обучении

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|D) = \arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}) =$$

Все примеры независимы

$$= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) \prod_{\boldsymbol{x}\in D} p(\boldsymbol{x}|\boldsymbol{\theta}) =$$

$$= \arg\max_{\boldsymbol{\theta}} \left(\log p(\boldsymbol{\theta}) + \sum_{\boldsymbol{x}\in D} \log p(\boldsymbol{x}|\boldsymbol{\theta}) \right)$$

Априорная вероятность = регуляризация

Правдоподобие модели

Линейная регрессия

- Линейная модель:

$$a(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{i=1}^{d} w_i x_i = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$$

- Целевая (истинная) зависимость:

$$t(\boldsymbol{x}) = a(\boldsymbol{x}, \boldsymbol{w}) + \varepsilon$$

- Пусть шум распределен нормально вдоль линейной модели с центром в нуле:

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- Тогда модель линейной регрессии имеет вид:

$$p(y | \boldsymbol{x}, \boldsymbol{w}, \sigma^2) = \mathcal{N}(y | a(\boldsymbol{x}, \boldsymbol{w}), \sigma^2)$$

Линейная регрессия

- MLE:

$$\boldsymbol{\theta}_{\boldsymbol{MLE}} = \arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \log p(D|\boldsymbol{\theta}) =$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{l} \log p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta})$$

- Negative Log Likelihood (NLL):

$$\boldsymbol{\theta}_{\boldsymbol{MLE}} = \arg\min_{\boldsymbol{\theta}} \left(-\sum_{i=1}^{l} \log p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \right)$$

Линейная регрессия

- Плотность нормального распределения:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- MLE:

$$\boldsymbol{\theta}_{\boldsymbol{MLE}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{l} \log p(y_i | \boldsymbol{x}_i, \boldsymbol{\theta}) =$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{l} \log\left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{y_i - \langle \boldsymbol{w}, \boldsymbol{x}\rangle}{\sigma}\right)^2}\right] =$$

MSE (без $\frac{1}{l}$)

$$= \arg\max_{\boldsymbol{\theta}} \left[-\frac{l}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \boxed{\sum_{i=1}^{l} (y_i - \langle \boldsymbol{w}, \boldsymbol{x}\rangle)^2}\right]$$

Линейная регрессия: регуляризация

- Априорное предположение о весовых коэффициентах:

$$p(\boldsymbol{w}) = \prod_{j=1}^{d} \mathcal{N}\left(w_j | 0, \tau^2\right)$$

- MAP-гипотеза:

$$\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{w}} \sum_{i=1}^{l} \log \mathcal{N}(y_i | w_0 + \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle, \sigma^2) + \sum_{j=1}^{d} \log \mathcal{N}\left(w_j | 0, \tau^2\right)$$

$$\boldsymbol{\theta}_{MAP} = \arg\min_{\boldsymbol{w}} \frac{1}{N} \sum_{i=1}^{l} \left(y_i - (w_0 + \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)\right)^2 + \frac{\sigma^2}{\tau^2} \sum_{j=1}^{d} w_j^2$$

– гребневая (ridge) регрессия (L_2-регуляризация, weight decay)

Линейная регрессия: регуляризация

- Априорное предположение о весовых коэффициентах:

$$p(\boldsymbol{w}) = \prod_{j=1}^{d} Lap(w_j | 0, b),$$

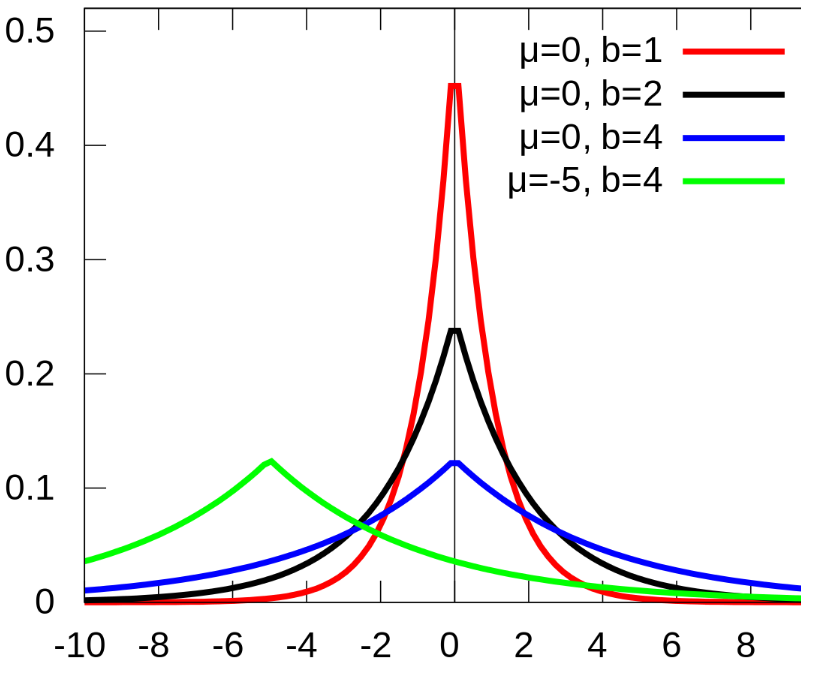
где $Lap(\mu, b) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}}$

– распределение Лапласа

$$\boldsymbol{\theta}_{MAP} = \arg\min_{\boldsymbol{w}} \frac{1}{N} \sum_{i=1}^{l} \left(y_i - (w_0 + \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)\right)^2 +$$

$$+ \frac{2\sigma^2}{b} \sum_{j=1}^{d} |w_j|$$

– L_1-регуляризация (LASSO regression)



Наивный байесовский классификатор

- Условная вероятность класса:

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})}, \qquad \boldsymbol{x} = (x_1, \dots, x_d)$$

- «Наивное» предположение о независимости признаков:

$$p\left(x_j | y, x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d\right) = p(x_j|y)$$

- Наивный байесовский классификатор (Naive Bayes classifier):

$$p(y|\boldsymbol{x}) = \frac{p(y) \prod_{j=1}^{d} p(x_j|y)}{p(x_1, \dots, x_d)}$$

Наивный байесовский классификатор

- Поскольку $p(x_1, \dots, x_d)$ – константа при заданном $\boldsymbol{x}$, то

$$p(y|\boldsymbol{x}) \propto p(y) \prod_{j=1}^{d} p(x_j|y)$$

$$\hat{y} = \arg\max_{y} p(y) \prod_{j=1}^{d} p(x_j|y)$$

- Вероятность класса:

$$p(y = c) = \frac{l_c}{l},$$

где l – общее количество примеров,
l_c – количество примеров класса c

Наивный байесовский классификатор

- Проблема переполнения при перемножении вероятностей:

$$\hat{y} = \arg\max_y p(y) \prod_{j=1}^{d} p(x_j|y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y \left[\log p(y) + \sum_{j=1}^{d} \log p(x_j|y) \right]$$

Наивный байесовский классификатор

- Вероятность $p(\boldsymbol{x}|y)$ зависит от предположений относительно этого распределения – разные варианты наивного байесовского классификатора
- scikit-learn:
  - GaussianNB
  - CategoricalNB
  - MultinomialNB
  - BernoulliNB
  - ComplementNB

Наивный байесовский классификатор

- **GaussianNB** (непрерывные признаки):

$$p(x_i|y=c) = \frac{1}{\sqrt{2\pi\sigma_{ic}^2}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_{ic}}{\sigma_{ic}}\right)^2},$$

где μ_{ic}, σ_{ic}^2 – среднее значение и дисперсия признака x_i для примеров класса c

Наивный байесовский классификатор

- **CategoricalNB** (номинальные/категориальные признаки):
  - можно дискретизировать непрерывные признаки
  - нумерация категорий с нуля

$$p(x_i = t | y = c, \alpha) = \frac{N_{tic} + \alpha}{N_c + \alpha n_i},$$

где N_{tic} – количество примеров, принадлежащих классу c, в которых $x_i = t$,

N_c – количество примеров, принадлежащих классу c,

n_i – количество категорий для признака x_i,

α – сглаживающий параметр (smoothing parameter, $\alpha = 1$)

Наивный байесовский классификатор

- **MultinomialNB** (счетные признаки, например, слова):

$$p(x_i|y = c) = \frac{N_{ic} + \alpha}{N_c + \alpha n},$$

где N_{ic} – количество раз, сколько признак x_i встретился в примерах, принадлежащих классу c,

N_c – суммарное количество раз, сколько все признаки встретились в примерах, принадлежащих классу c,

α – сглаживающий параметр (smoothing parameter, $\alpha = 1$),

n – количество признаков (размер словаря)

Наивный байесовский классификатор

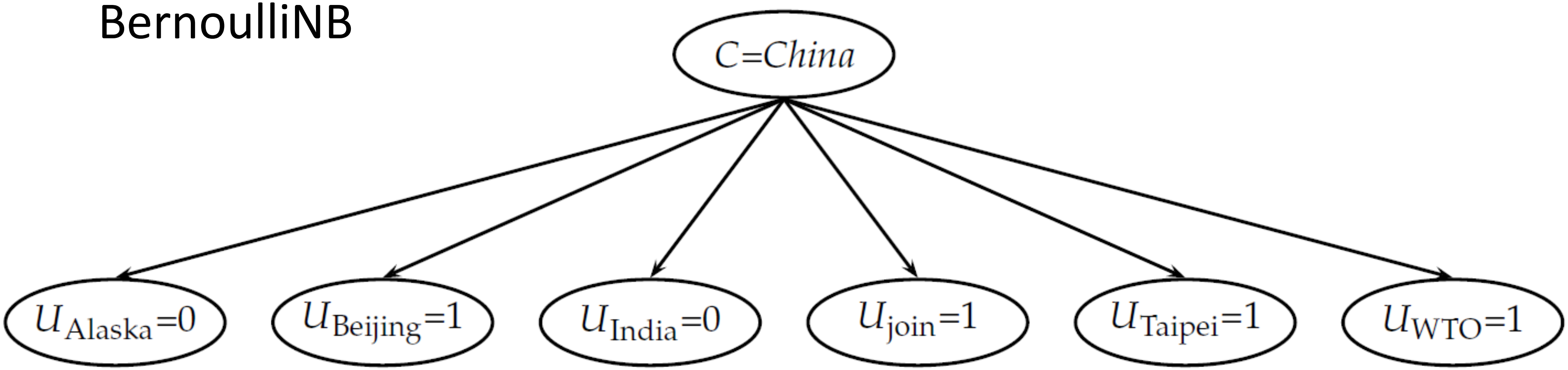
- **BernoulliNB** (бинарные признаки)
  - можно бинаризовать признаки (one-hot encoding)

$$p(x_i|y=c) = p(i|y=c)x_i + \big(1 - p(i|y=c)\big)(1 - x_i)$$

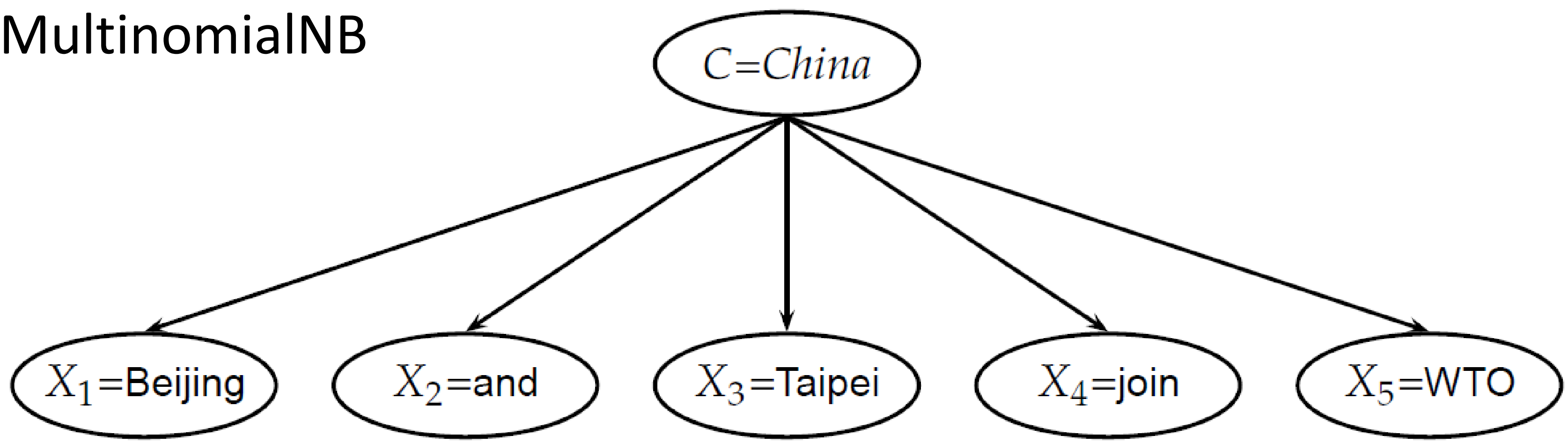
- При вычислении $p(i|y=c)$ используется аддитивное сглаживание

Наивный байесовский классификатор

BernoulliNB



MultinomialNB



Наивный байесовский классификатор

- **ComplementNB** (счетные признаки, например, слова)
  - вариант `MultinomialNB` для несбалансированных датасетов
  - часто превосходит `MultinomialNB` в текстовой классификации

$$p(x_i|y=c) = \frac{N_{i\tilde{c}} + \alpha_i}{N_{\tilde{c}} + \alpha},$$

где $N_{i\tilde{c}}$ – количество раз, сколько признак x_i встретился в примерах, **не** принадлежащих классу c,

$N_{\tilde{c}}$ – суммарное количество раз, сколько все признаки встретились в примерах, **не** принадлежащих классу c,

α_i – сглаживающий параметр (smoothing parameter, $\alpha_i = 1$),

$\alpha = \sum_i \alpha_i$

Наивный байесовский классификатор

- **ComplementNB**
- Поскольку находятся вероятности для классов, отличных от данного:

$$\hat{y} = \arg\max_{y} \left[\log p(y) - \sum_{j=1}^{d} \log p(x_j|y) \right]$$

Минус

Наивный байесовский классификатор

- GaussianNB: непрерывные признаки
- CategoricalNB: категориальные признаки
- MultinomialNB: счетные признаки
  - количество слов, TF-IDF
- BernoulliNB: бинарные признаки
- ComplementNB: вариант MultinomialNB для несбалансированных датасетов

Наивный байесовский классификатор

- Преимущества:
  - Простота понимания и реализации
  - Высокое качество для многих задач
  - Высокая скорость
  - Интерпретируемость
  - Может работать с небольшим количеством примеров
  - Может работать с большими данными (`partial_fit`)
- Недостаток:
  - Предположение о независимости часто не выполняется

Выводы

- Байесовский вывод – метод статистического вывода, в котором новые наблюдения используются для обновления вероятности гипотезы
- Байесовский вывод основан на теореме Байеса и понимании вероятности как степени уверенности в истинности суждения (*байесовская вероятность* vs. *частотная вероятность*)
- В байесовском подходе к машинному обучению ставится задача определения апостериорной вероятности $p(\boldsymbol{\theta}|D)$, которая на практике сводится к нахождению правдоподобия модели $\log p(D|\boldsymbol{\theta})$ и регуляризаторов $\log p(\boldsymbol{\theta})$

Литература

- Kevin P. Murphy. Probabilistic Machine Learning. The MIT Press, 2022
- David Barber. Bayesian Reasoning and Machine Learning. Cambridge University Press. 2012
  - онлайн-версия
- Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer, 2006
- Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. СПб.: Питер, 2018