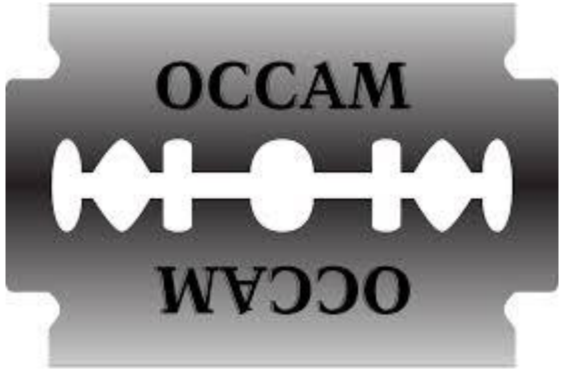
Математика в машинном обучении: краткий обзор

Александр Дьяконов

План

- Общие факты: бритва Оккама, «бесплатный сыр», футбольный оракул
- ТВиМС: распределения, условная плотность, оценки, ММП, оценка плотности
- Теория информации
- Проклятие размерности
- Сингулярное разложение матрицы (SVD)
- Матричное дифференцирование
- Статистические гипотезы

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…» // Альберт Эйнштейн

Теорема о бесплатном сыре (No Free Lunch Theorem)

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

Сложность

Чуть позже коснёмся

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 ×10 (т.е. обзвонить 160 человек), то перед финалом 10 человек, которым безошибочно сказали 4 исхода!

аналогично, если бы мы случайно давали прогнозы…

Важно правильно формировать выборку и ставить эксперимент!

Сведения из ТВиМС

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

Есть и другой подход к пониманию вероятности!
ЗБЧ: частота → вероятность

→
теория вероятностей

←
математическая статистика

Сведения из ТВиМС

Как задать распределение с.в. ξ

Если принимает значения $x_1, x_2, \ldots$, то вероятностями

$$p_1 = \mathbf{P}(\xi = x_1),\ p_2 = \mathbf{P}(\xi = x_2), \ldots$$

$$\sum_i p_i = 1,\ p_i \geq 0$$

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

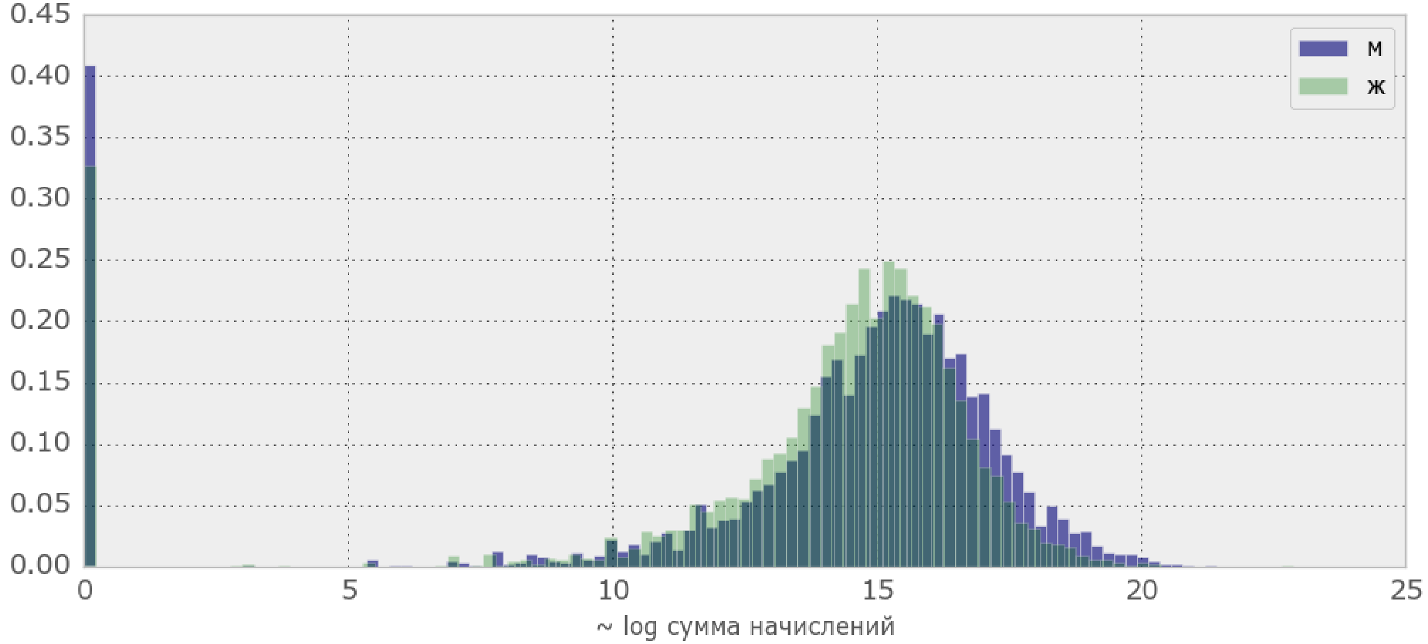
$$\mathbf{P}(a \le \xi \le b) = \int_{a}^{b} p(x)\partial x$$

Связь плотности и вероятности

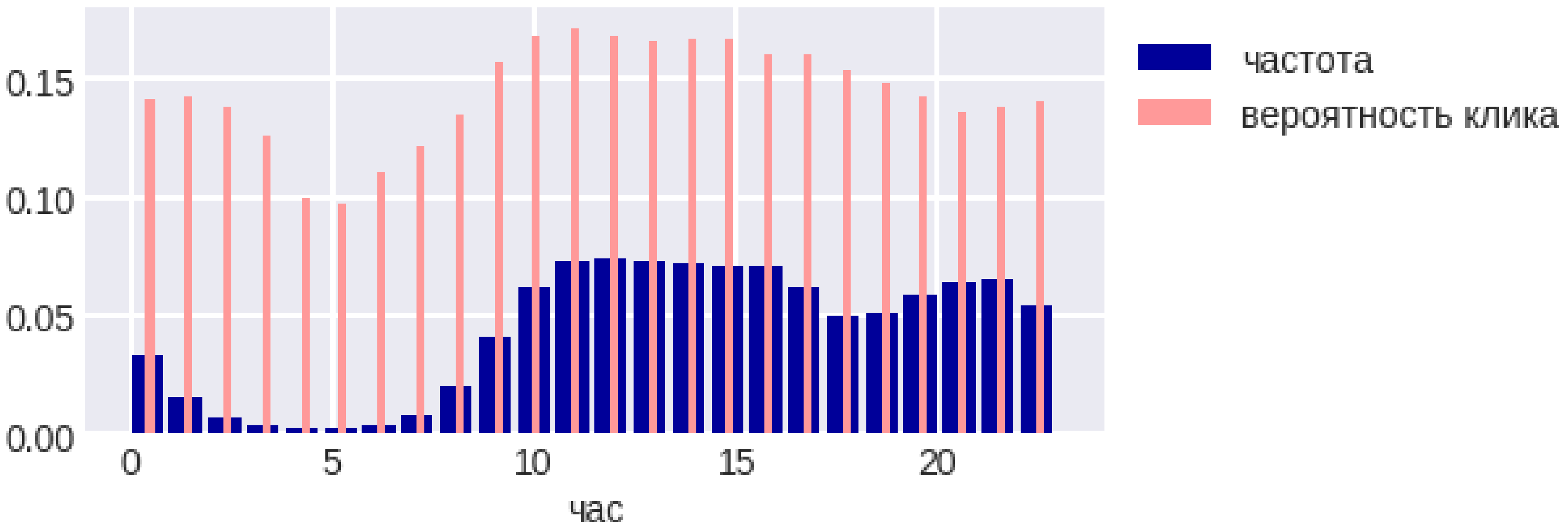
$$\mathbf{P}(x-\varepsilon \le \xi \le x+\varepsilon) = \int_{x-\varepsilon}^{x+\varepsilon} p(z)\partial z \approx 2\varepsilon p(x)$$

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

Примеры распределений из жизни: сбербанк



Примеры распределений из жизни: тикетлэнд



Сведения из ТВиМС

Пусть с.в. имеет плотность $p(x)$

Математическое ожидание (~центр масс) **–**

$$\mathbf{E}X = \int x p(x) \partial x$$

Дисперсия (средний квадрат отклонения от МО) –

$$\mathbf{E}(X - \mathbf{E}X)^2 = \int (x - \mathbf{E}X)^2 p(x) \partial x$$

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

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

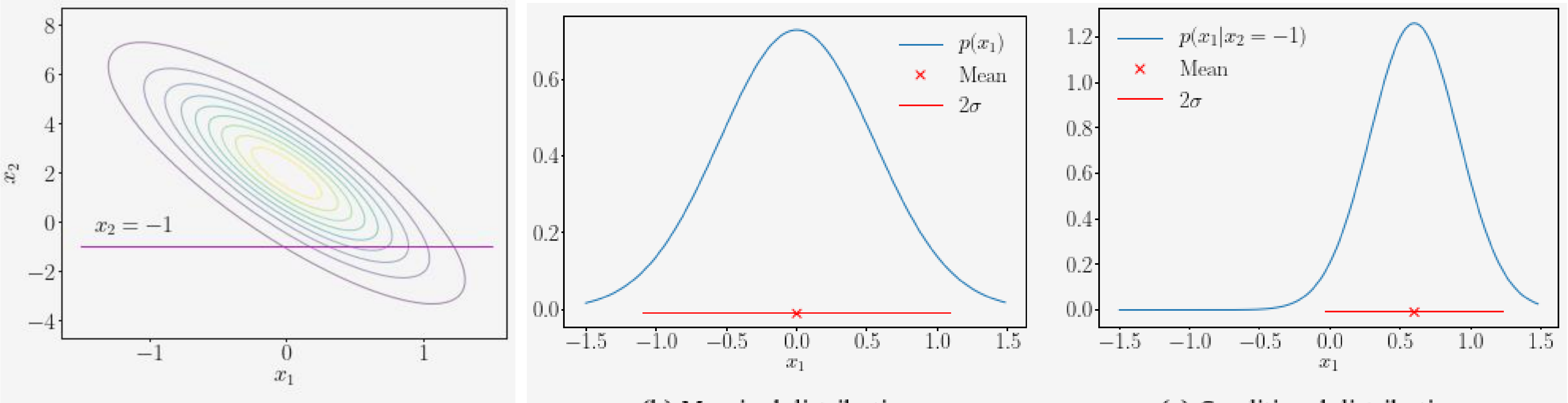
Сведения из ТВиМС

Маргинализация плотности по неизвестной компоненте

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

Обуславливание плотности по известной компоненте

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



(a) Bivariate Gaussian. (b) Marginal distribution. (c) Conditional distribution.

Правило произведения

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

Точечное оценивание

Зачем нужно?

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

Выборка $\{x_1,\ldots,x_m\}$

(независимые одинаково распределённые случайные величины)

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

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

Это тоже случайная величина!

примеры

Требования к статистике

1) Значение должно быть близко к истинному значению параметров модели θ

Смещение $\mathrm{bias}(\hat{\theta}) = \mathbf{E}\hat{\theta} - \theta$

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

Асимптотически несмещённая оценка $\mathrm{bias}(\hat{\theta}) \to 0$

Для нормального распределения несмещённые оценки:

$$\hat{\mu} = \frac{x_1 + \ldots + x_m}{m}$$

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

Требования к статистике

2) Оценка не должна сильно варьировать
в зависимости от выборки

$$\mathrm{var}(\hat{\theta}) \to \min$$

Пример:

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

Требования к статистике

3) с ростом числа наблюдений должна быть сходимость

Состоятельность (Consistency) –

$$\hat{\theta} \xrightarrow[P]{} \theta$$

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

Пример:

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

Оценка Maximum Likelihood Estimation (MLE / ММП)

$$\{y_1,\ldots,y_m\}$$

независимые, одинаково распределённые

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

м.б. смещённая

1) состоятельная

2) асимптотически эффективная (среди асимптотически нормальных**)**
и асимптотически нормальная

эффективность вводится в классе оценок:

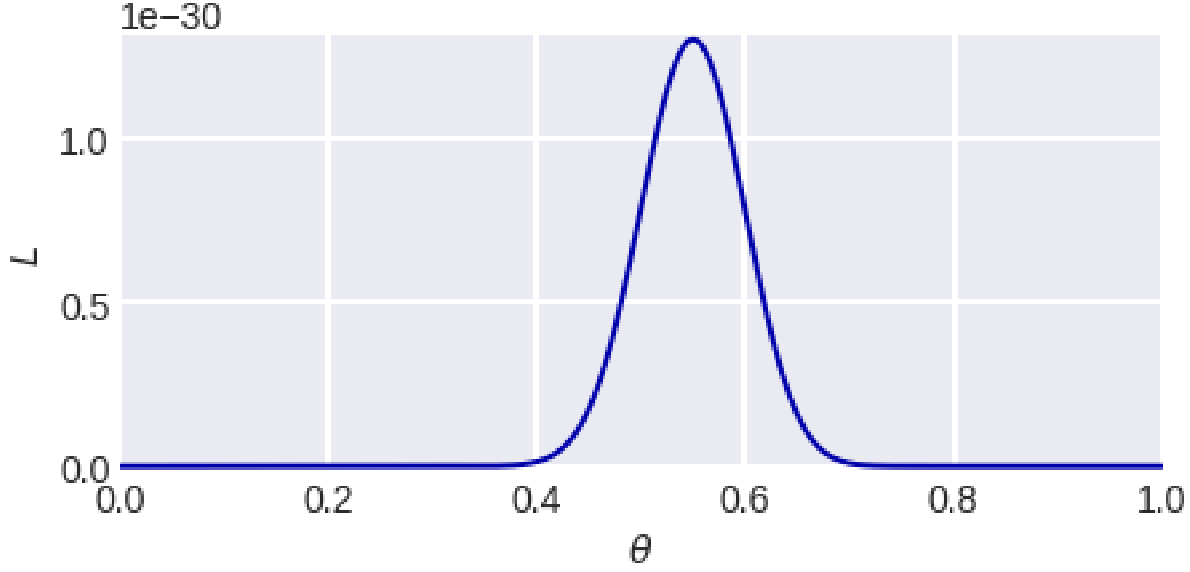
$$\mathbf{E}\,|\hat{\theta} - \theta|^2 \le \mathbf{E}\,|\hat{\theta}' - \theta|^2$$

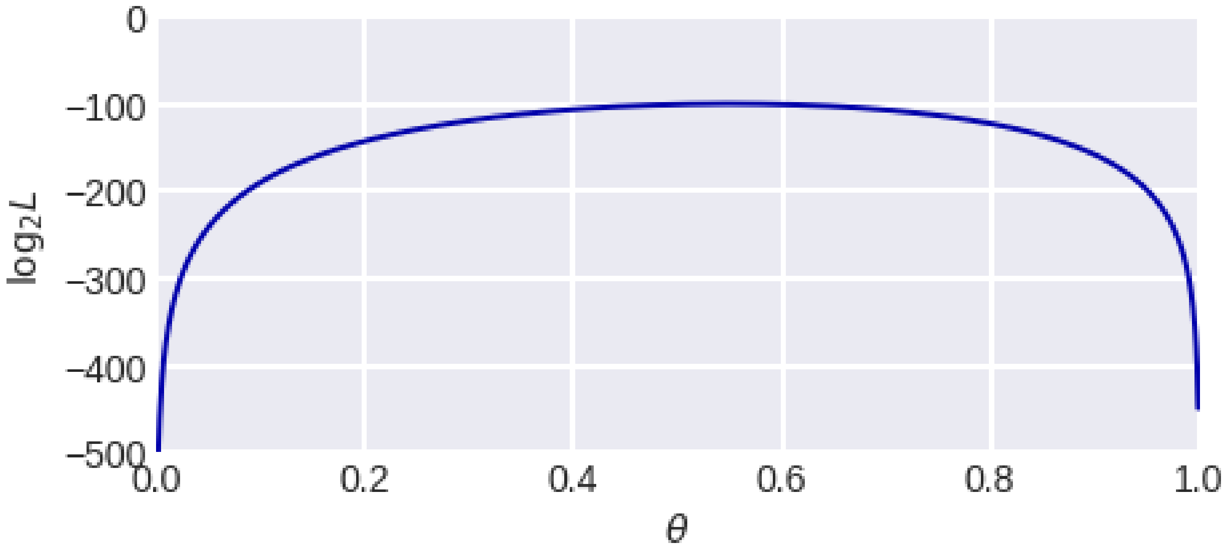
эффективная – несмещенная оценка, имеющая наименьшую дисперсию из всех возможных несмещенных оценок данного параметра

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

кроме MLE есть ещё, например, метод моментов, MAP (будет)

Откуда берётся дивергенция Кульбака-Лейблера

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

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

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{m} q(x_i\,|\,\theta) = \arg\max_{\theta} \sum_{i=1}^{m} \log q(x_i\,|\,\theta) =$$

$$= \arg\max_{\theta} \frac{1}{m}\sum_{i=1}^{m} \log q(x_i\,|\,\theta) - \frac{1}{m}\sum_{i=1}^{m} \log p(x_i)$$

$$\sim \arg\min_{\theta} \frac{1}{m}\sum_{i=1}^{m} \log \frac{p(x_i)}{q(x_i\,|\,\theta)}$$

$$\sim \arg\min_{\theta} \underbrace{\int \log \frac{p(x)}{q(x\,|\,\theta)} p(x)\partial x}_{\mathrm{KL}(p\|q_\theta)}$$

Дивергенция Кульбака-Лейблера

$$\mathrm{KL}(p\,||\,q_\theta) = \mathbf{E}_p[\log p(x) - \log q(x\,|\,\theta)] = \int \log \frac{p(x)}{q(x\,|\,\theta)} p(x) \partial x$$

кстати, если хотим минимизировать,

то достаточно (истинное распределение не знаем)

$$\mathbf{E}_p[\log q(x\,|\,\theta)] \to \max$$

а это и есть метод максимального правдоподобия!

Попытка совместить распределение-оценку с истинным…

Энтропия (Entropy)	$H(p) = \mathbf{E}_{x\sim p}[-\ln p(x)]$
Перекрёстная энтропия (CrossEntropy)	$H(p,q) = \mathbf{E}_{x\sim p}[-\ln q(x)]$
KL-дивергенция	$\mathrm{KL} = H(p,q) - H(p) = \mathbf{E}_{x\sim p} \ln \frac{p(x)}{q(x)} \geq 0$

Взаимная информация

$$I(x,y) = \mathrm{KL}(p(x,y)\,||\,p(x)p(y)) =$$

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

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

Ковариация и корреляция

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

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

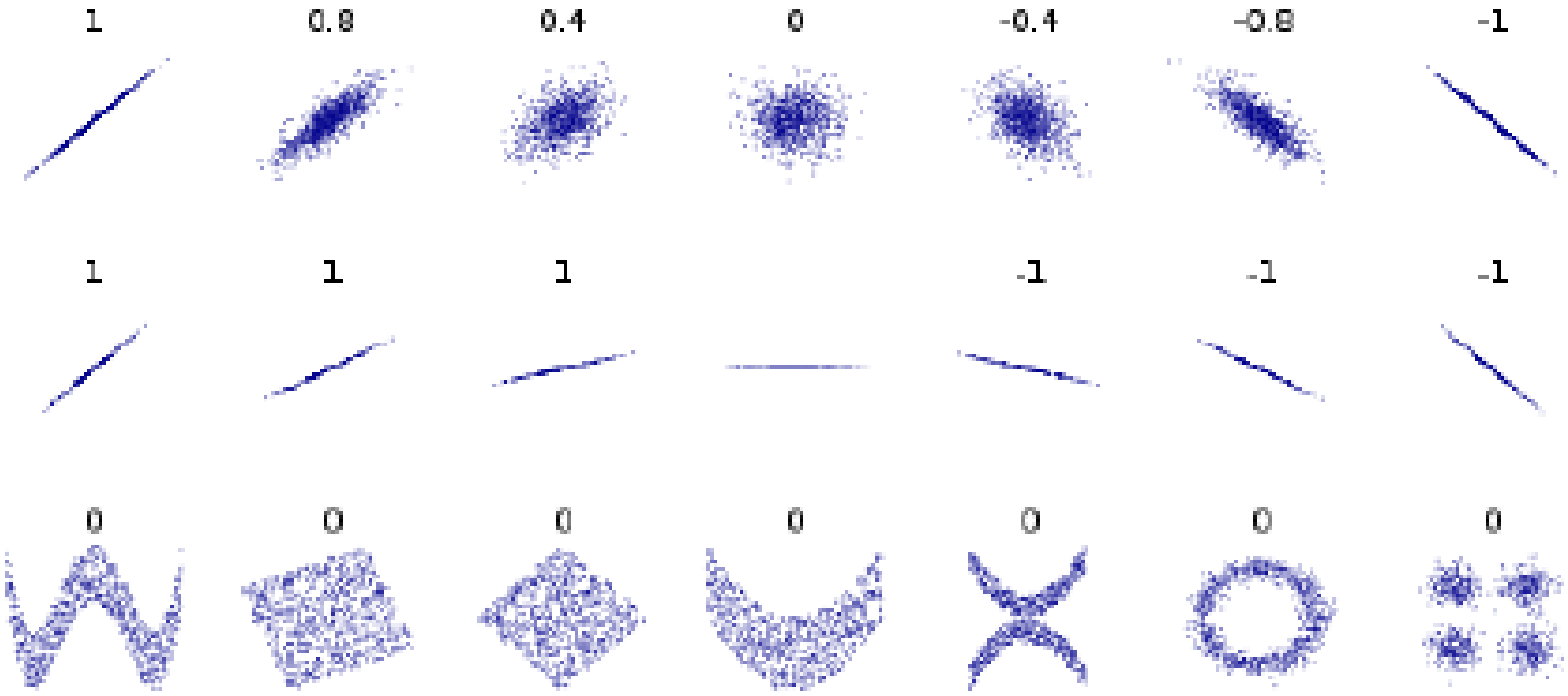
Корреляционный коэффициент Пирсона (Pearson):

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

определяет меру линейной зависимости между с.в.

Независимые переменные некоррелированны.
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

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

$$= 1 - \frac{6}{m^3 - m}\sum_{i=1}^{m}\left(\mathrm{rank}(x_i) - \mathrm{rank}(y_i)\right)^2$$

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

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

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: гистограммный подход


```
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:

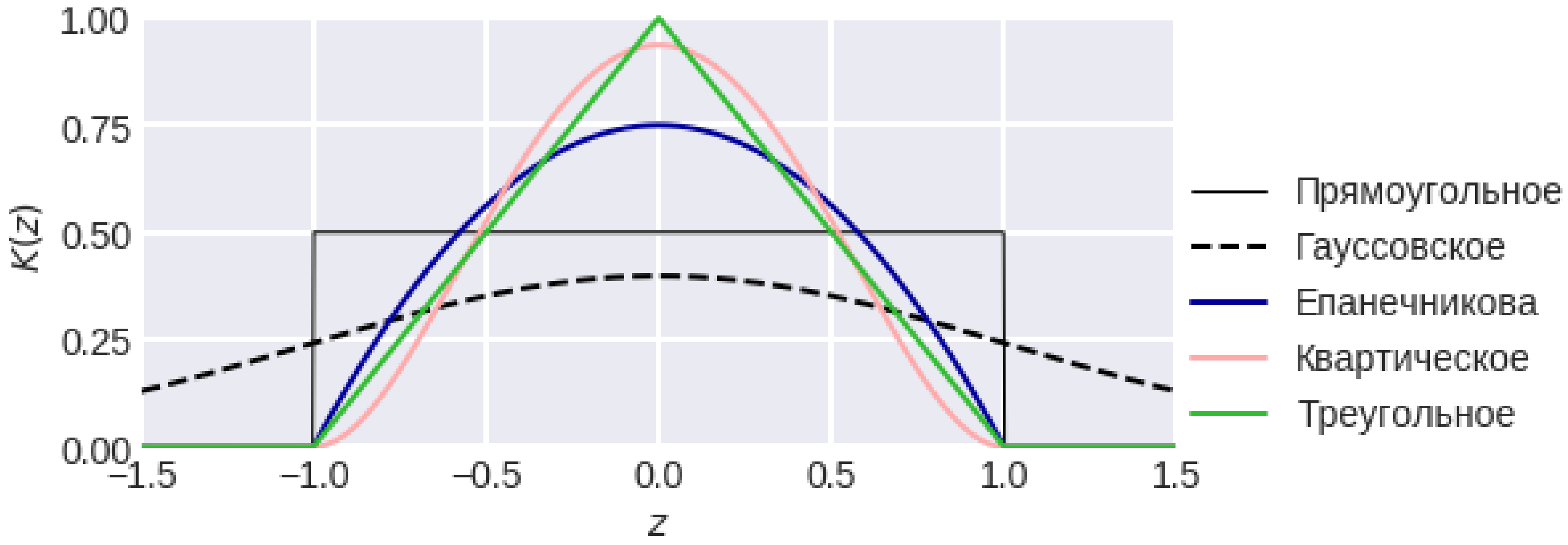
$$K(z)\geq 0$$

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

плотность = среднее локальных плотностей

есть теорема о сходимости

Различные виды ядер (одномерных)



Прямоугольное / tophat

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

Гауссовское / gaussian

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

Треугольное / linear

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

Квартическое

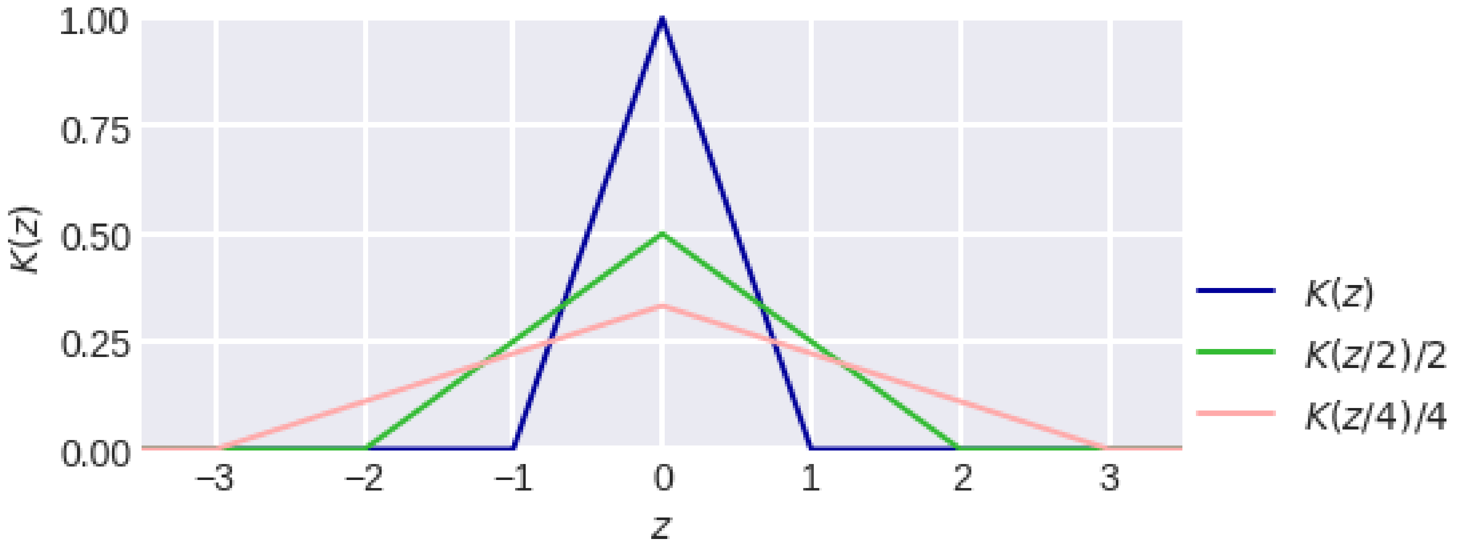
$$K(z)=\frac{15}{16}(1-z^2)^2 I[|z|\le 1]$$

Епанечникова / epanechnikov

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

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

Различные виды ядер (одномерных)



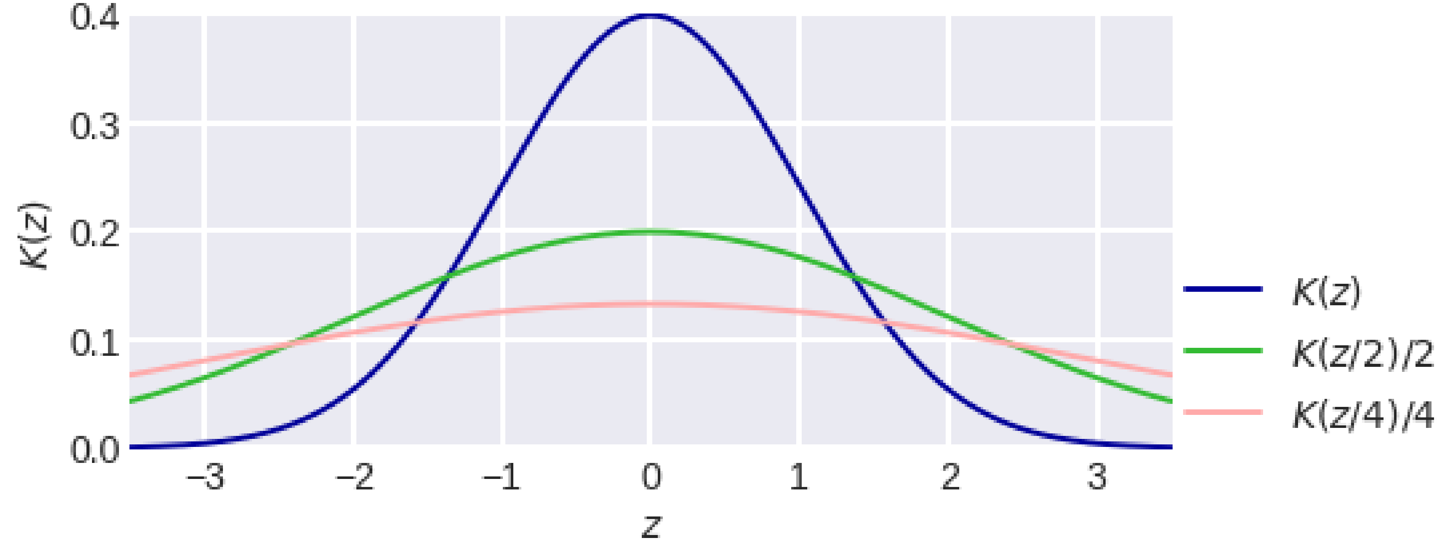
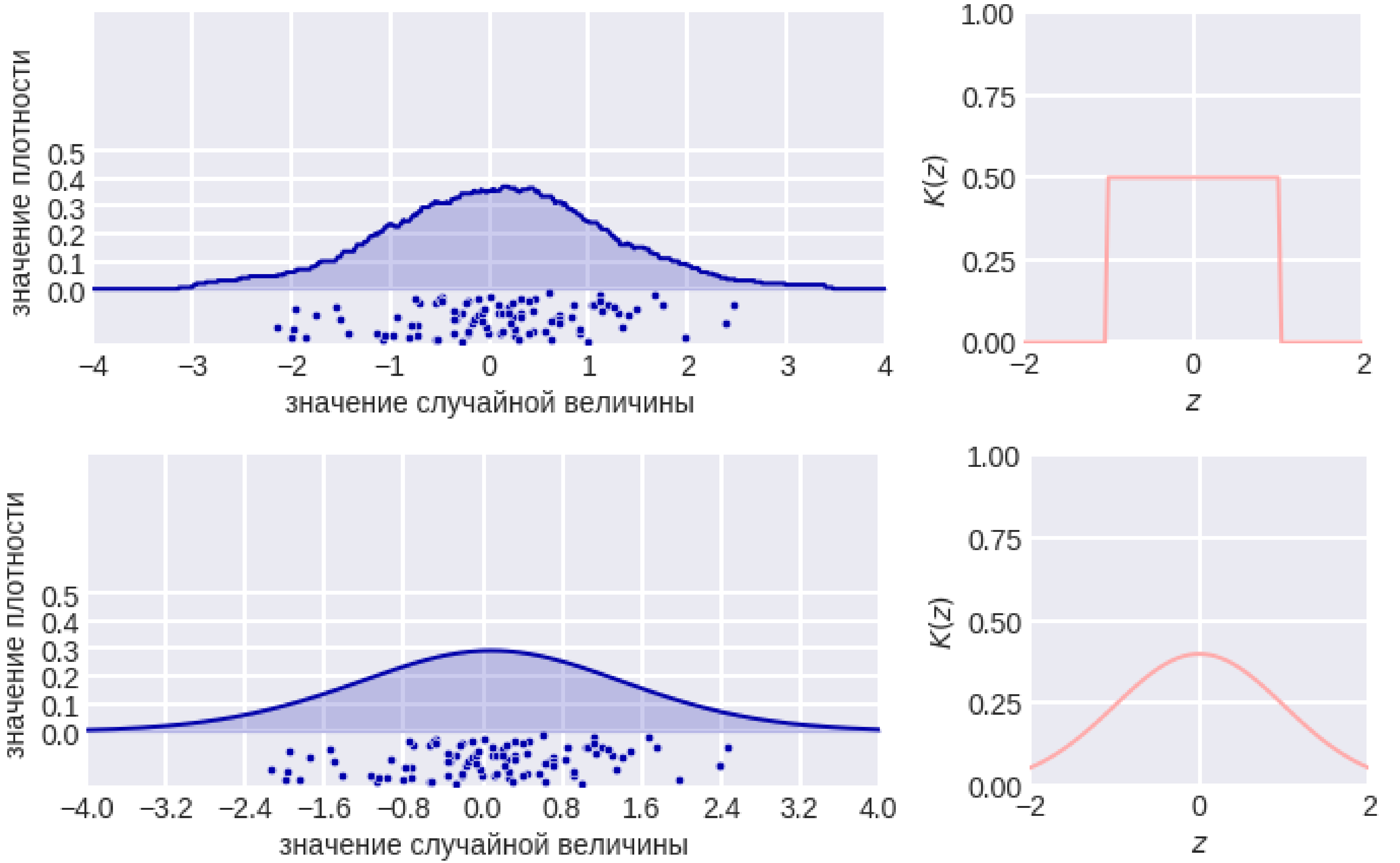


иллюстрация масштабирования

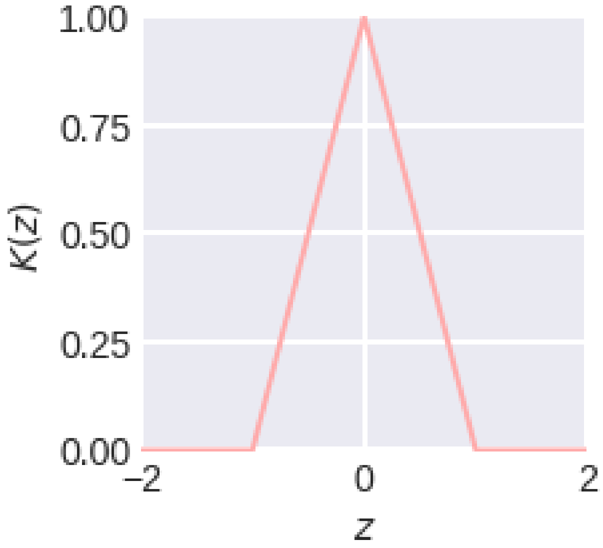
$$\frac{1}{h}K\left(\frac{z}{h}\right)$$

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

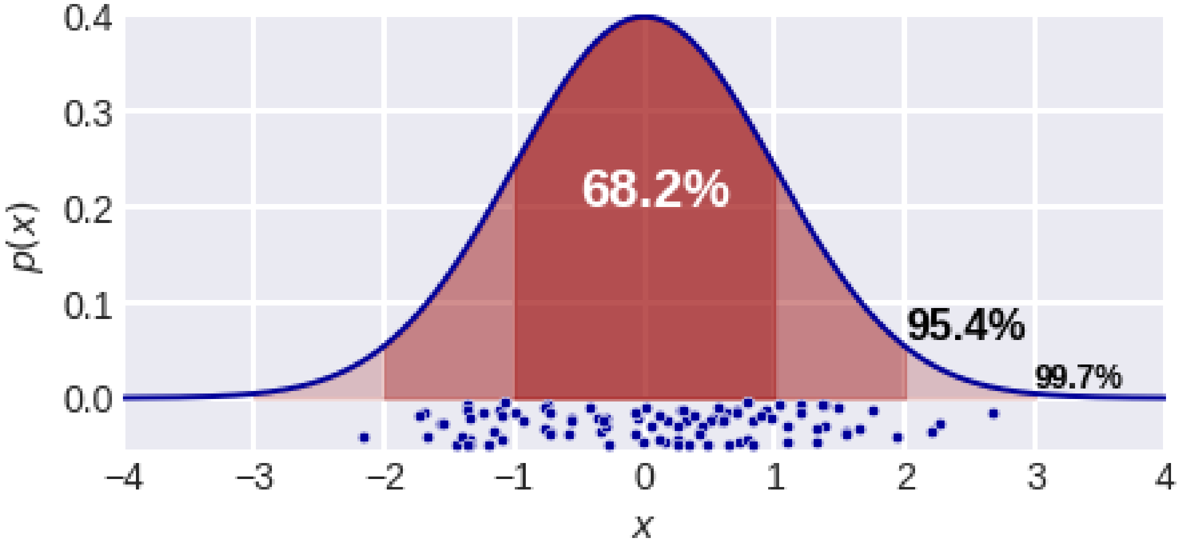
Дьяконов, А. Г. Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab (практикум на эвм кафедры математических методов прогнозирования). — МАКСПресс, 2010. — 278 с.
http://www.machinelearning.ru/wiki/images/7/7e/Dj2010up.pdf

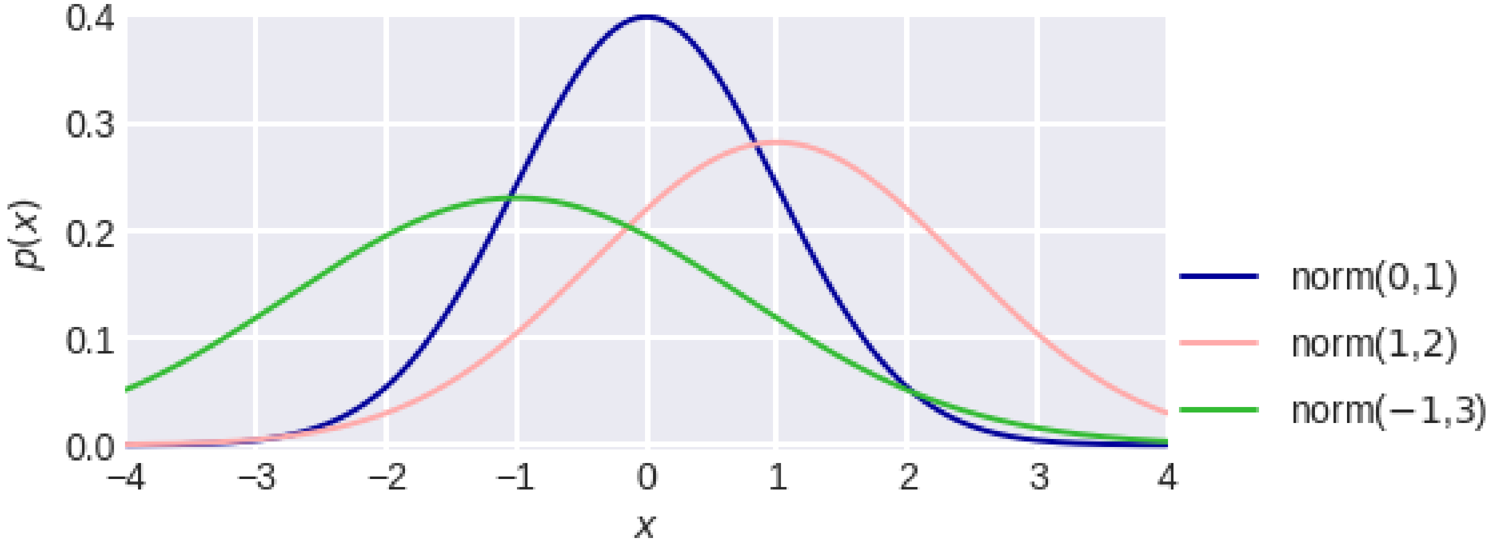
Дьяконов А.Г. Прогноз поведения клиентов супермаркетов с помощью весовых схем оценок вероятностей и плотностей // Бизнес-информатика. 2014. № 1 (27). С. 68–77
https://bijournal.hse.ru/data/2014/04/15/1320713004/8.pdf

Пример распределения – нормальное

Одномерное нормальное распределение

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





«около центра данных очень много»

Нет так полезна на практике, как в теории…

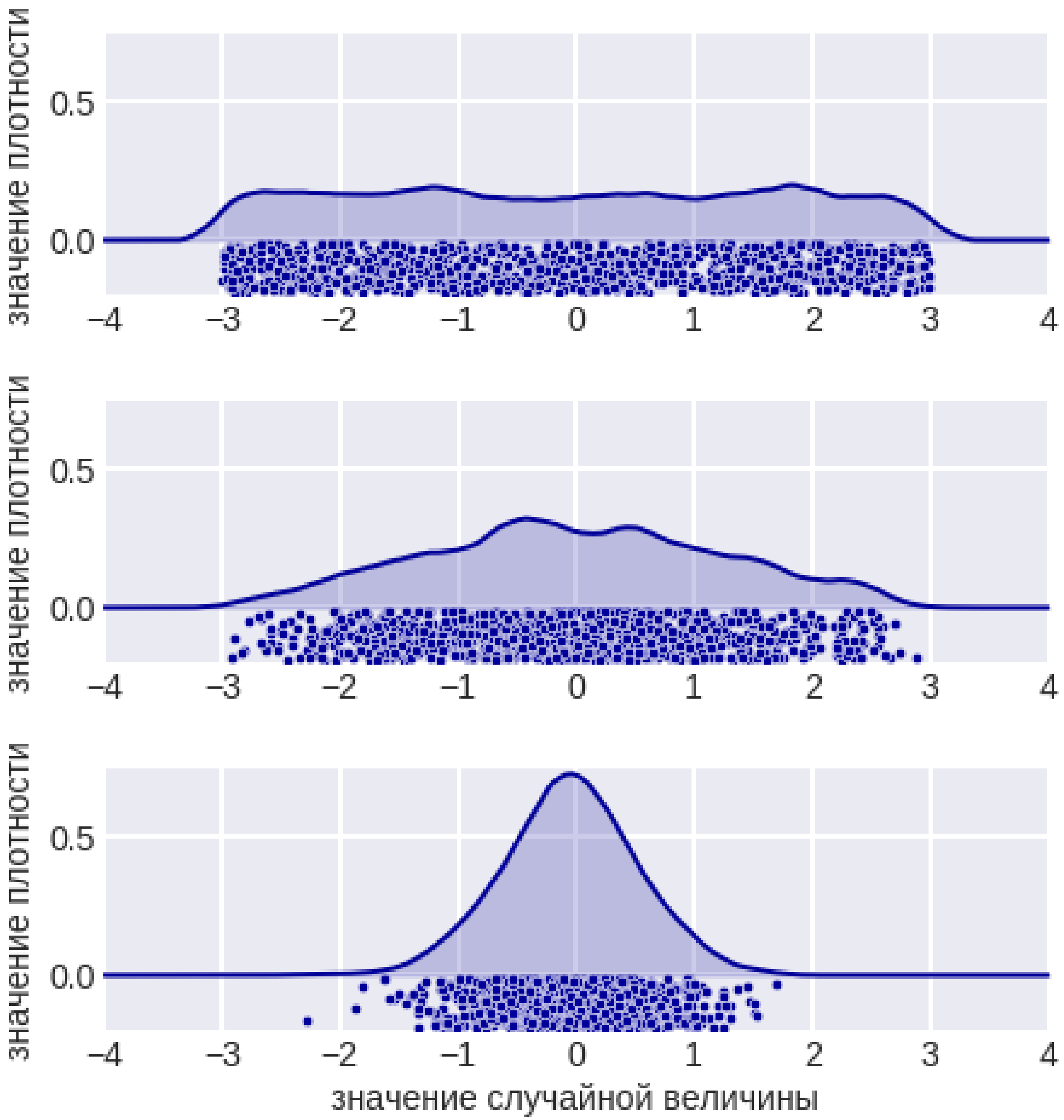
Центральная предельная теорема

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

$$\sqrt{m}\,\frac{\frac{1}{m}\sum_{i=1}^{m}\xi_i-\mu}{\sigma}\to\mathrm{norm}(0,1)$$ **по распределению.**

пример почему так важно нормальное распределение

Иллюстрация ЦПТ: плотности, оцененные по Парзену



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

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

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

похоже на нормальное?

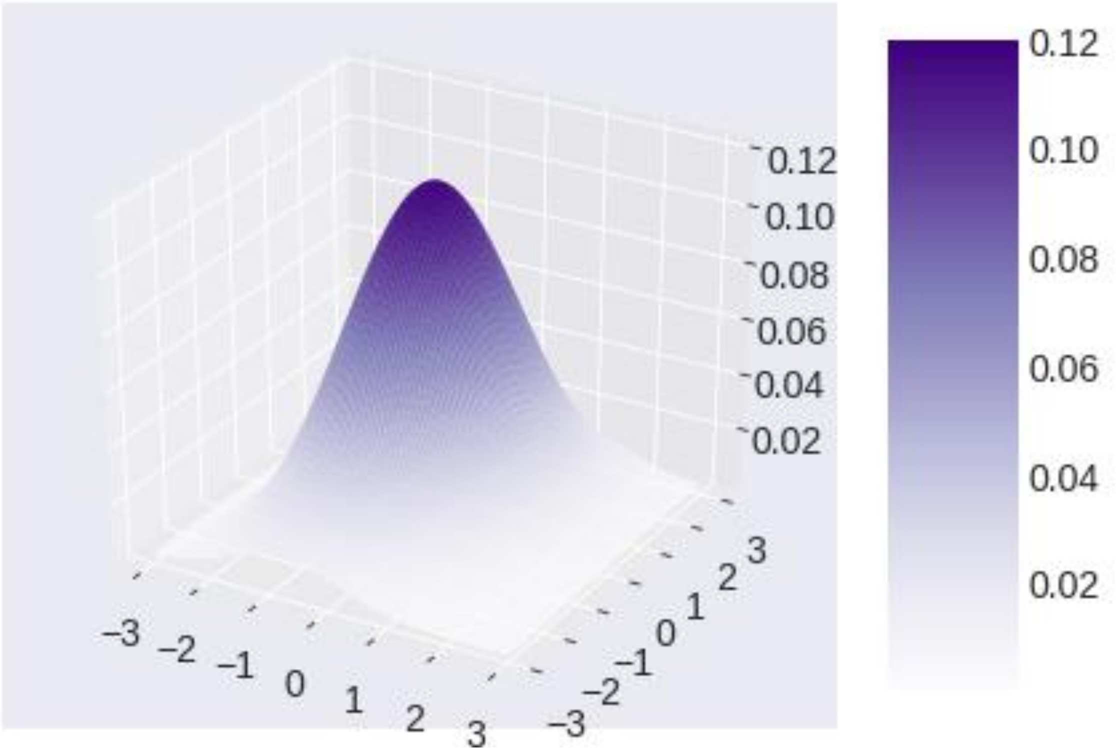
Многомерное нормальное (гауссовское) распределение

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

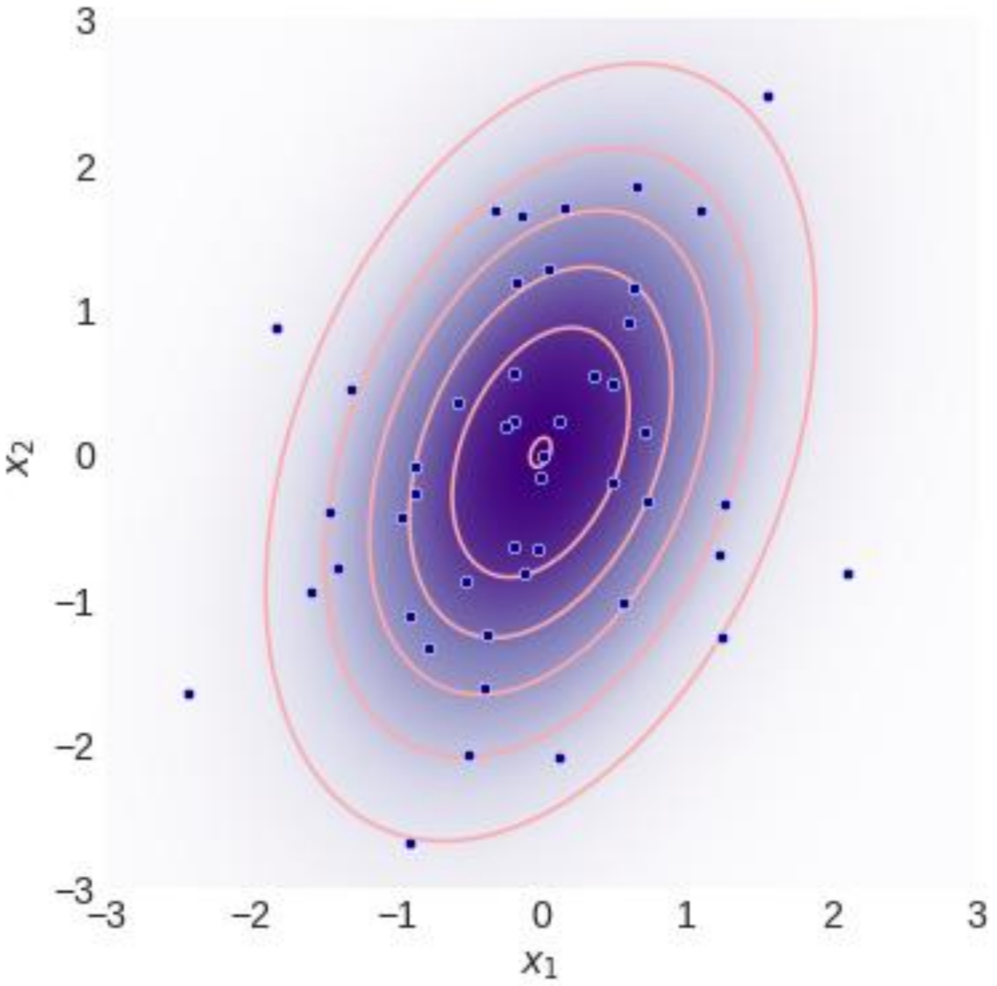
$$\mu=\mathbf{E}x$$

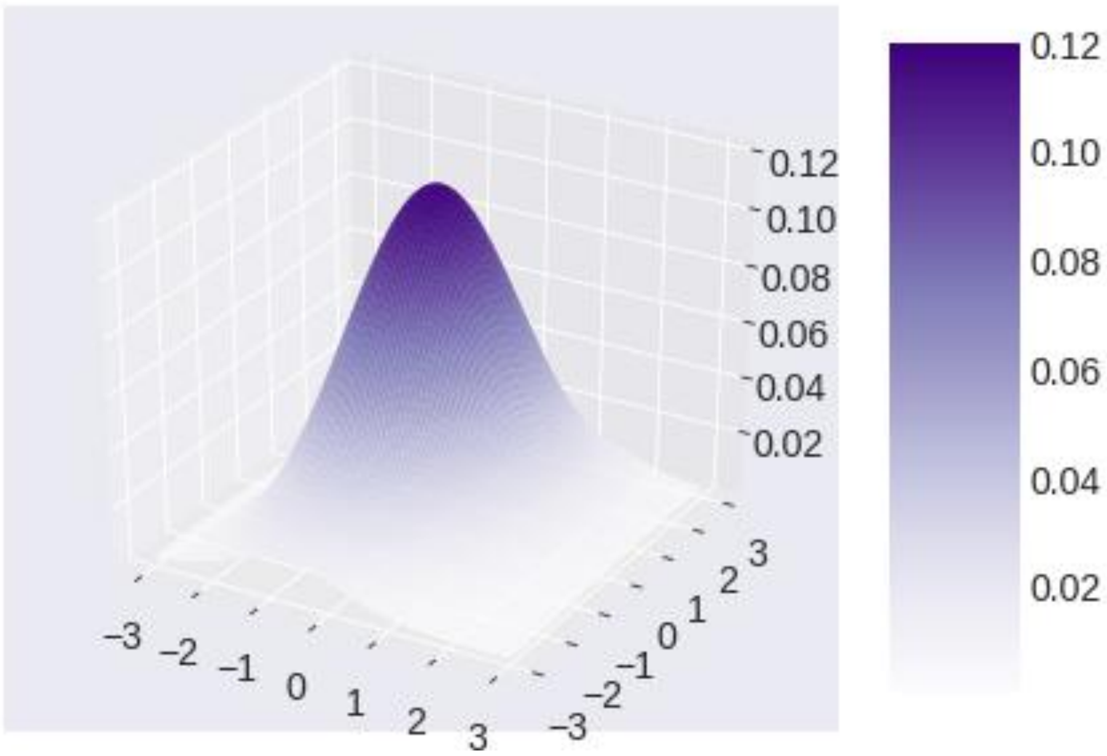
$$\Sigma=\mathrm{cov}(x)=$$

$$=\begin{bmatrix}\mathrm{var}(X_1) & \mathrm{cov}(X_1,X_2) & \ldots & \mathrm{cov}(X_1,X_n)\\ \mathrm{cov}(X_2,X_1) & \mathrm{var}(X_2) & \ldots & \mathrm{cov}(X_2,X_n)\\ \ldots & \ldots & \ldots & \ldots\\ \mathrm{cov}(X_n,X_1) & \mathrm{cov}(X_n,X_2) & \ldots & \mathrm{var}(X_n)\end{bmatrix}$$

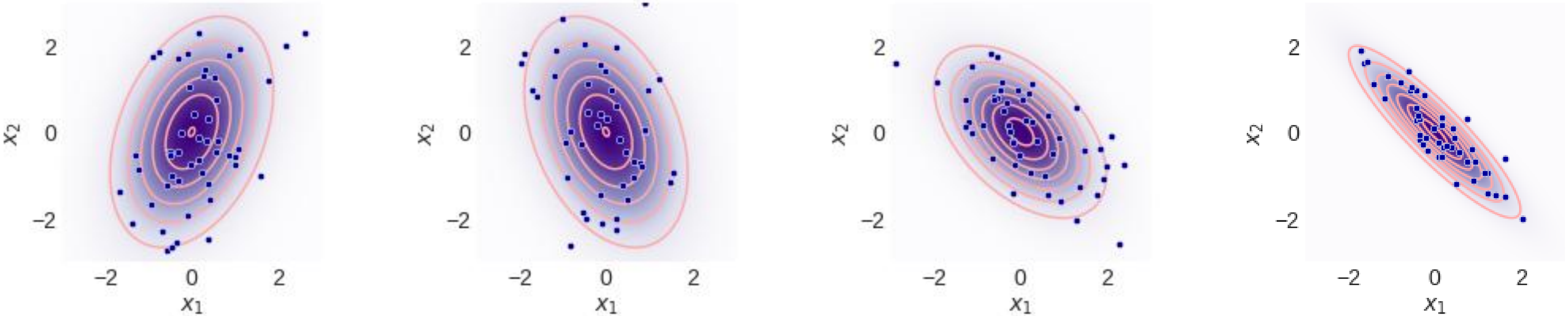


Многомерное нормальное (гауссовское) распределение





Многомерное нормальное (гауссовское) распределение



$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$

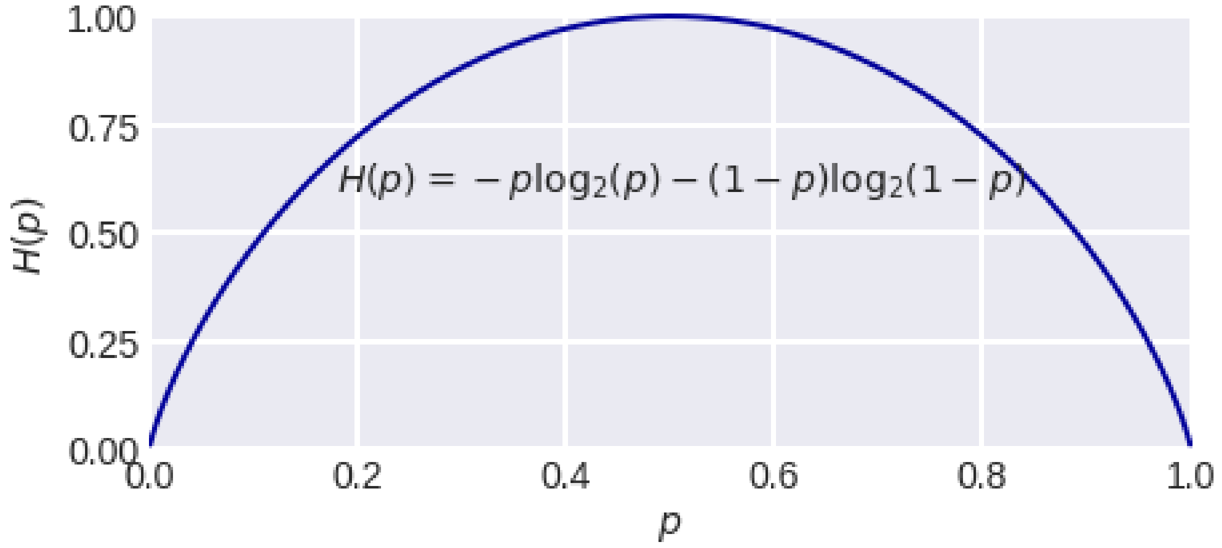
Теория информации

Информационная энтропия (Entropy) – **мера неопределённости некоторой системы**

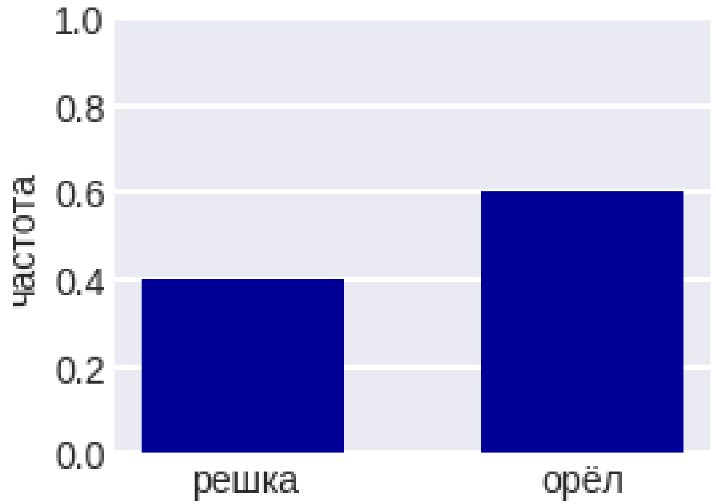
$$x \sim (x_1, p_1), (x_2, p_2), \ldots$$

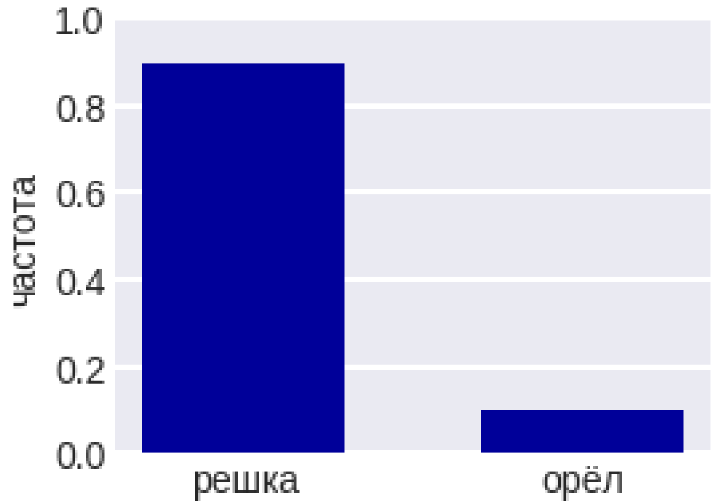
$$H(x) = -\sum_t p_t \log p_t \qquad H(p) = -p\log_2 p - (1-p)\log_2(1-p)$$

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 $\mathbb{R}^n$

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

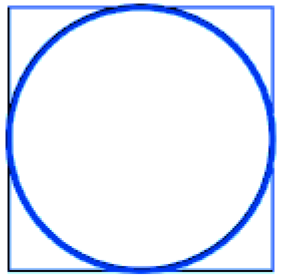
$$n = 1$$

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

100% от объёма описанного параллелипипеда

$$n = 2$$

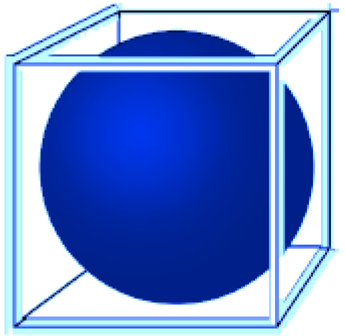
$$\mathrm{vol}(r) = \pi r^2$$



79%

$$n = 3$$

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



52%

$V_{шар} / V_{пар} \to 0$

Проклятие размерности (The Curse of Dimensionality)

Весь объём сосредоточен «на краю» шара

$$\frac{\mathrm{vol}(r+\varepsilon)}{\mathrm{vol}(r)} = \left(1+\frac{\varepsilon}{r}\right)^n \xrightarrow[n\to+\infty]{} +\infty$$

слой пыли – самая большая составляющая в запылённом многомерном шаре

скорее всего, соседи будут с краю...

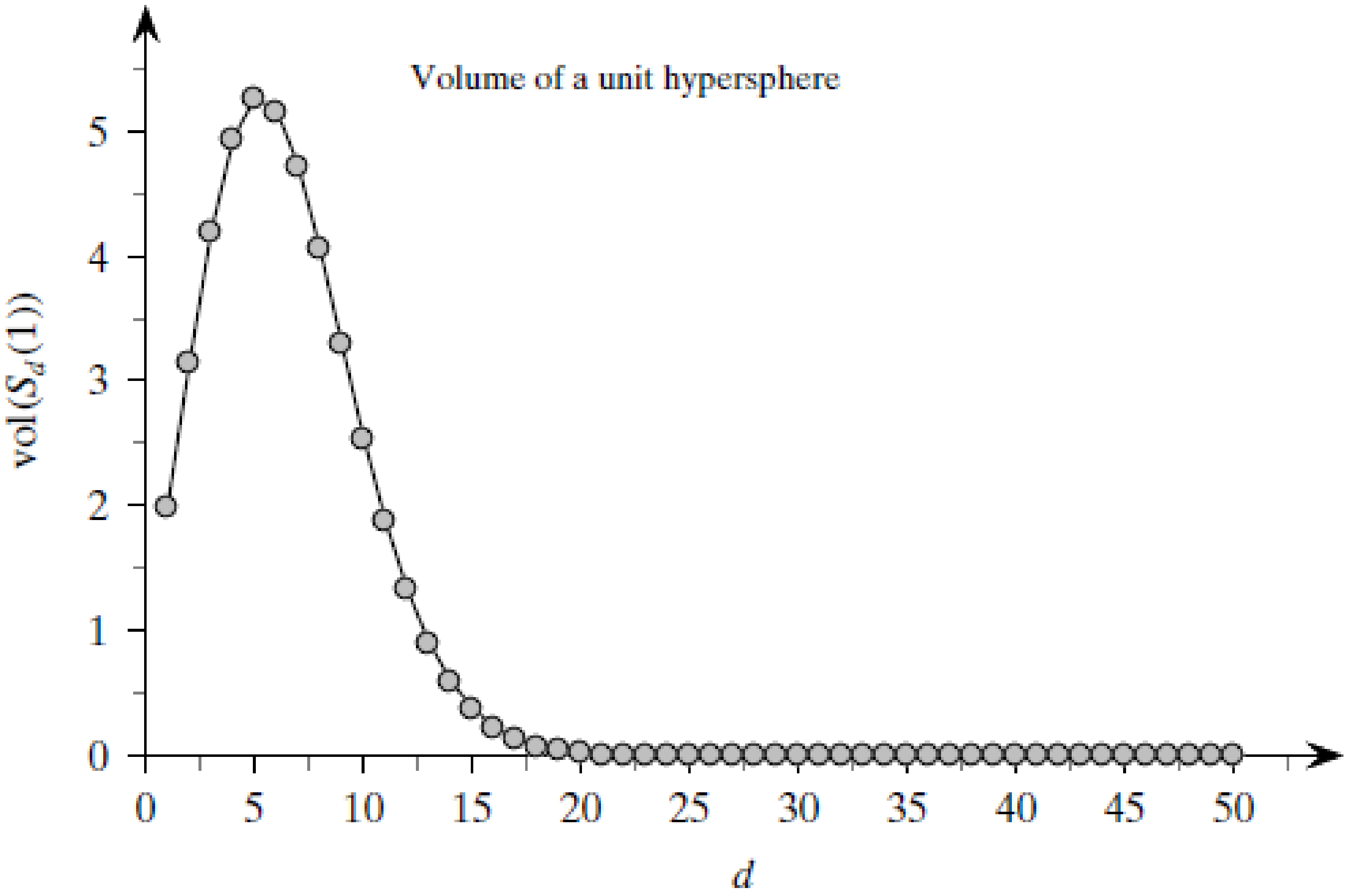
но это в предположении, что объекты «равномерно» разбросаны по пространству, а в реальности лежат около поверхностей малых размерностей

ДЗ доказать перечисленные факты

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

Проклятие размерности (The Curse of Dimensionality)

Объём единичного шара при росте размерности



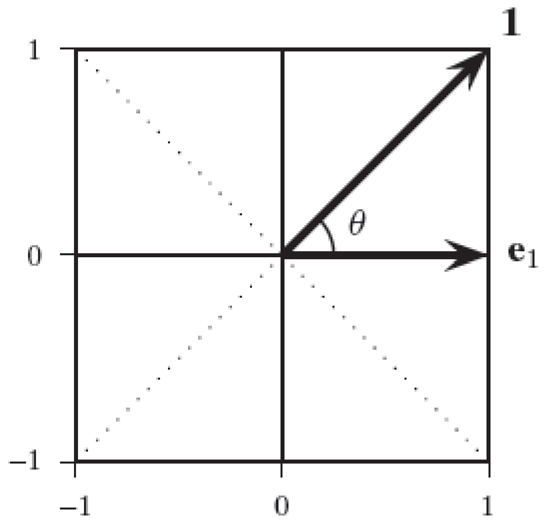
нет соответствия интуиции

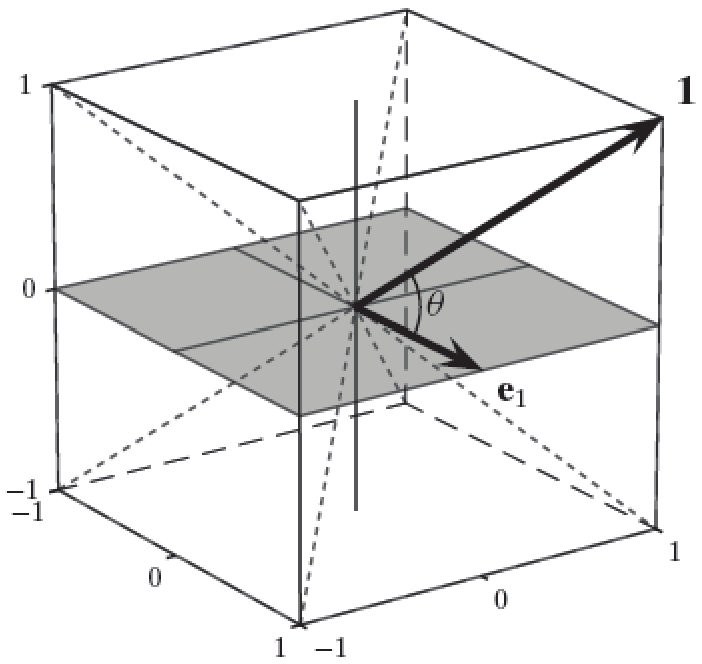
Проклятие размерности (The Curse of Dimensionality)

Угол между диагональю и первым базисным вектором

$$\cos\theta = \frac{(1,0,\ldots,0)^{\mathrm{T}}\cdot(1,\ldots,1)}{\|(1,0,\ldots,0)\|\cdot\|(1,\ldots,1)\|} = \frac{1}{\sqrt{n}} \to 0$$

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





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

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

но при росте размерности «по сути» они отличаются

Проклятие размерности (The Curse of Dimensionality)

необходимость больших данных с ростом числа признаков

например, в оценке плотности (отсюда разные наивные Байесы…)

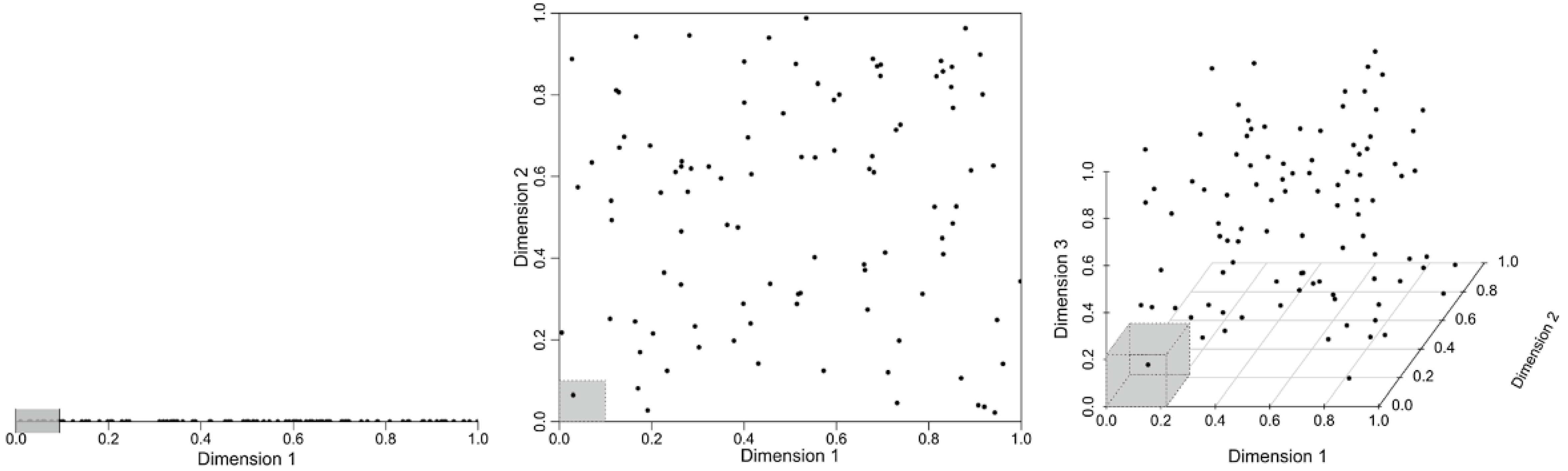


Fig. 2 The curse of dimensionality. Adding dimensions stretches the points apart making high-dimensional data extremely sparse and uniformly distributed

Debie, E., Shafi, K. Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses. Pattern Anal Applic 22, 519–536 (2019). https://doi.org/10.1007/s10044-017-0649-0

Сингулярное разложение матрицы (SVD)

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

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

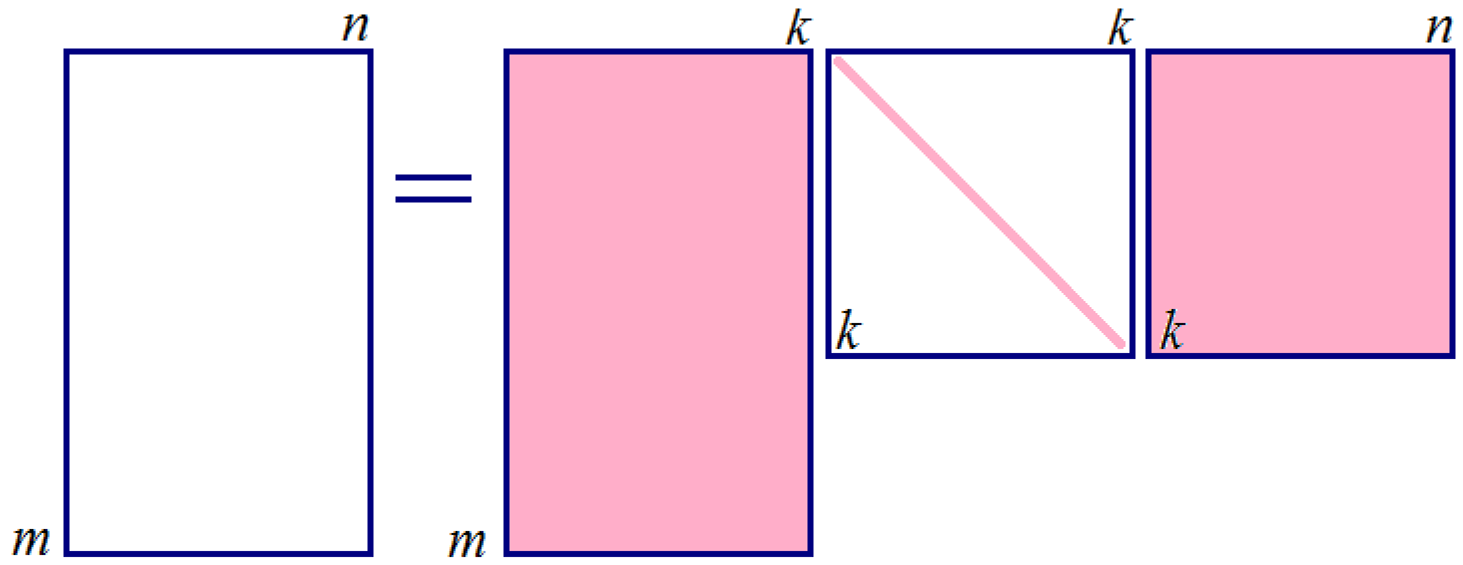
$$X = \sum_{i=1}^{k}\lambda_i u_i v_i^{\mathrm{T}}$$

где $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_k)$

$$\lambda_1 \geq \ldots \geq \lambda_k > 0$$

$$U^{\mathrm{T}}U = I$$

$$V^{\mathrm{T}}V = I$$



Сингулярное разложение матрицы (SVD)

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

поэтому

$$X^{\mathrm{T}}XV = V\Lambda^2$$

и матрица V **состоит из с.в. матрицы** $X^{\mathrm{T}}X$**,**

которым соответствуют с.з. $\lambda_1^2 \geq \ldots \geq \lambda_k^2 > 0$

$\lambda_1 \geq \ldots \geq \lambda_k > 0$ **– сингулярные числа**

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

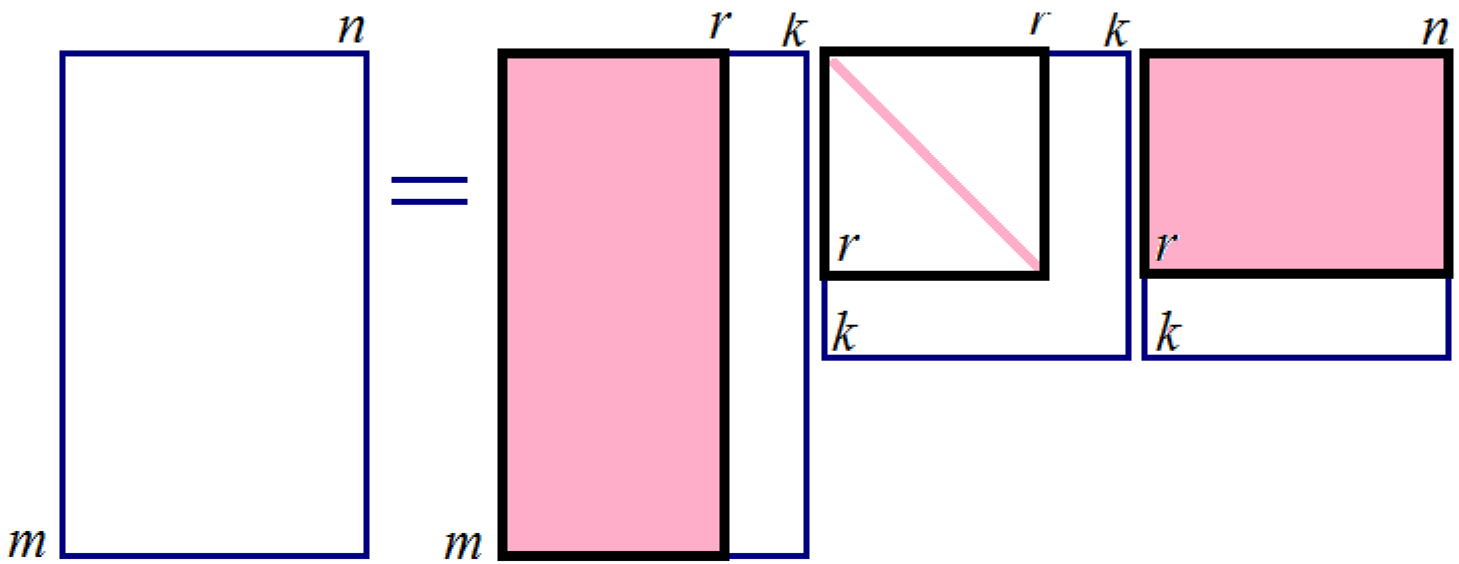
Усечённое сингулярное разложение матрицы (Truncated SVD)

что будет если

$$X = U\Lambda V^{\mathrm{T}}$$

$$\Lambda' = \mathrm{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$$

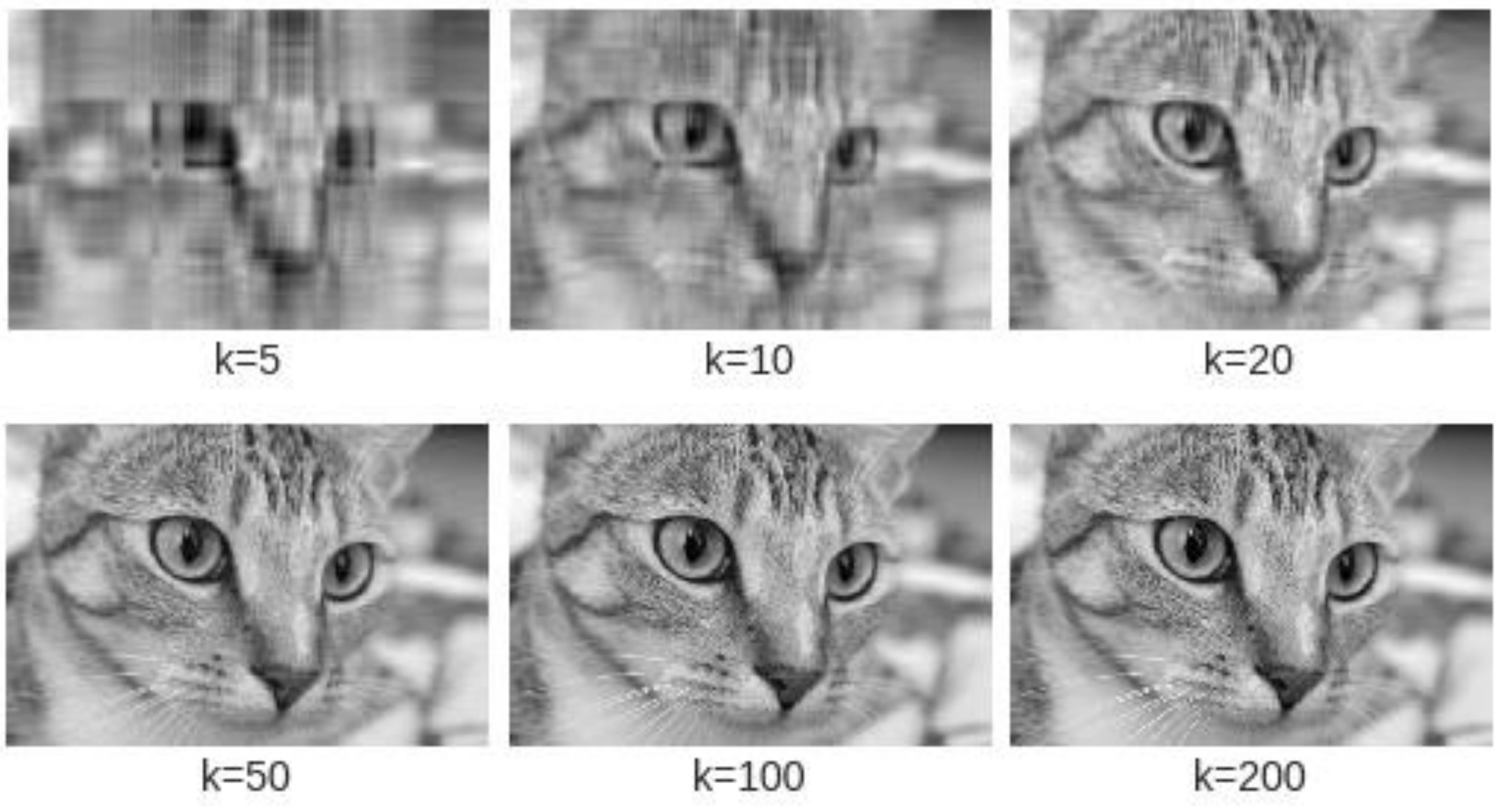
$$U\Lambda V^{\mathrm{T}} = \sum_{i=1}^{r} \lambda_i u_i v_i^{\mathrm{T}} \equiv X' = \arg\min_{H:\mathrm{rank}\, H = r} \| X - H \|_2^2 = \sum_{i=r+1}^{k} \lambda_i^2$$



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 и сжатие изображений с помощью SVD



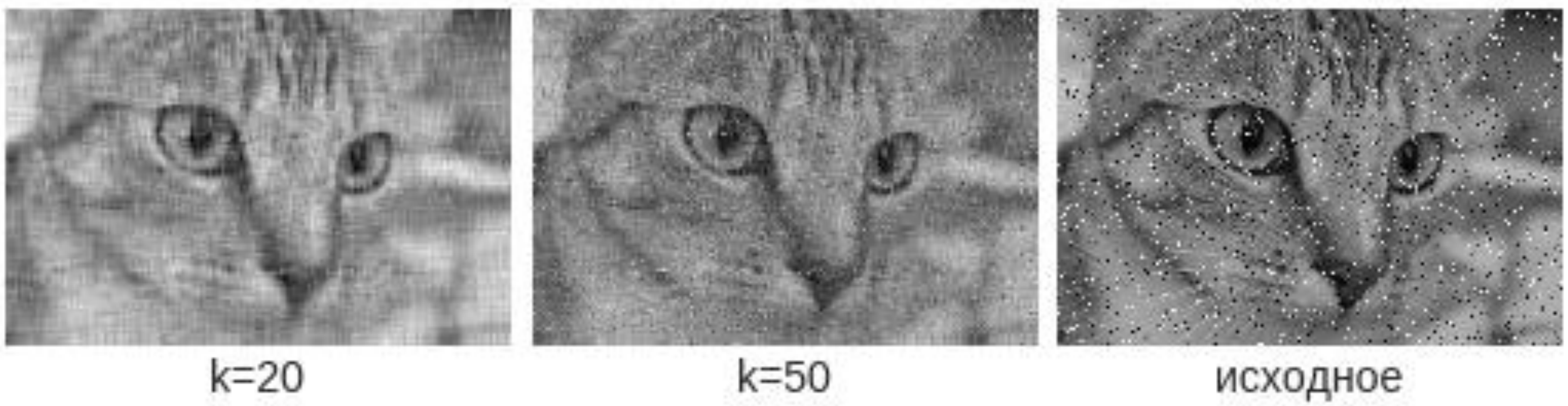
Изначальный размер изображения 300×451 = 135 300

300×50 + 50 + 50×451 = 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



Матричное дифференцирование

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

$$\frac{\partial(A_1W_1 + \ldots + A_nW_n)}{\partial(W_1,\ldots,W_n)} = \begin{bmatrix} \dfrac{\partial(A_1W_1 + \ldots + A_nW_n)}{\partial W_1} \\ \vdots \\ \dfrac{\partial(A_1W_1 + \ldots + A_nW_n)}{\partial W_n} \end{bmatrix} = \begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix}$$

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