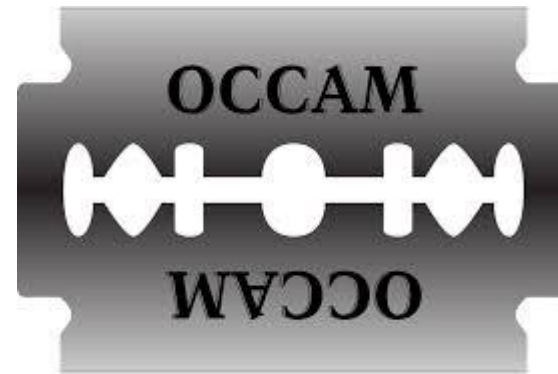


Математика в машинном обучении: краткий обзор

Александр Дьяконов

21 сентября 2020 года

Бритва Оккама



«Entia non sunt multiplicanda sine necessitate»

(лат. сущности не следует умножать без необходимости)

Из всех гипотез, объясняющих данные, надо выбирать простейшую...

1, 2, 3, ? ...

«Объяснение должно быть наипростейшим, но не проще...» Эйнштейн

Теорема о бесплатном сыре (No Free Lunch Theorem)

**В среднем (по всем возможным порождающим распределениям)
у всех алгоритмов процент ошибок одинаков...**

Сложность

Простота алгоритма – MDL, порядок полинома, ...

Простота модели – VC-размерность, ...

Футбольный оракул

исход матча	предсказания при обзвоне
0	0000000011111111
1	00001111-----
0	-----0011-----
0	-----01-----

Будете ли Вы верить предсказаниям?

Если начать с 1/16 финала и распараллелить $\times 10$ (т.е. обзвонить 160 человек),
то перед финалом 10 человек, которым безошибочно сказали 4 исхода!

Сведения из ТВиМС

Вероятность события \sim доля испытаний, завершившихся наступлением события, при бесконечном числе экспериментов.

Есть и другой подход к определению вероятности!

ЗБЧ: частота \rightarrow вероятность



\rightarrow
теория вероятностей

\leftarrow
**математическая
статистика**



Сведения из ТВиМС

Как задать распределение с.в. ξ

Если принимает значения x_1, x_2, \dots , то вероятностями

$$p_1 = \mathbf{P}(\xi = x_1), p_2 = \mathbf{P}(\xi = x_2), \dots$$
$$\sum_i p_i = 1, p_i \geq 0$$



Сведения из ТВиМС

Если $\xi \in \mathbb{R}$, то функцией распределения

$$p(x) : F_{\xi}(x) = \int_{-\infty}^x p(z) \partial z$$

удобна тем, что

$$\mathbf{P}(a \leq \xi \leq b) = \int_a^b p(x) \partial x$$



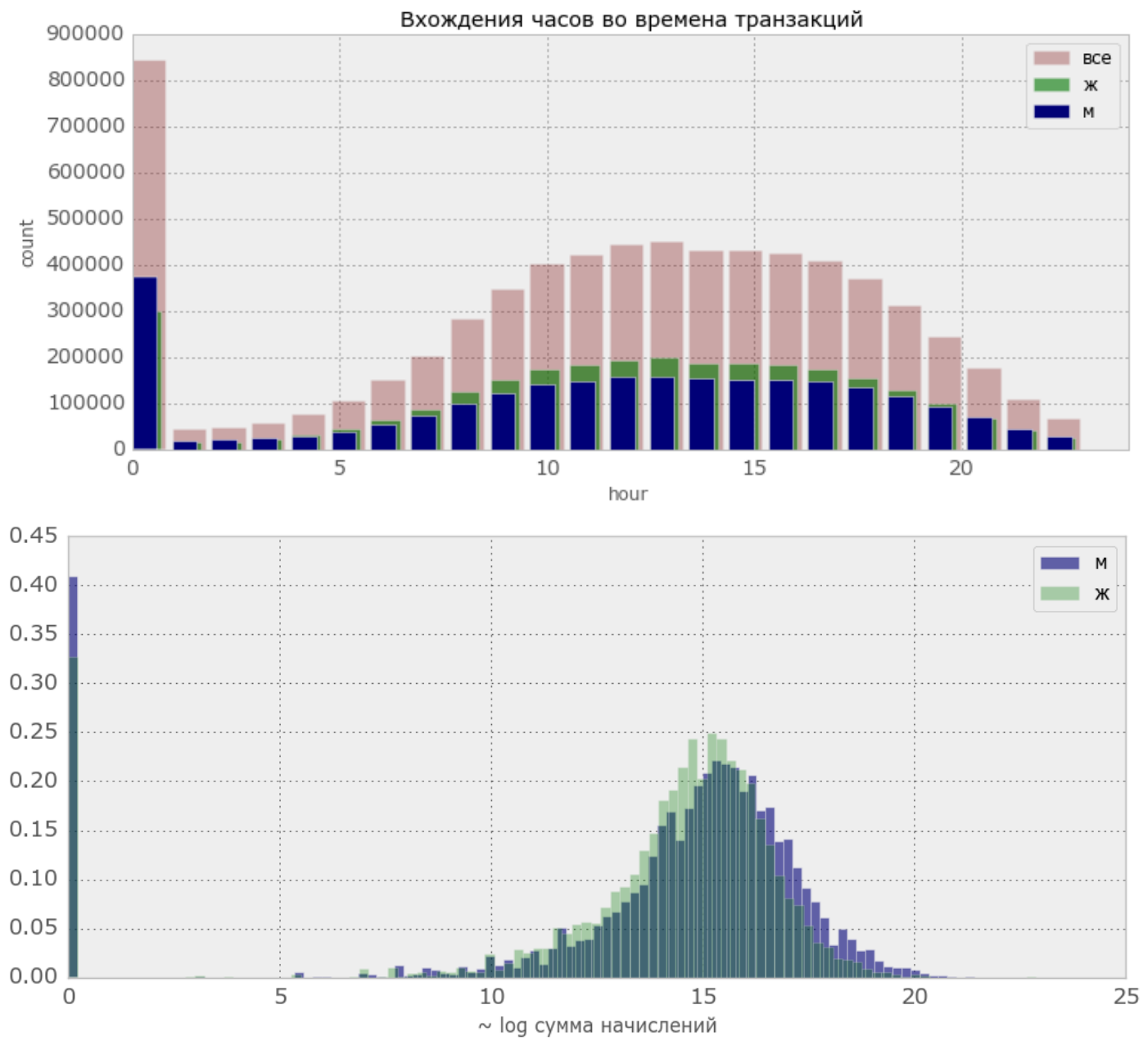
Связь плотности и вероятности

$$\mathbf{P}(x - \varepsilon \leq \xi \leq x + \varepsilon) = \int_{x-\varepsilon}^{x+\varepsilon} p(z) dz \approx 2\varepsilon p(x)$$

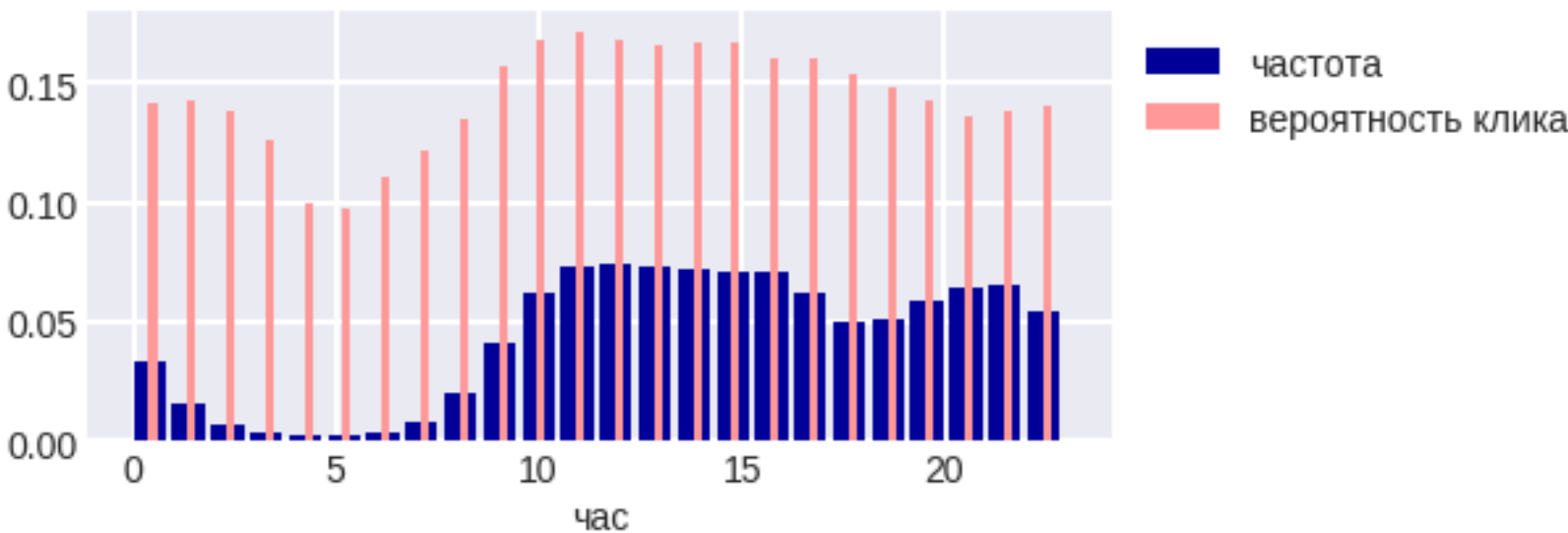
$$\frac{\mathbf{P}(\xi \in [x_1 - \varepsilon, x_1 + \varepsilon])}{\mathbf{P}(\xi \in [x_2 - \varepsilon, x_2 + \varepsilon])} = \frac{p(x_1)}{p(x_2)}$$



Примеры распределений из жизни: сбербанк



Примеры распределений из жизни: тикетлэнд



Сведения из ТВиМС

Пусть с.в. имеет плотность $p(x)$

Математическое ожидание (~центр масс) –

$$EX = \int xp(x)dx$$

Дисперсия (средний квадрат отклонения от МО) –

$$E(X - EX)^2 = \int (x - EX)^2 p(x)dx$$

можно рассматривать и другие средние и отклонения
квантиль, медиана, мода

Сведения из ТВиМС

Условная плотность –

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

Очевидный пересчёт

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

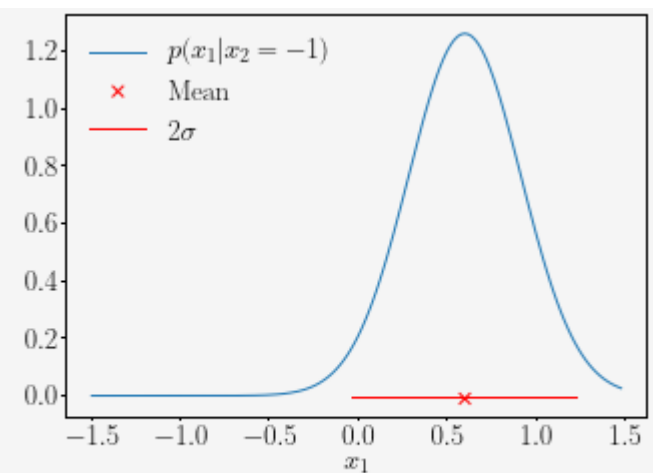
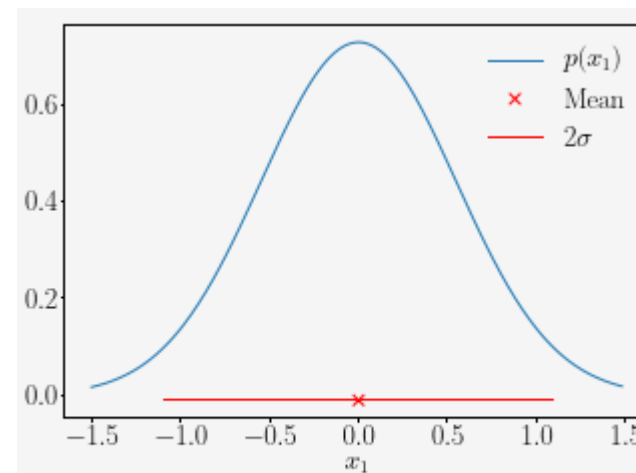
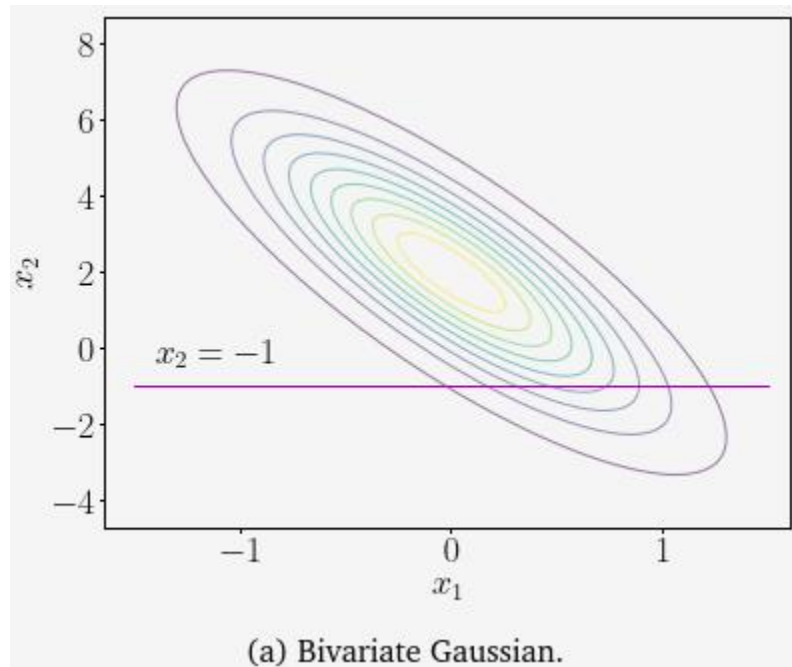
Сведения из ТВиМС

Маргинализация плотности по неизвестной компоненте

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

Обуславливание плотности по известной компоненте

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



Правило произведения

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

Точечное оценивание

Зачем нужно?

**Наша же цель найти (оценить?)
истинные значения параметров модели...**

Выборка $\{x_1, \dots, x_m\}$

(независимые одинаково распределённые случайные величины)

Статистика (точечная оценка) – (измеримая) функция от выборки

$$\hat{\theta} = g(x_1, \dots, x_m)$$

Это тоже случайная величина!

примеры

Требования к статистике

1) Надо, чтобы значение было близко к истинному значению параметров модели θ

Смещение $\text{bias}(\hat{\theta}) = \mathbb{E}\hat{\theta} - \theta$

Несмещённая (unbiased) оценка $\text{bias}(\hat{\theta}) = 0$

Асимптотически несмещённая оценка $\text{bias}(\hat{\theta}) \rightarrow 0$

Для нормального распределения несмещённые оценки:

$$\hat{\mu} = \frac{x_1 + \dots + x_m}{m}$$
$$\hat{\sigma}^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \hat{\mu})^2$$

Требования к статистике

**2) Надо, чтобы оценка не сильно варьировалась
в зависимости от выборки**

$$\text{var}(\hat{\theta}) \rightarrow \min$$

Пример:

$$\text{var}(\hat{\mu}) = \frac{\sigma^2}{m}$$

Требования к статистике

3) с ростом числа наблюдений была сходимость

Состоятельность (Consistency): $\hat{\theta} \xrightarrow{P} \theta$

$$\forall \varepsilon > 0 \mathbf{P}(|\hat{\theta} - \theta| > \varepsilon) \rightarrow 0 \text{ при } m \rightarrow \infty$$

Пример: оценка $\hat{\mu} = x_1$ несмещённая, но не является состоятельной

Оценка Maximum Likelihood Estimation (MLE / ММП)

$$\{y_1, \dots, y_m\}$$

независимые, одинаково распределённые

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} p(D | \theta) = \arg \max_{\theta} \prod_i p(y_i | \theta)$$

м.б. смещённая

1) состоятельная

2) асимптотически эффективная и асимптотически нормальная

эффективность вводится в классе оценок:

$$\mathbf{E} |\hat{\theta} - \theta|^2 \leq \mathbf{E} |\hat{\theta}' - \theta|^2$$

эффективная – несмещённая оценка, имеющая наименьшую дисперсию из всех возможных несмещённых оценок данного параметра

ММП – среди асимптотически нормальных

есть ещё, например, метод моментов

MAP (будет)

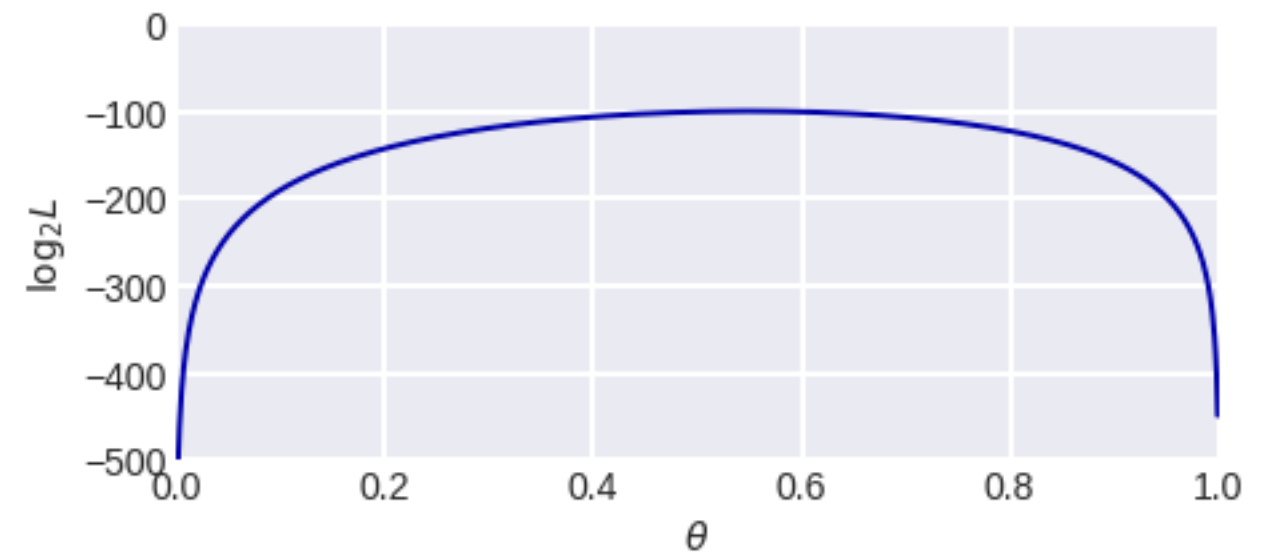
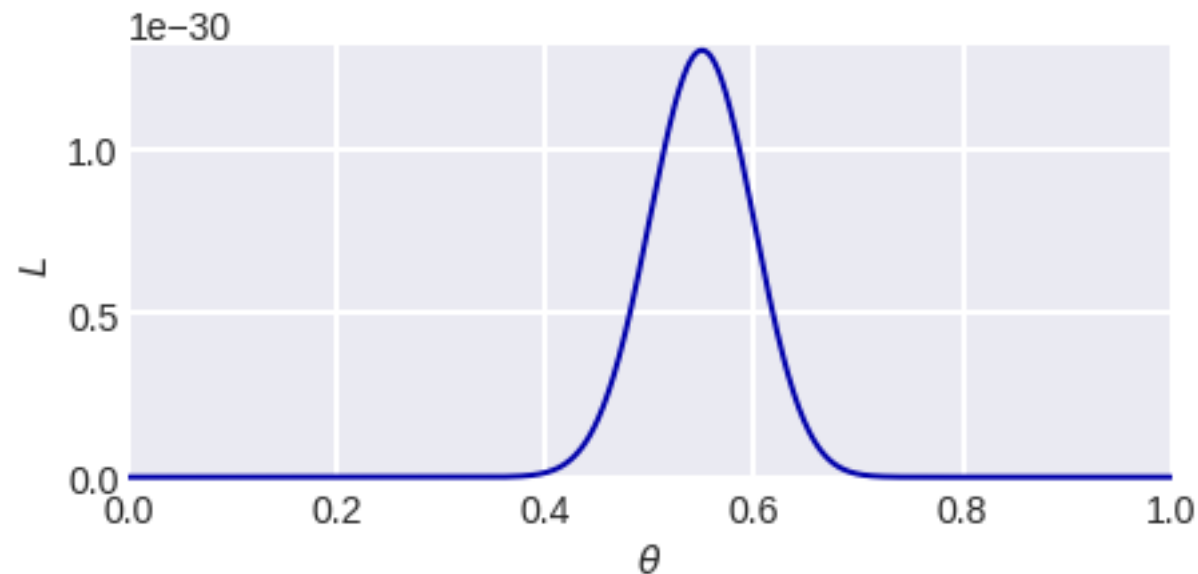
Правдоподобие (Likelihood): графики

Подбрасывание монеты
число бросков – $n = 100$
выпадение орла – $m = 55$

$$L(\theta) = \theta^m (1 - \theta)^{n-m}$$

$$L(0.5) \approx 7.9 \cdot 10^{-31}$$

$$\log L(0.5) = -100$$



часто берут логарифм

Откуда берётся дивергенция Кульбака-Лейблера

Пусть есть выборка $\{x_1, \dots, x_m\}$ из распределения с плотностью p

Мы пытаемся найти распределение $q(x | \theta)$ с параметрами θ ММП:

$$\begin{aligned}\hat{\theta} &= \arg \max_{\theta} \prod_{i=1}^m q(x_i | \theta) = \arg \max_{\theta} \sum_{i=1}^m \log q(x_i | \theta) = \\ &= \arg \max_{\theta} \frac{1}{m} \sum_{i=1}^m \log q(x_i | \theta) - \frac{1}{m} \sum_{i=1}^m \log p(x_i) \\ &\sim \arg \min_{\theta} \frac{1}{m} \sum_{i=1}^m \log \frac{p(x_i)}{q(x_i | \theta)} \\ &\sim \arg \min_{\theta} \underbrace{\int \log \frac{p(x)}{q(x | \theta)} p(x) \partial x}_{\text{KL}(p \| q_{\theta})}\end{aligned}$$

Дивергенция Кульбака-Лейблера

$$\begin{aligned}\text{KL}(p \parallel q_\theta) &= \mathbf{E}_p[\log p(x) - \log q(x \mid \theta)] = \\ &= \int \log \frac{p(x)}{q(x \mid \theta)} p(x) \partial x\end{aligned}$$

**кстати, если хотим минимизировать,
то достаточно (истинное распределение не знаем)**

$$\mathbf{E}_p[\log q(x \mid \theta)] \rightarrow \max$$

а это и есть метод максимального правдоподобия!

Попытка совместить распределение-оценку с истинным...

Взаимная информация

$$I(x, y) = \text{KL}(p(x, y) \parallel p(x)p(y)) =$$

$$= \iint p(x, y) \ln \frac{p(x, y)}{p(x)p(y)} \partial x \partial y$$

$$I(x, y) = H(x) - H(x \mid y) = H(y) - H(y \mid x)$$

Ковариация и корреляция

$$\text{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}X)(Y - \mathbf{E}Y)]$$

$$\text{var}(X) = \text{cov}(X, X) = \mathbf{E}[(X - \mathbf{E}X)^2]$$

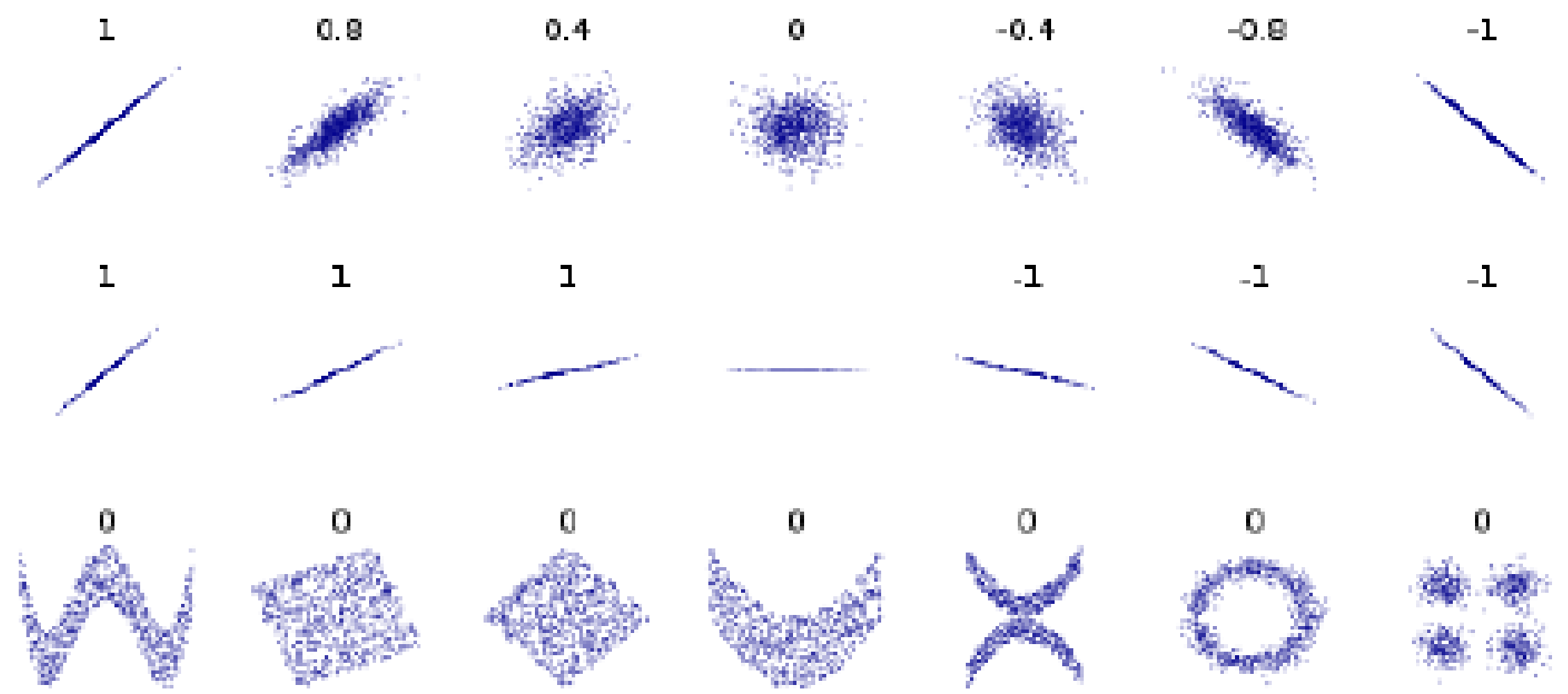
Корреляционный коэффициент Пирсона (Pearson):

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X) \text{var}(Y)}} \in [-1, +1]$$

определяет меру **линейной зависимости** между с.в.

**Независимые переменные некоррелированы.
Обратное неверно.**

Ковариация и корреляция



На корреляционный коэффициент влияют выбросы!

Ковариация и корреляция

Коэффициент корреляции Спирмена (Spearman) – определяет меру монотонной зависимости

= коэффициент корреляции Пирсона между рангами

$$r(\{x_i\}, \{y_i\}) = \frac{\sum_{i=1}^m \left(\text{rank}(x_i) - \frac{m+1}{2} \right) \left(\text{rank}(y_i) - \frac{m+1}{2} \right)}{\frac{1}{12}(m^3 - m)} =$$
$$= 1 - \frac{6}{m^3 - m} \sum_{i=1}^m (\text{rank}(x_i) - \text{rank}(y_i))^2$$

последняя формула – если нет совпадающих рангов

Зависимость бинарных величин

	$a = 0$	$a = 1$
$y = 0$	m_{00}	m_{01}
$y = 1$	m_{10}	m_{11}

φ -коэффициент

$$\varphi = \frac{m_{11}m_{00} - m_{10}m_{01}}{\sqrt{m_{1*}m_{0*}m_{*1}m_{*0}}}$$

Зависимость бинарной и вещественной Point-biserial correlation coefficient

$$r_{pb} = \frac{\text{mean}(x \mid y = 1) - \text{mean}(x \mid y = 0)}{\text{std}(x)} \frac{\sqrt{m_1 m_0}}{m}$$

Оценка плотности

1. Непараметрические методы

нет априорной гипотезы о распределении

2. Параметрические методы

распределение известно с точностью до параметров

$p(y | \theta)$ – ММП – см. выше

3. Смеси распределений

$$p(x) = \sum_{t=1}^k \pi_t p_t(x | \theta_t)$$

$$\sum_{t=1}^k \pi_t = 1, \pi_t \geq 0$$

ЕМ-алгоритм – будет дальше

Оценка плотности: непараметрические методы

Гистограммный подход



Парзеновский подход



Оценка плотности



```
plt.figure(figsize=(7, 3))
plt.hist(x, color='#000099', bins=30, width=0.17, normed=True)
plt.hist(x, color='#990000', bins=10, width=0.6, normed=True, alpha=0.4)
plt.grid(lw=2)
plt.xlabel('значение случайной величины')
plt.ylabel('значение плотности')
plt.xlim([-4, 4])
```

какие недостатки?

Парзеновский подход

$$\frac{1}{mh} \sum_{i=1}^m K\left(\frac{x - x_i}{h}\right)$$

функция ядра / окна:

$$K(z) \geq 0$$

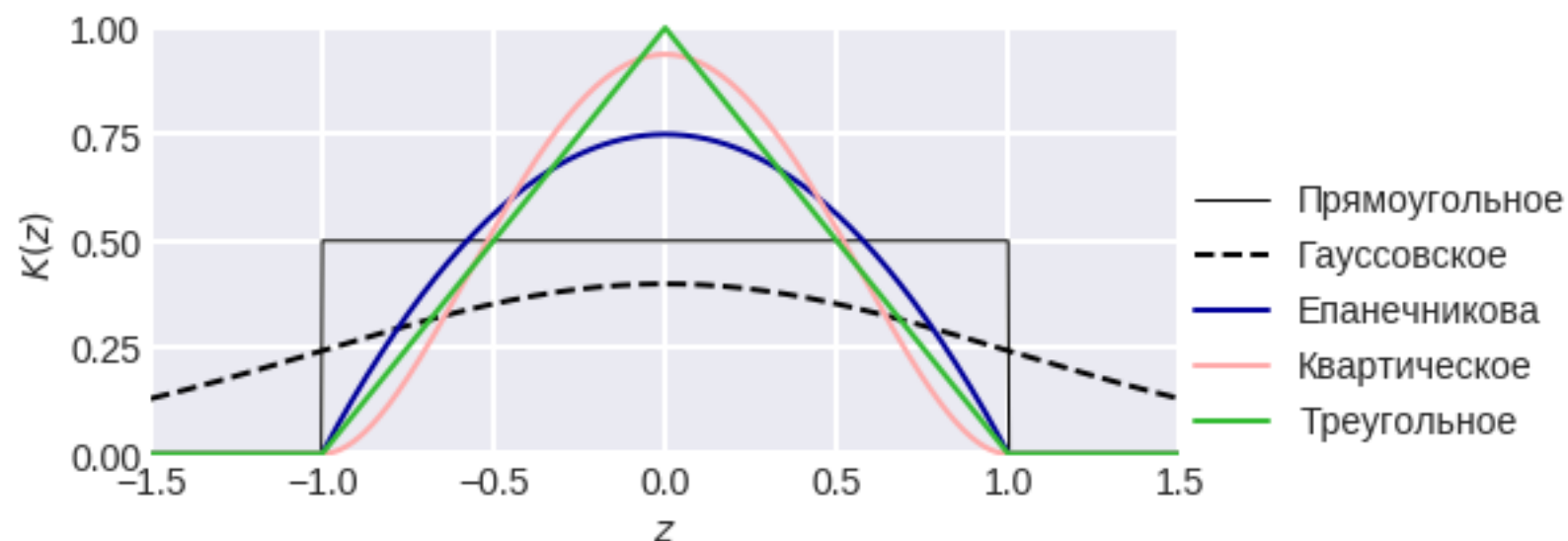
$$\int_{-\infty}^{+\infty} K(z) dz = 1$$



плотность = среднее локальных плотностей

Есть теорема о сходимости

Различные виды ядер (одномерных)

**Треугольное / linear**

$$K(z) = \max(\min(1 - z, 1 + z), 0)$$

Квартическое

$$K(z) = \frac{15}{16} (1 - z^2)^2 I[|z| \leq 1]$$

Прямоугольное / tophat

$$K(z) = \frac{1}{2} I[|z| \leq 1]$$

Гауссовское / gaussian

$$K(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^T z}{2}\right)$$

Епанечникова / епанечников

$$K(z) = \frac{3}{4} (1 - z^2) I[|z| \leq 1]$$

https://scikit-learn.org/stable/auto_examples/neighbors/plot_kde_1d.html

Различные виды ядер (одномерных)

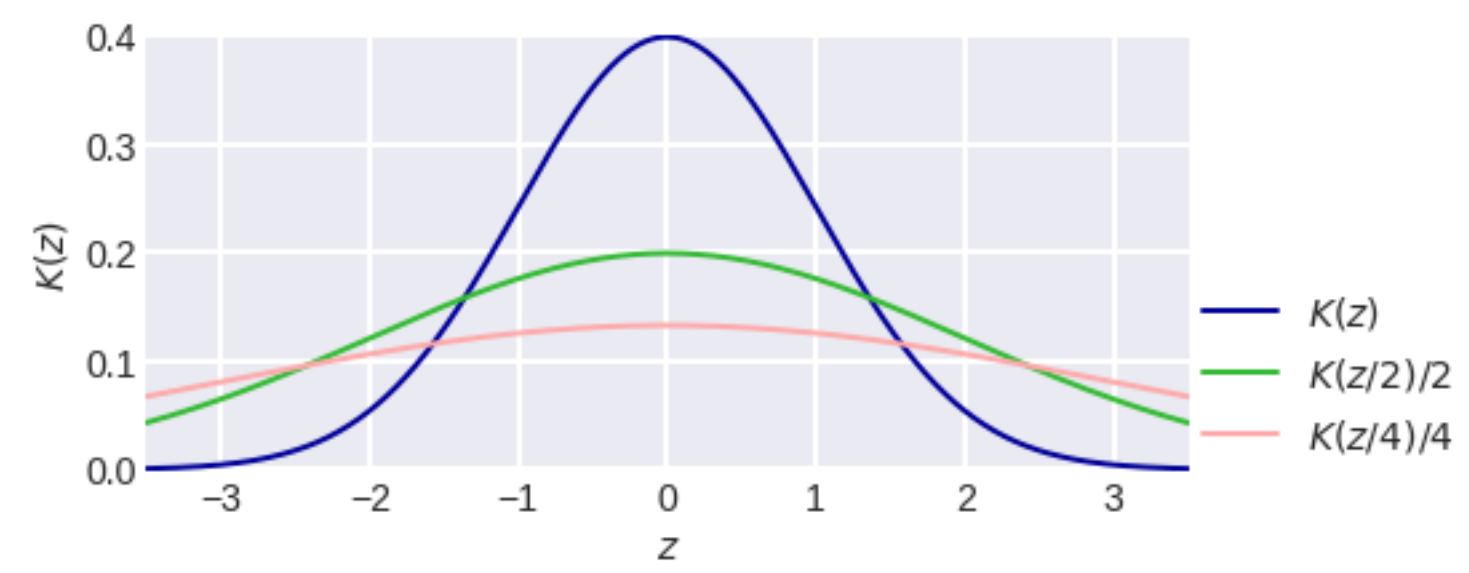
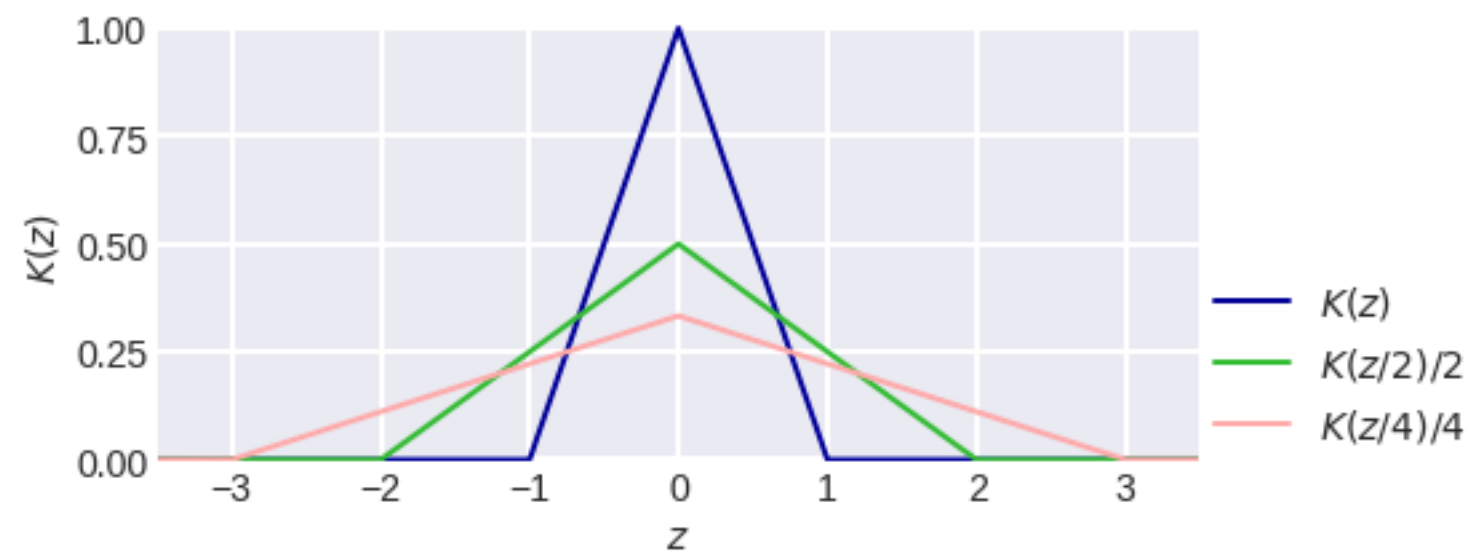
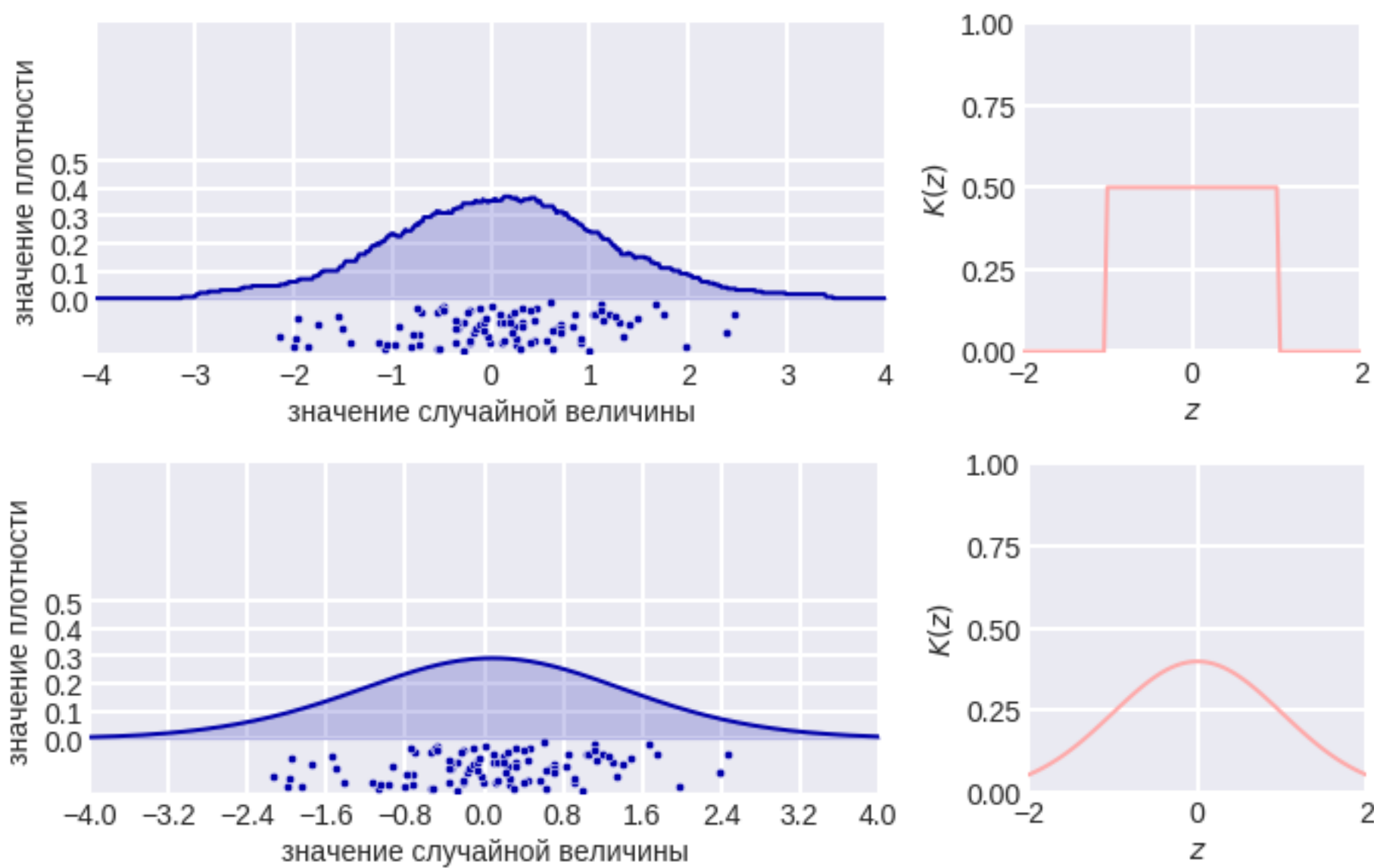


иллюстрация масштабирования

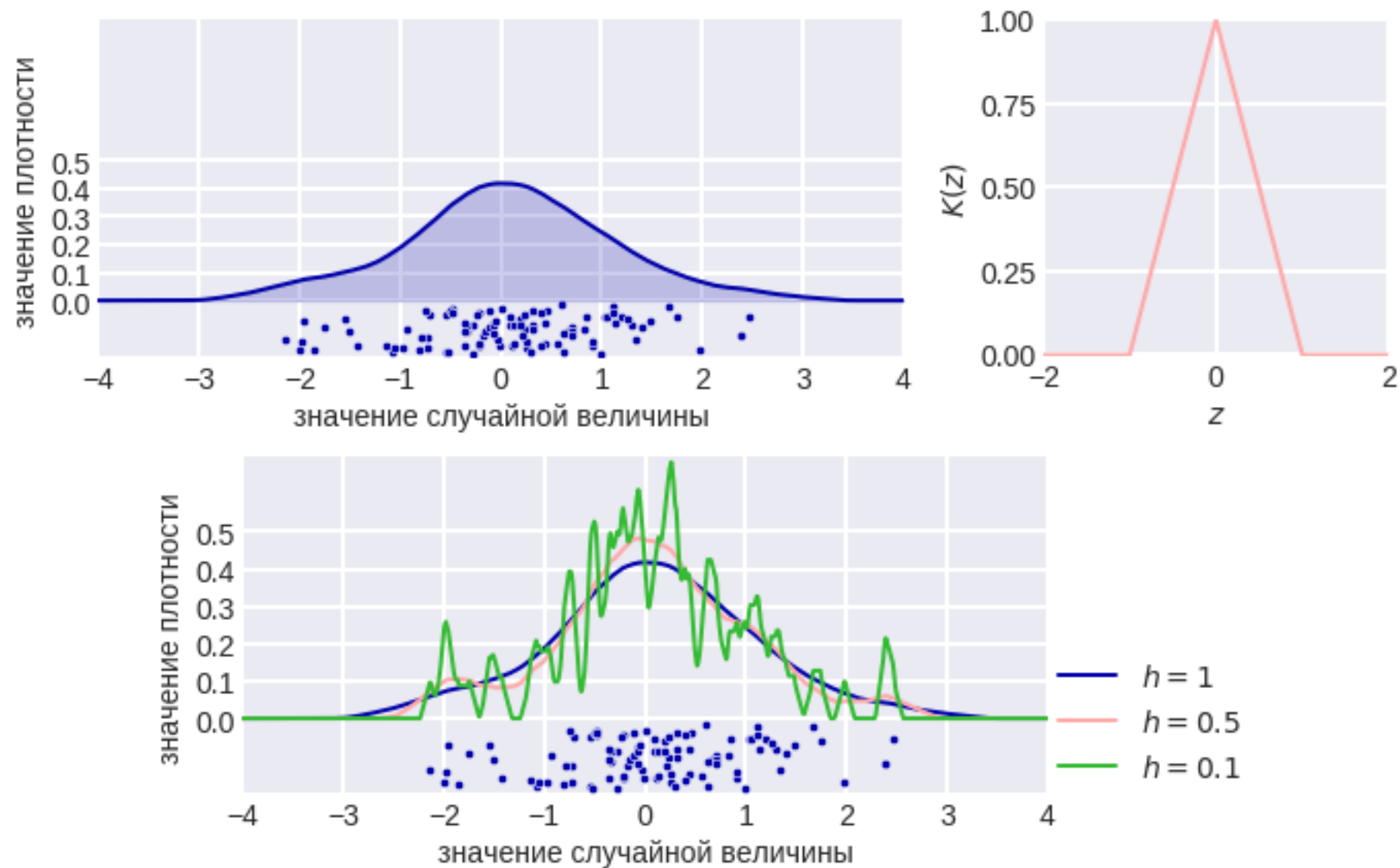
$$\frac{1}{h} K\left(\frac{z}{h}\right)$$

многомерные ядра можно получать в виде произведения одномерных

Парзеновский подход



Парзеновский подход



Парзеновский подход



```
from scipy.stats import gaussian_kde
density = gaussian_kde(x)
xs = np.linspace(-4, 4, 100)
density.covariance_factor = lambda : .3
density._compute_covariance()

i = np.abs(xs-1) <= 0.5
plt.plot(xs, density(xs))
```

Парзеновский подход

Дьяконов, А. Г. Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab (практикум на эвм кафедры математических методов прогнозирования). — МАКСПресс, 2010. — 278 с.

<http://www.machinelearning.ru/wiki/images/7/7e/Dj2010up.pdf>

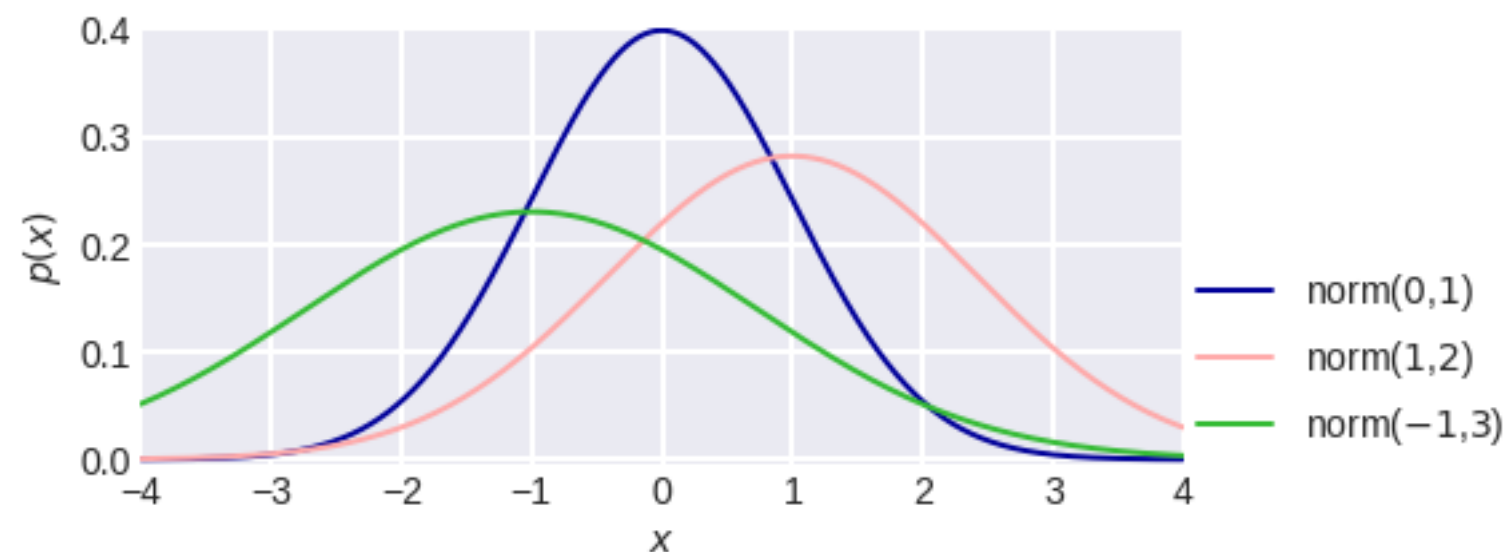
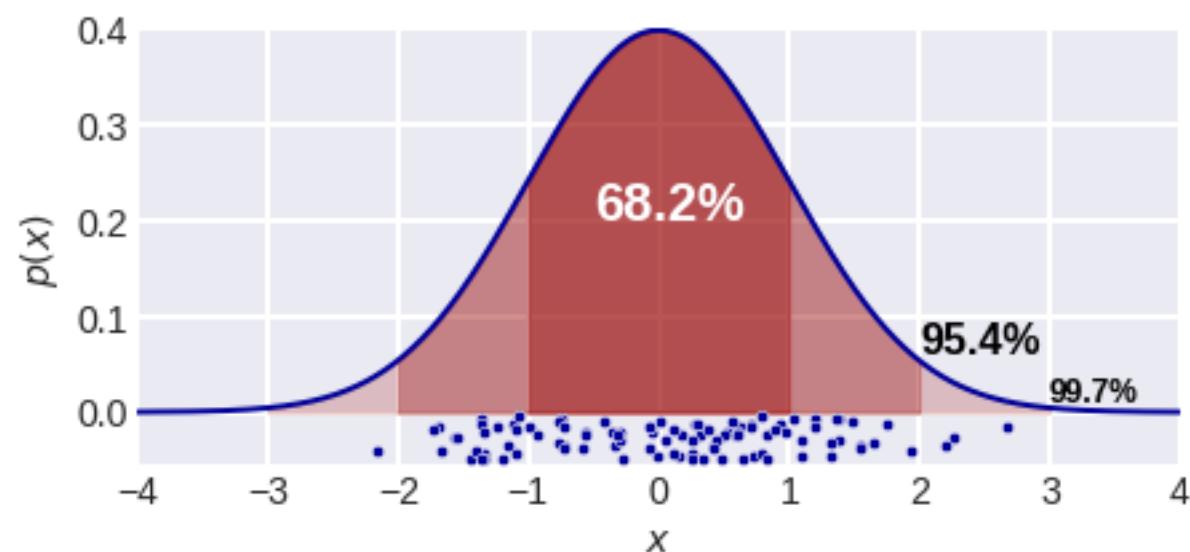
Дьяконов А.Г. Прогноз поведения клиентов супермаркетов с помощью весовых схем оценок вероятностей и плотностей // Бизнес-информатика. 2014. № 1 (27). С. 68–77

<https://bijournal.hse.ru/data/2014/04/15/1320713004/8.pdf>

Пример распределения – нормальное

Одномерное нормальное распределение

$$\text{norm}(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



«около центра данных очень много»

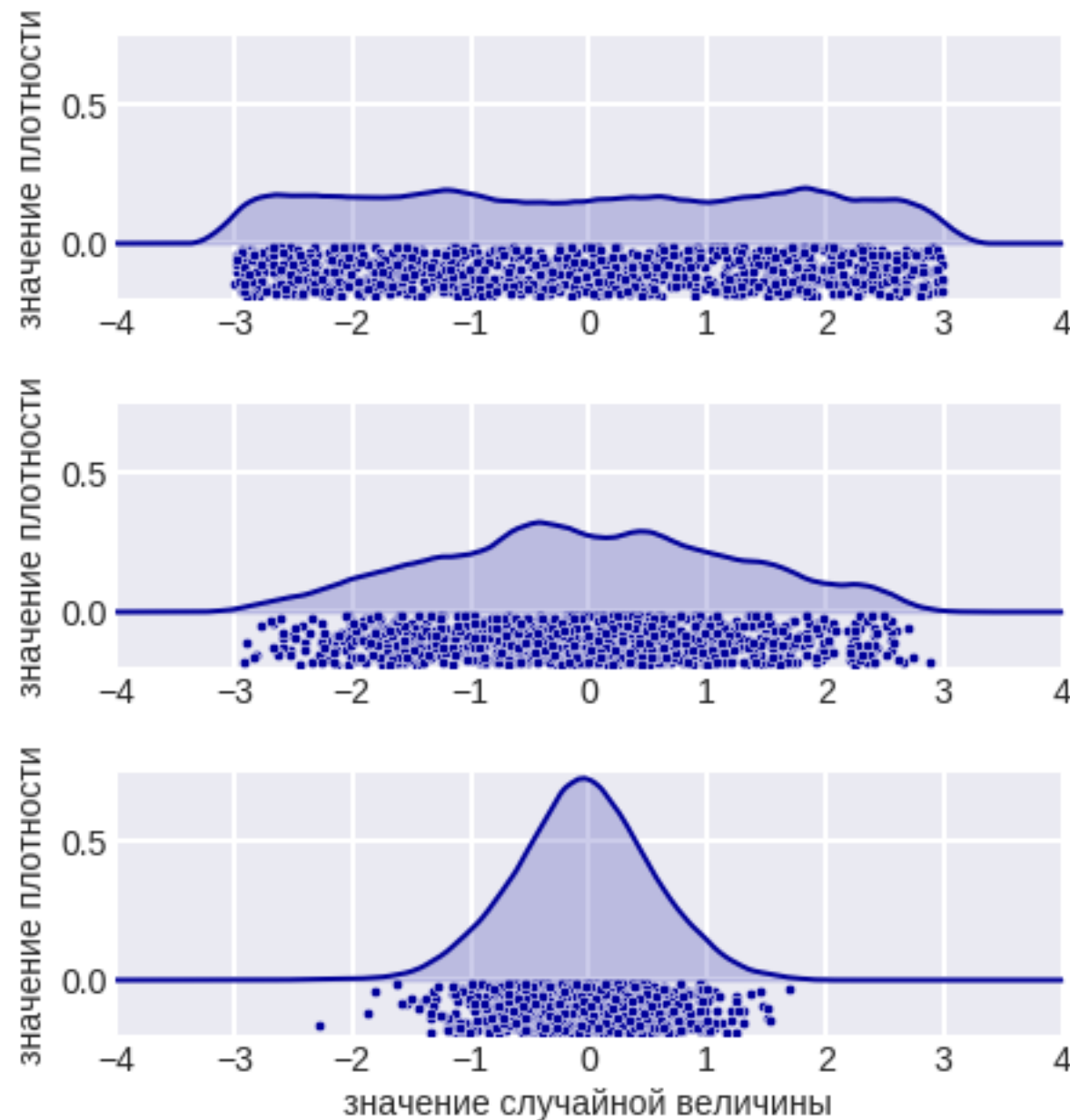
Нет так полезна на практике, как в теории...

Центральная предельная теорема

**о усреднении независимых одинаково распределённых
с конечными м.о. и дисперсией с.в.**

$$\sqrt{m} \frac{\frac{1}{m} \sum_{i=1}^m \xi_i - \mu}{\sigma} \rightarrow \text{norm}(0, 1) \text{ по распределению}$$

Иллюстрация ЦПТ



$$\xi_i \sim U[-3, 3]$$

$$\frac{\xi_1 + \xi_2}{2}$$

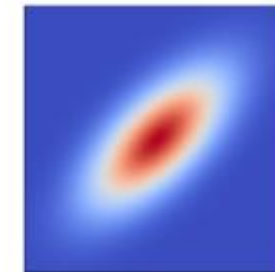
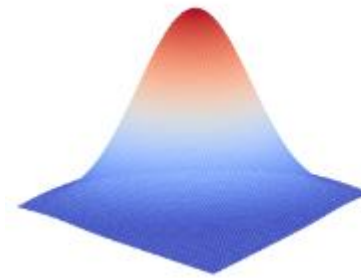
$$\frac{\xi_1 + \dots + \xi_{10}}{10}$$

плотности оценены по Парзену; **похоже на нормальное?**

Многомерное нормальное (гауссовское) распределение

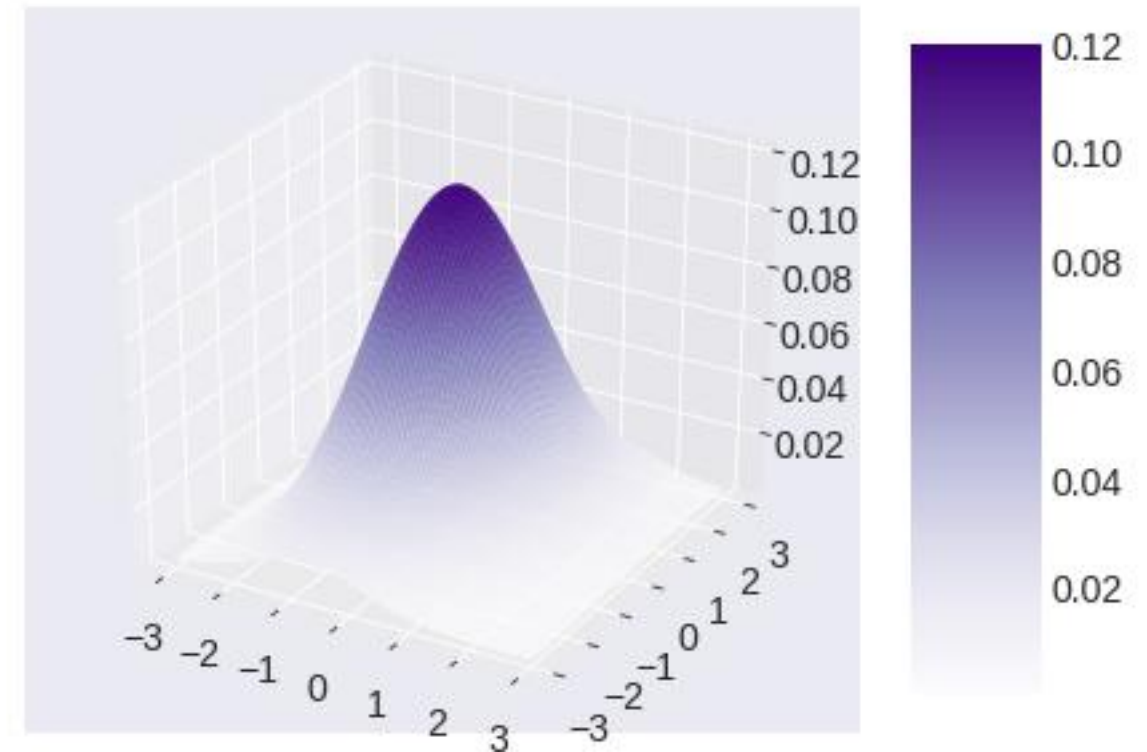
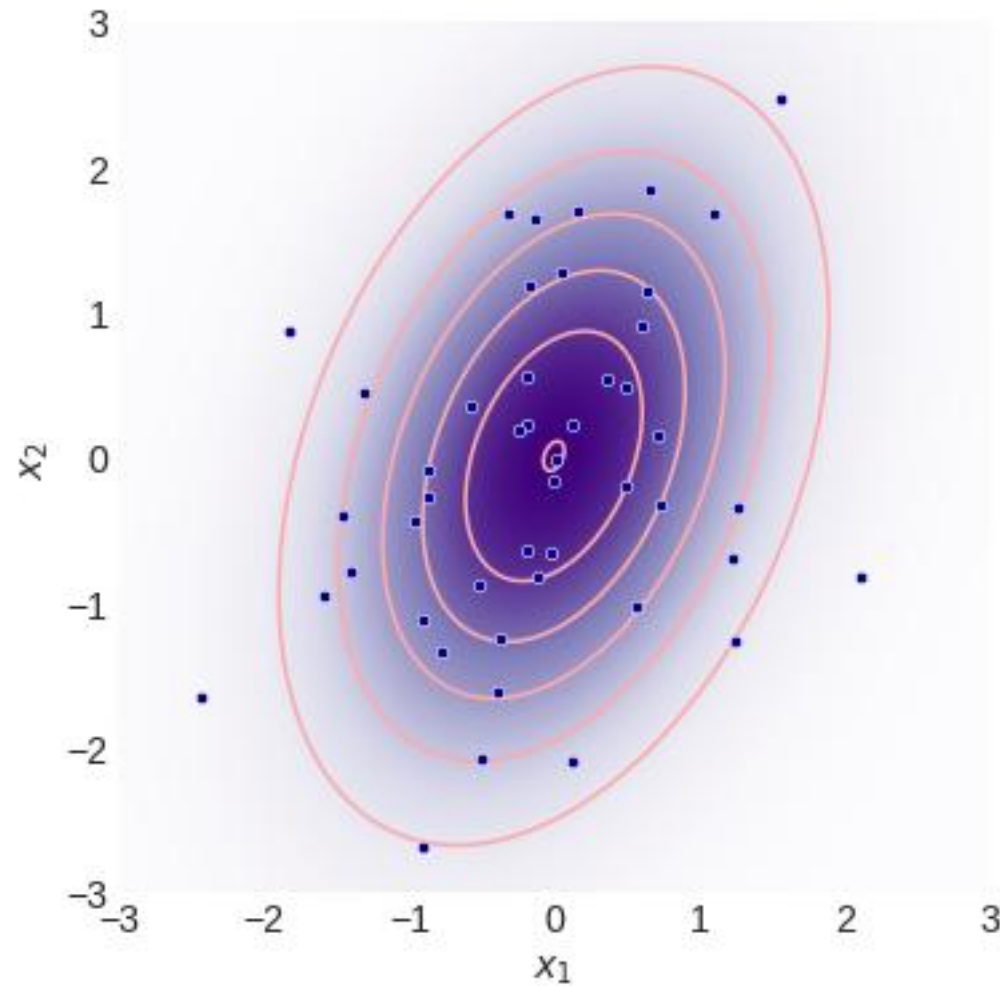
$$\text{norm}(\mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right)$$

$$\mu = \mathbf{E}x$$

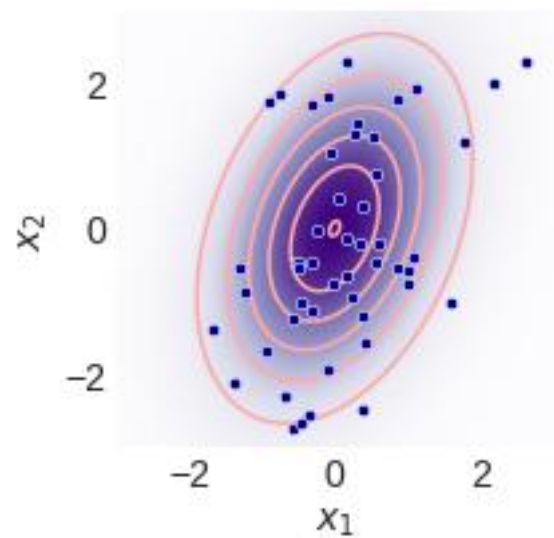


$$\Sigma = \text{cov}(x) = \begin{bmatrix} \text{var}(X_1) & \text{cov}(X_1, X_2) & \dots & \text{cov}(X_1, X_n) \\ \text{cov}(X_2, X_1) & \text{var}(X_2) & \dots & \text{cov}(X_2, X_n) \\ \dots & \dots & \dots & \dots \\ \text{cov}(X_n, X_1) & \text{cov}(X_n, X_2) & \dots & \text{var}(X_n) \end{bmatrix}$$

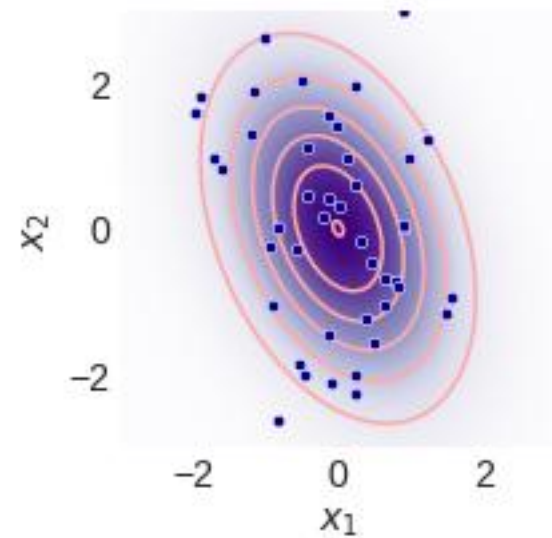
Многомерное нормальное (гауссовское) распределение



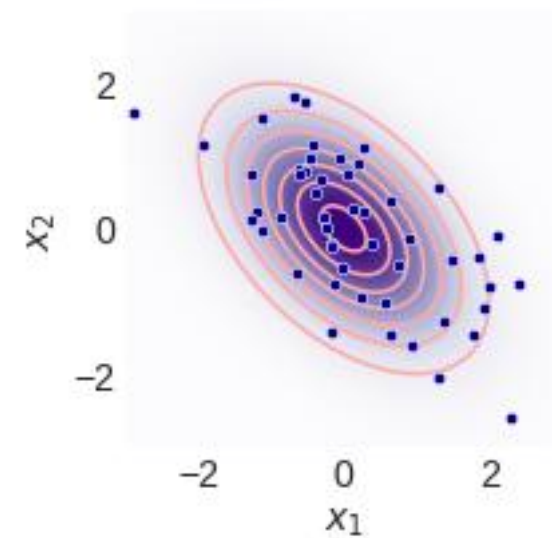
Многомерное нормальное (гауссовское) распределение



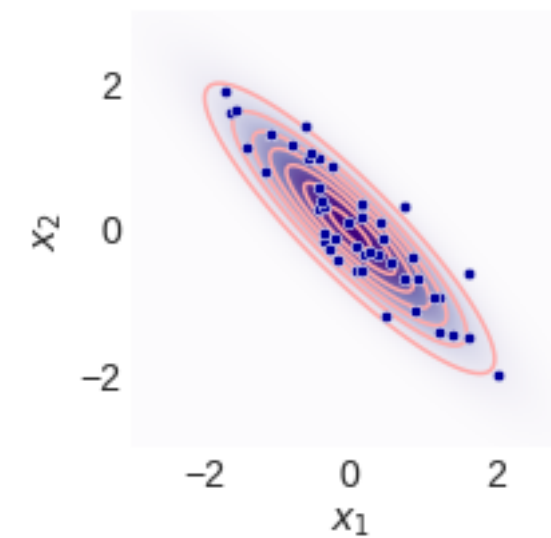
$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 2 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$

Теория информации

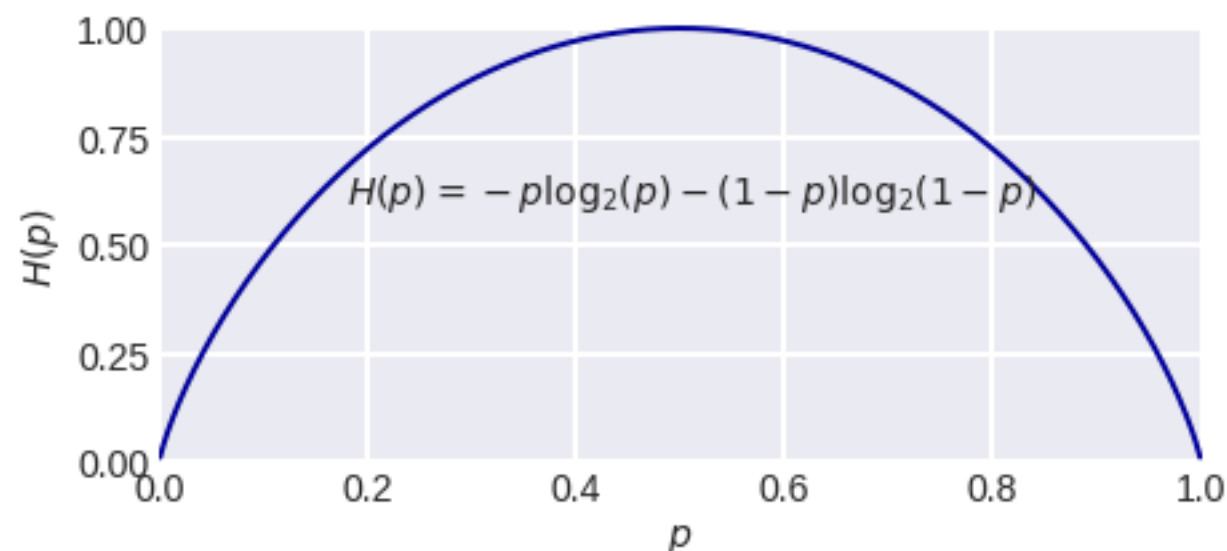
Информационная энтропия (Entropy) – мера неопределённости некоторой системы

$$x \sim (x_1, p_1), (x_2, p_2), \dots$$

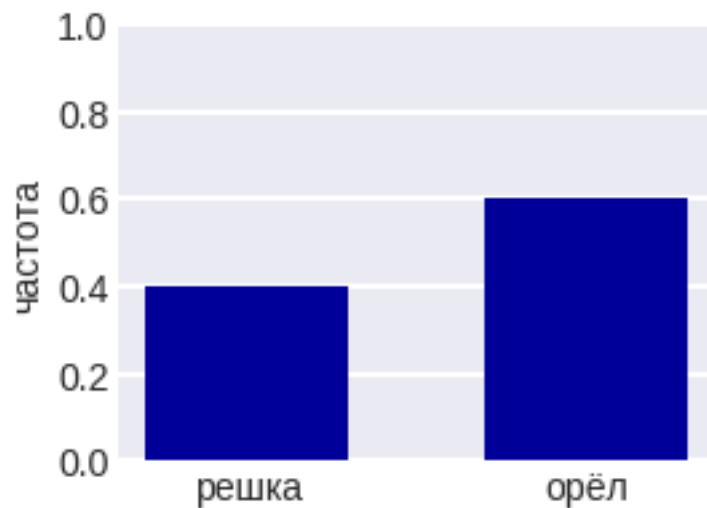
$$H(x) = -\sum_t p_t \log p_t$$

$$H(p) = -p \log_2 p - (1-p) \log_2 (1-p)$$

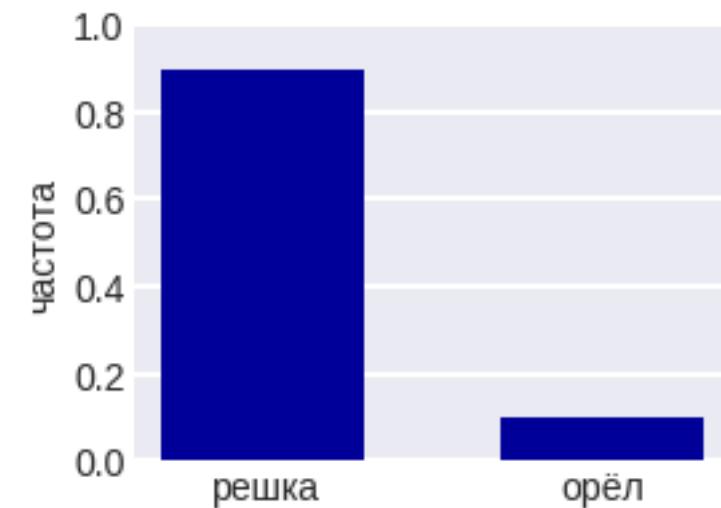
Что зависит от основания логарифма?



Теория информации



$$-\frac{4}{10} \log_2 \frac{4}{10} - \frac{6}{10} \log_2 \frac{6}{10} \approx 0.97$$



$$-\frac{9}{10} \log_2 \frac{9}{10} - \frac{1}{10} \log_2 \frac{1}{10} \approx 0.47$$

Результат подбрасывания честной монеты – 1 бит информации

Проклятие размерности

Объём шара радиуса r в \mathbb{R}^n

$$\text{vol}(r) = \frac{\pi^{n/2}}{\Gamma(n/2 + 1)} r^n$$

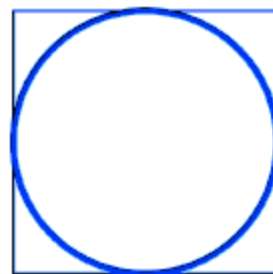
$$n = 1$$

$$\text{vol}(r) = 2r$$



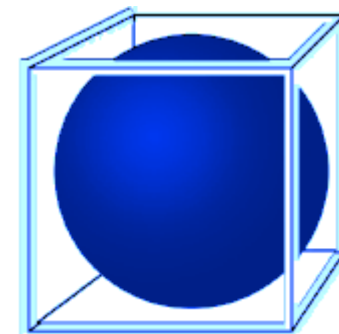
$$n = 2$$

$$\text{vol}(r) = \pi r^2$$



$$n = 3$$

$$\text{vol}(r) = \frac{4}{3} \pi r^3$$



ДЗ доказать

Проклятие размерности (The Curse of Dimensionality)

Весь объём сосредоточен «на краю» шара

$$\frac{\text{vol}(r + \varepsilon)}{\text{vol}(r)} = \left(1 + \frac{\varepsilon}{r}\right)^n \xrightarrow{n \rightarrow +\infty} +\infty$$

100% → 79% → 52% → ... → 0

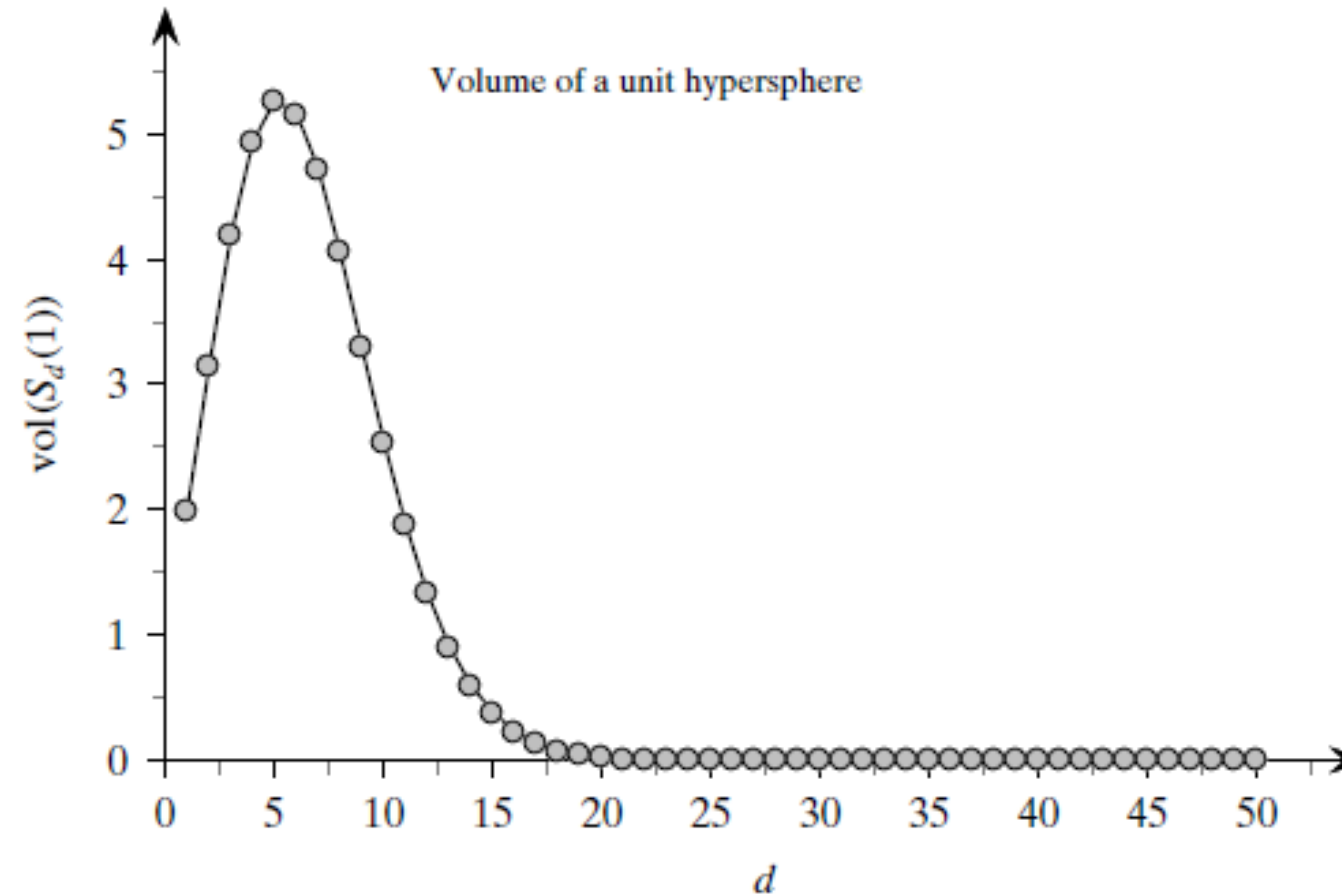
скорее всего, соседи будут с краю...

но это в предположении, что объекты «равномерно» разбросаны по пространству,
а в реальности лежат около поверхностей малых размерностей

<http://mc-stan.org/users/documentation/case-studies/curse-dims.html>

Проклятие размерности (The Curse of Dimensionality)

Объём единичного шара при росте размерности



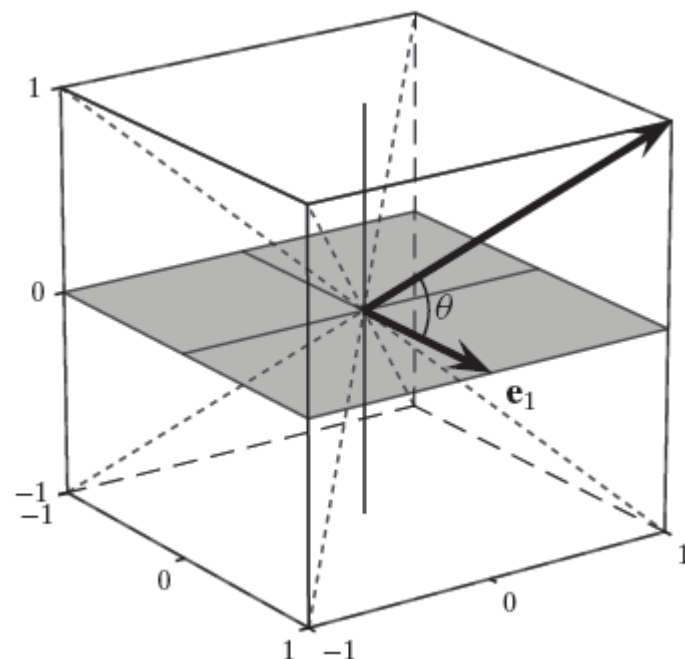
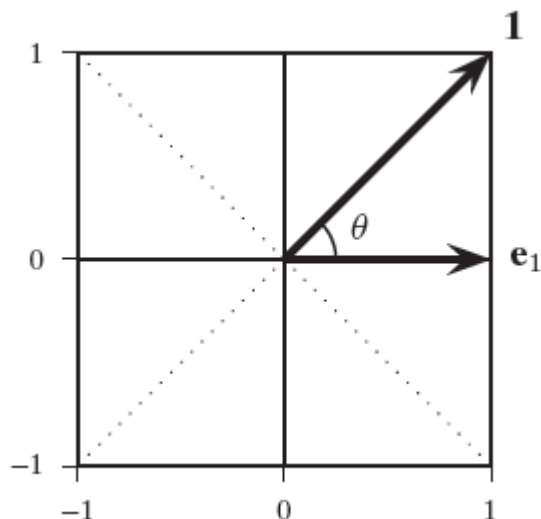
НЕТ СООТВЕТСТВИЯ ИНТУИЦИИ

Проклятие размерности (The Curse of Dimensionality)

Угол между диагональю и первым базисным вектором

$$\cos \theta = \frac{(1, 0, \dots, 0)^T \cdot (1, \dots, 1)}{\| (1, 0, \dots, 0) \| \cdot \| (1, \dots, 1) \|} = \frac{1}{\sqrt{n}} \rightarrow 0$$

$$\lim_{n \rightarrow \infty} \theta = \frac{\pi}{2}$$



Проклятие размерности (The Curse of Dimensionality)

**У нормально распределённых данных
при росте размерности «масса» смещается в хвосты**

ДЗ обосновать

**Как следствие из результата о нормах,
метрики в конечномерном пространстве эквивалентны**

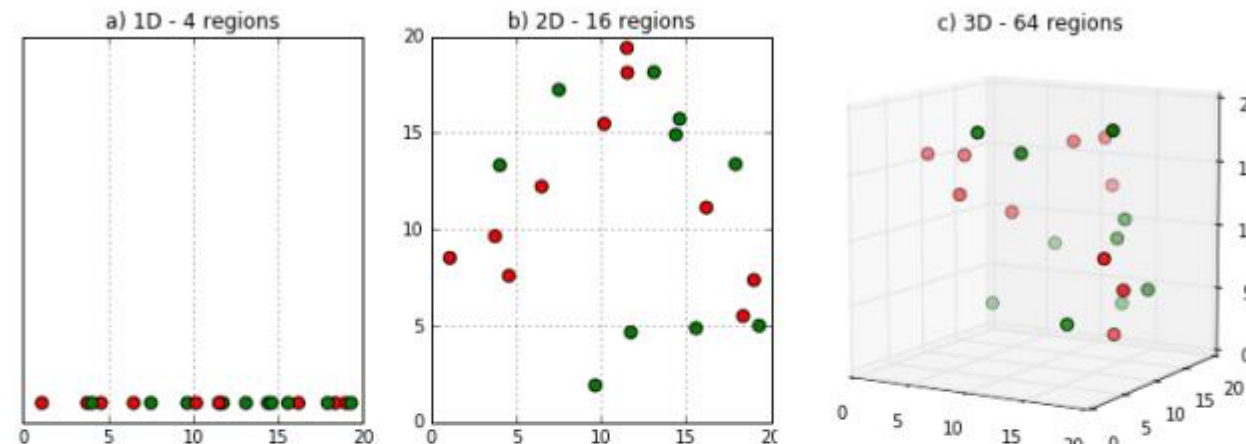
$$\forall z \ c\rho(x, z) \leq d(x, z) \leq C\rho(x, z)$$

но при росте размерности «по сути» они отличаются

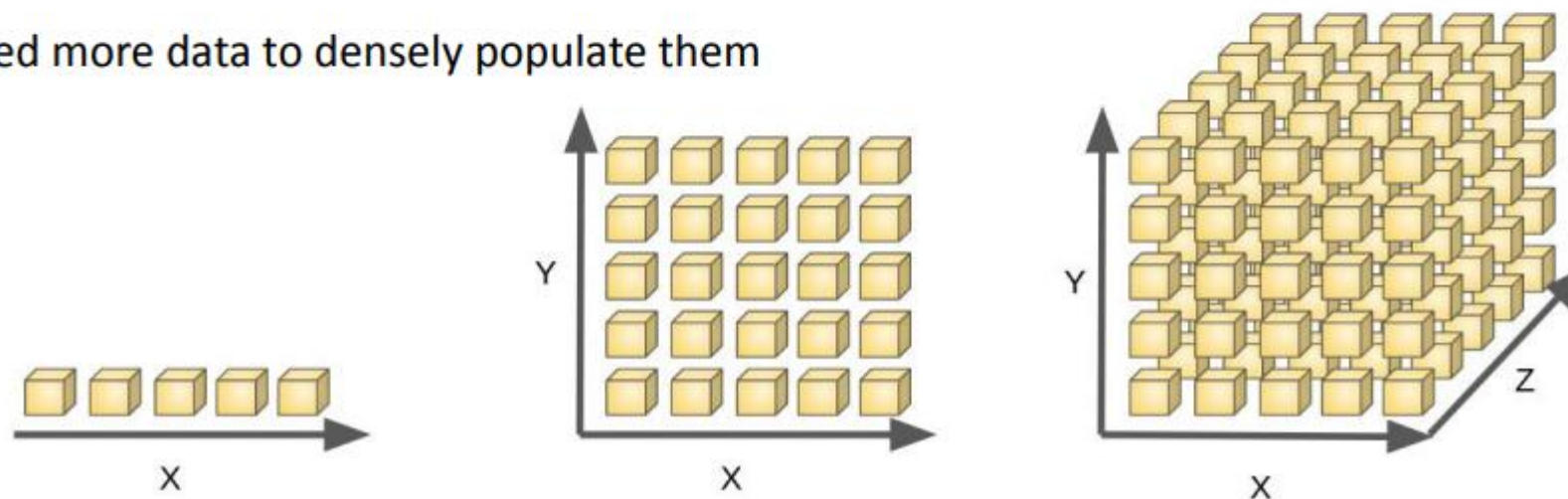
Проклятие размерности (The Curse of Dimensionality)

необходимость больших данных с ростом числа признаков

Regions become more sparsely populated given the same amount of data



Need more data to densely populate them



Сингулярное разложение матрицы (SVD)

любая $m \times n$ -матрица ранга k представляется в виде произведения

$$X_{m \times n} = U_{m \times k} \cdot \Lambda_{k \times k} \cdot V_{n \times k}^T$$

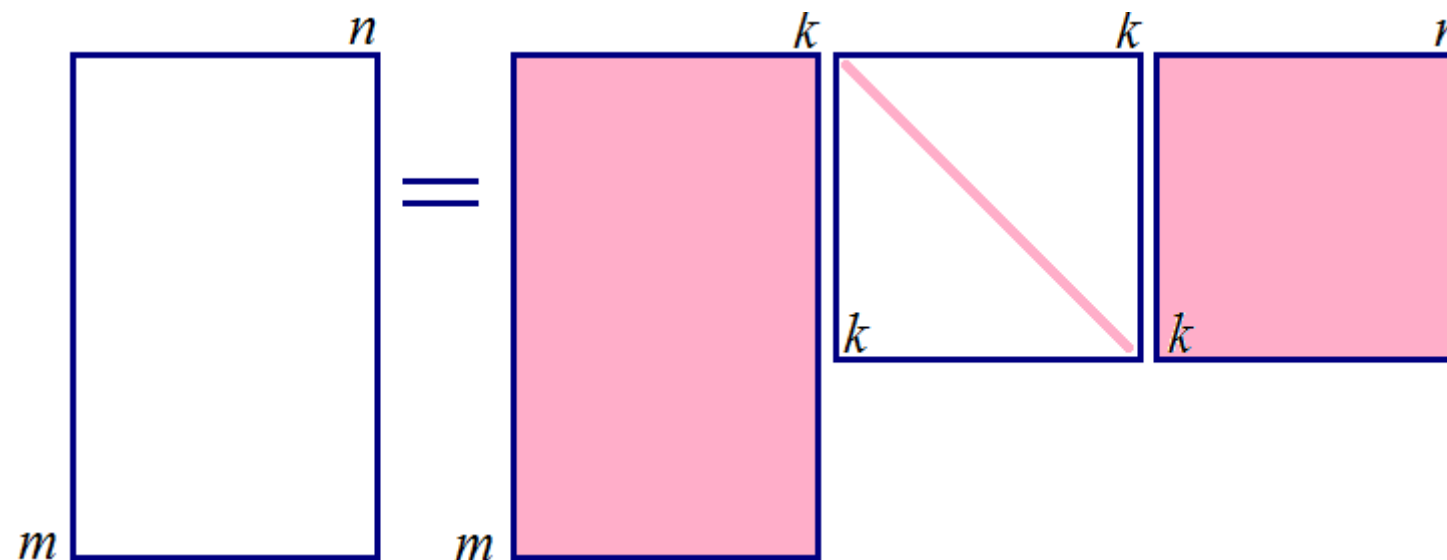
$$X = \sum_{i=1}^k \lambda_i u_i v_i^T$$

где $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_k)$

$$\lambda_1 \geq \dots \geq \lambda_k > 0$$

$$U^T U = I$$

$$V^T V = I$$



Сингулярное разложение матрицы (SVD)

$$X^T X = (U \Lambda V^T)^T U \Lambda V^T = V \Lambda^T U^T U \Lambda V^T = V \Lambda^2 V^T$$

поэтому

$$X^T X V = V \Lambda^2$$

**и матрица V состоит из с.в. матрицы $X^T X$,
которым соответствуют с.з. $\lambda_1^2 \geq \dots \geq \lambda_k^2 > 0$**

$\lambda_1 \geq \dots \geq \lambda_k > 0$ – сингулярные числа

аналогично матрица U состоит из с.в. матрицы XX^T с теми же с.з.

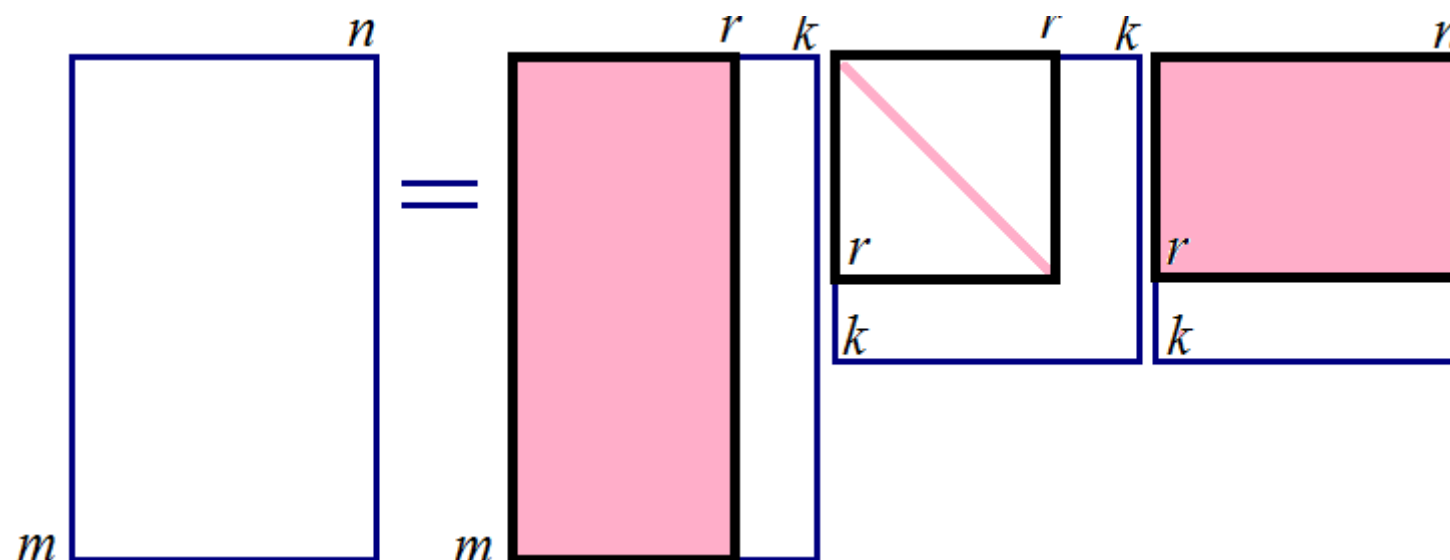
Усечённое сингулярное разложение матрицы (Truncated SVD)

что будет если

$$X = U \Lambda V^T$$

$$\Lambda' = \text{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0)$$

$$U \Lambda' V^T = \dots \sum_{i=1}^r \lambda_i u_i v_i^T \equiv X'$$



$$X' = \arg \min_{H: \text{rank } H = r} \| X - H \|_2^2 = \sum_{i=r+1}^k \lambda_i^2$$

Применение SVD

- Для матриц малого ранга – экономное хранение
- Для произвольных матриц – приближение и сжатие
 - Регуляризация
 - Основа некоторых методов рекомендаций
- Основа некоторых методов тематического моделирования
 - Основа некоторых методов сокращения размерности

Реконструкция изображений с помощью SVD



k=5



k=10



k=20



k=50



k=100



k=200

Изначальный размер изображения $300 \times 451 = 135\,300$

$300 \times 50 + 50 + 50 \times 451 = 37\,600$

Минутка кода

```
from numpy.linalg import svd
U, L, V = svd(image)
k = 5
plt.imshow(U[:, :k].dot(np.diag(L[:k])).dot(V[:k, :]),
            cmap=plt.cm.gray)
```

Устойчивость к шумам



k=20



k=50



исходное

Матричное дифференцирование

$$\frac{\partial(a^T w)}{\partial w} = \frac{\partial(w^T a)}{\partial w} = a$$

$$\frac{\partial(A_1 W_1 + \dots + A_n W_n)}{\partial(W_1, \dots, W_n)} = \begin{bmatrix} \frac{\partial(A_1 W_1 + \dots + A_n W_n)}{\partial W_1} \\ \vdots \\ \frac{\partial(A_1 W_1 + \dots + A_n W_n)}{\partial W_n} \end{bmatrix} = \begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix}$$

Книги

Математика в ML

Marc Peter Deisenroth, A Aldo Faisal, and Cheng Soon Ong "Mathematics for Machine Learning", 2019, <https://mml-book.github.io>

Последняя книга по математике в ML

Jean Gallier, Jocelyn Quaintance «Algebra, Topology, Differential Calculus, and Optimization Theory for Computer Science and Machine Learning» // Book in Progress, 1958 pp. (2019) <http://www.cis.upenn.edu/~jean/gbooks/geomath.html>