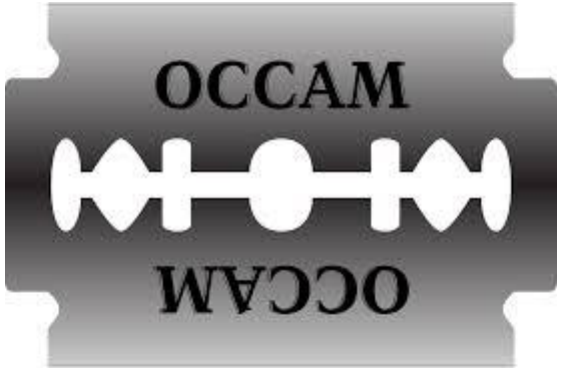
Машинное обучение и анализ данных

Математика в машинном обучении: краткий обзор

Дьяконов А.Г.

Московский государственный университет
имени М.В. Ломоносова (Москва, Россия)

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… Энштейн

Теорема о бесплатном сыре (No Free Lunch Theorem)

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

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?

Сведения из ТВиМС

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

Есть и другой подход к определению 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

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

удобна тем, что

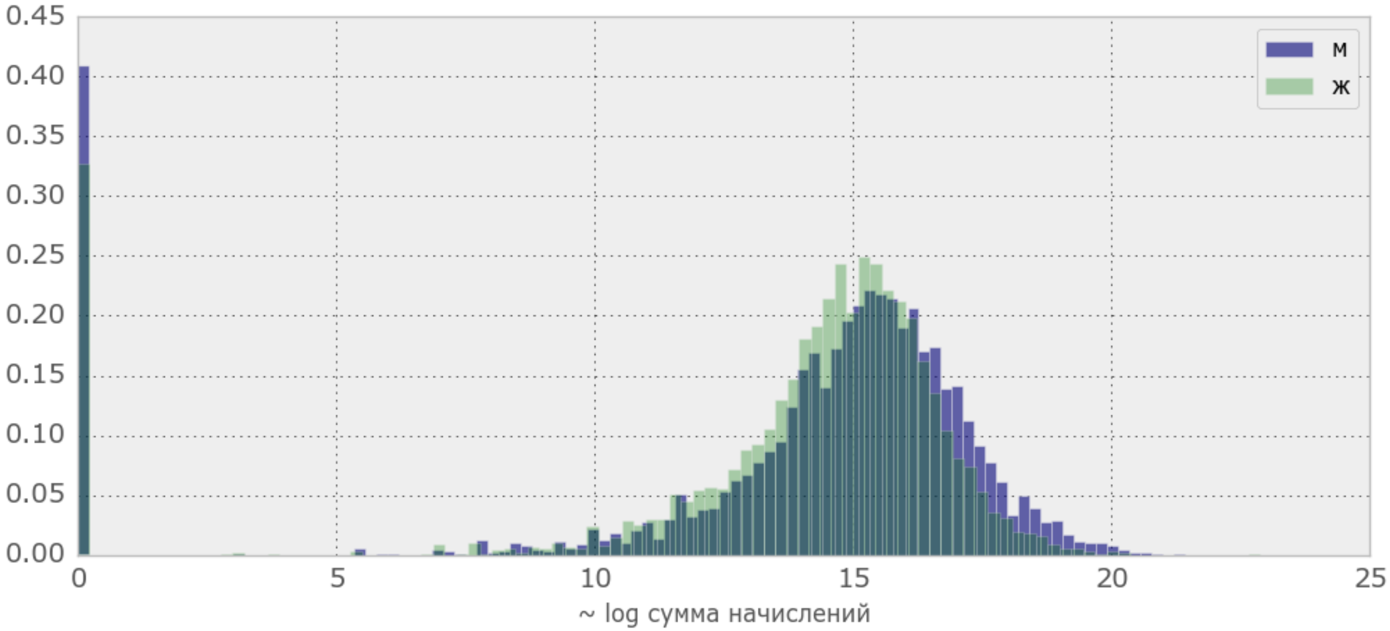
$$\mathbf{P}(a \le \xi \le b) = \int_{a}^{b} p(x)\partial x$$

Связь плотности и вероятности

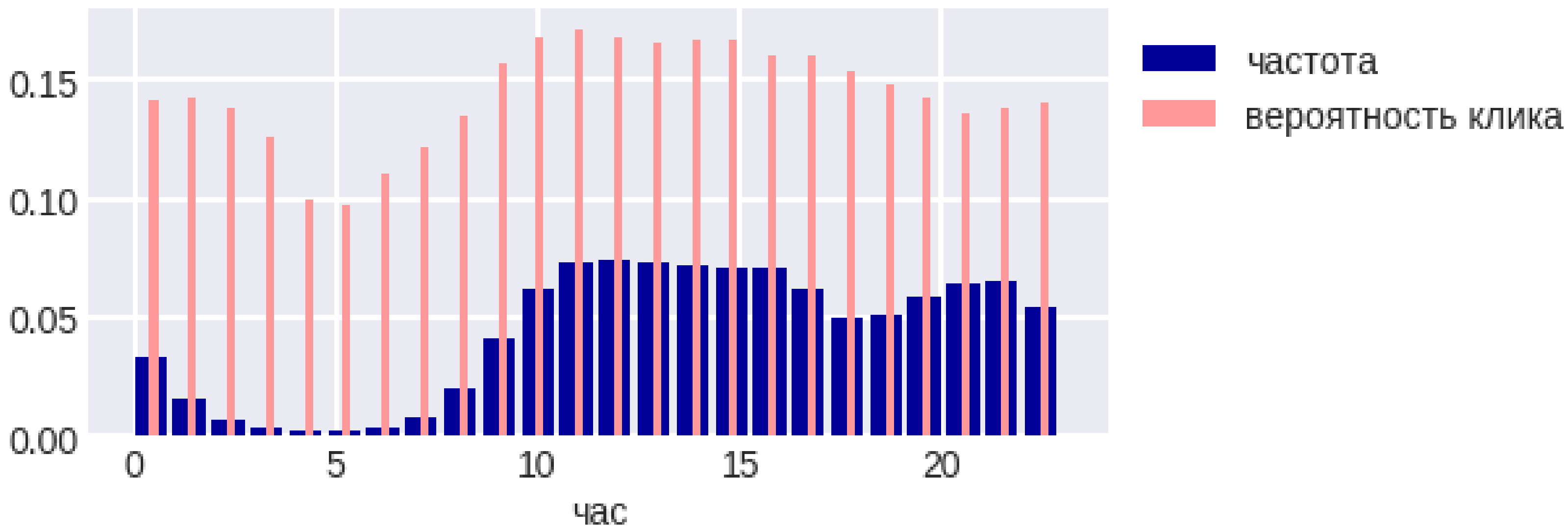
$$\mathbf{P}(x-\varepsilon \le \xi \le x+\varepsilon) = \int\limits_{x-\varepsilon}^{x+\varepsilon} p(z)\partial z \approx 2\varepsilon\, p(x)$$

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

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 –

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

Очевидный пересчёт

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

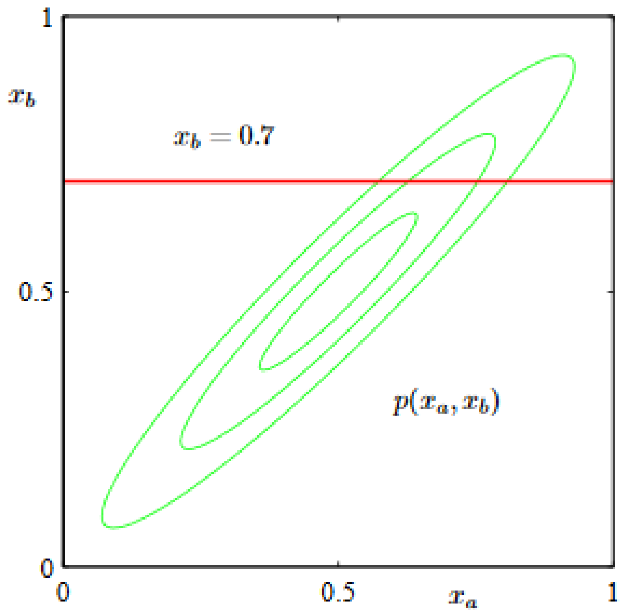
Сведения из ТВиМС

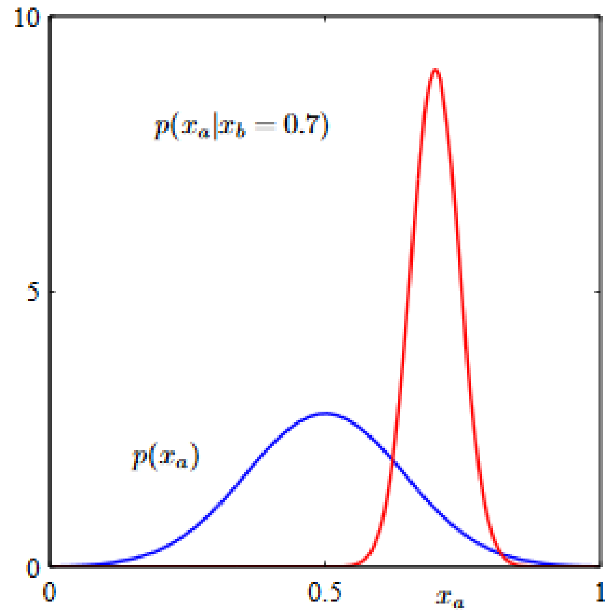
Маргинализация плотности по неизвестной компоненте

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

Обуславливание плотности по известной компоненте

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





Правило произведения

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

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

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) – функция от выборки

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

примеры

– это тоже случайная величина

Требования к статистике

1) Надо, чтобы значение было 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 $\mathrm{bias}(\hat{\theta}) \to 0$

Состоятельность (Consistency) $\mathbf{P}(\hat{\theta} = \theta) \to 1$ **при** $m \to \infty$

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

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

Пример:

$$\mathrm{Var}(\hat{\mu}) = \frac{\sigma}{m}$$

Требования к статистике

3) с ростом числа наблюдений была сходимость

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

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

Оценка MLE (ММП)

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

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

1) состоятельная

2) эффективная (грубо: из всех состоятельных самая лучшая)

Эффективность (Efficiency) ~ самая точная, которую можно получить для выборки этого размера

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:

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

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

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

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

Дивергенция Кульбака-Лейблера

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

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...

Ковариация и корреляция

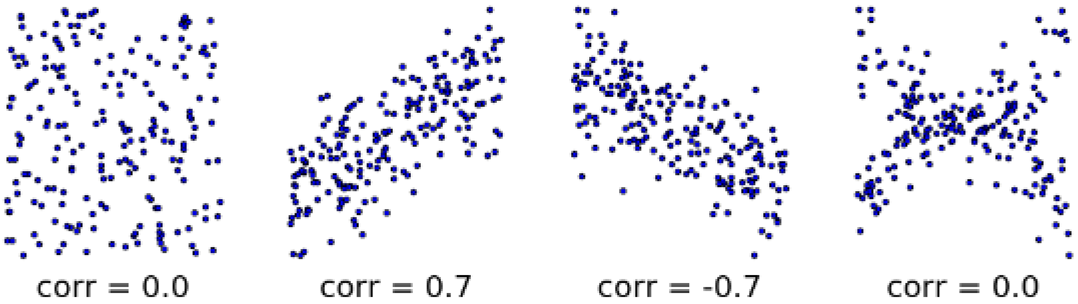
$$\mathrm{cov}(X,Y)=\mathbf{E}[(X-\mathbf{E}X)(Y-\mathbf{E}Y)]$$

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

Корреляционный коэффициент Пирсона:

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

определяет меру линейной зависимости между с.в.

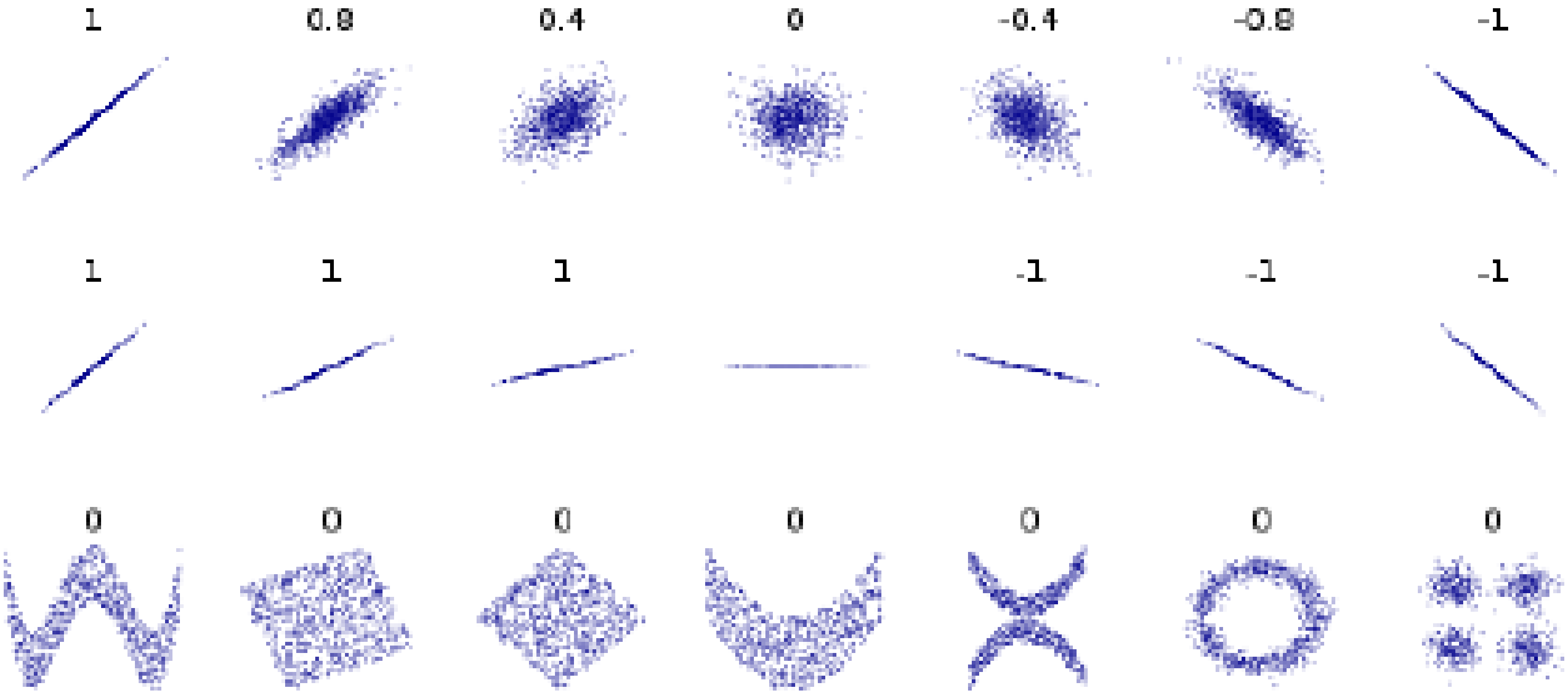


var = D

Независимые переменные некоррелированны.

Обратное неверно.

Ковариация и корреляция



Ковариация и корреляция

Коэффициент корреляции Спирмена – определяет меру монотонной зависимости

= коэффициент корреляции Пирсона между рангами

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

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

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

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

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

EM-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:

$$K(z)\geq 0$$

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

Примеры:

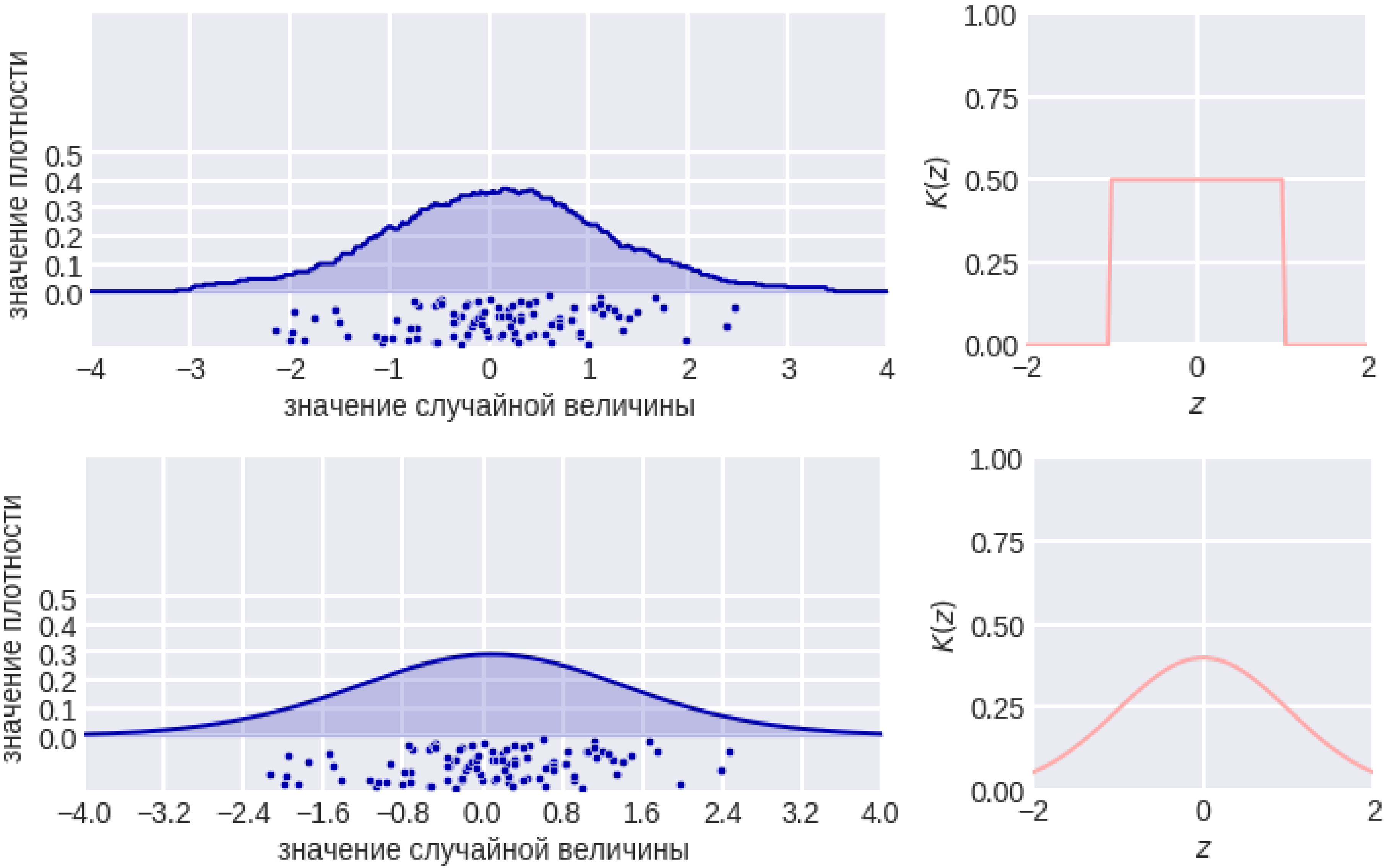
$$K(z)=\frac{1}{2}I[|z|\leq 1]$$

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

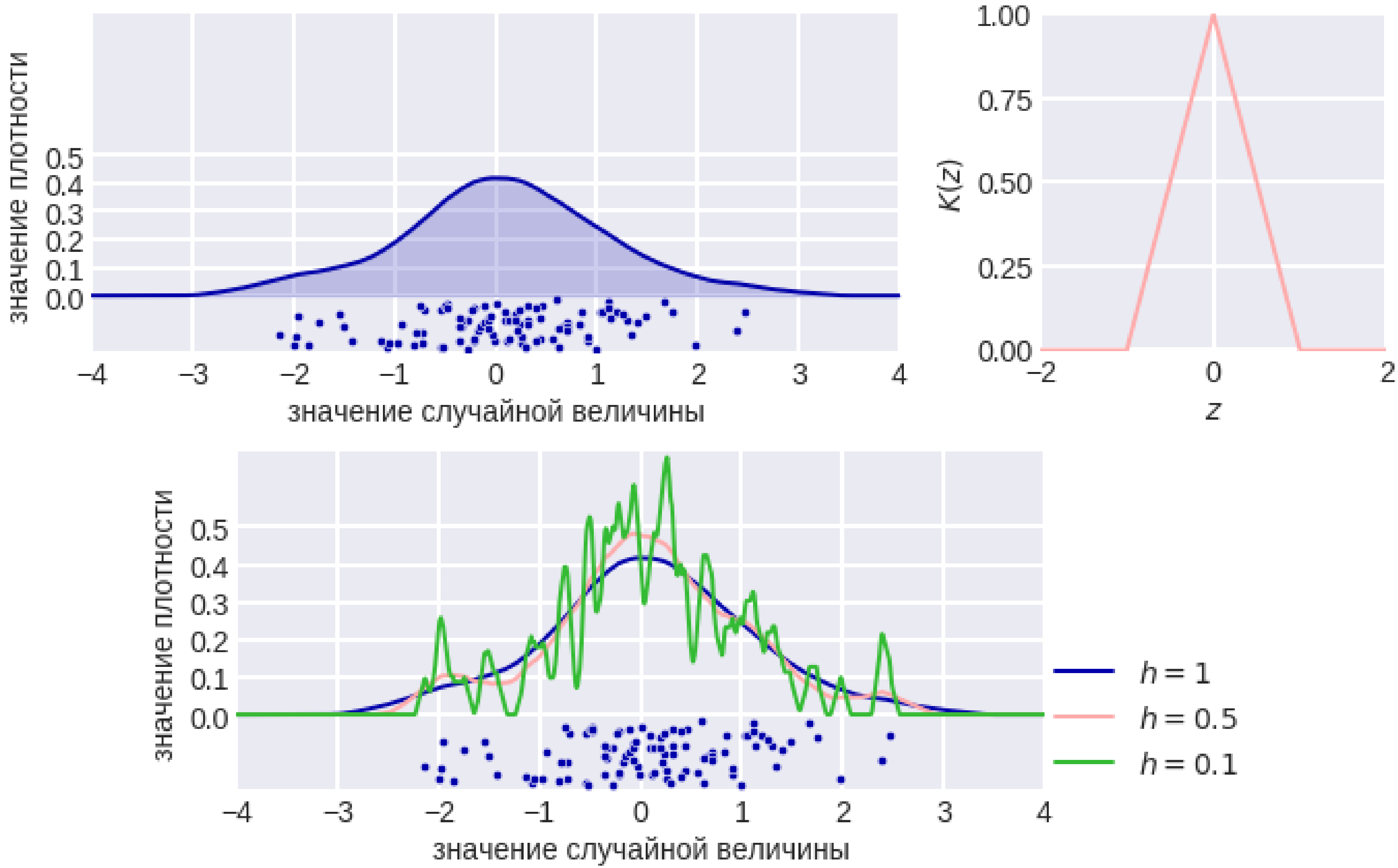
$$K(z)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^{\mathrm{T}}z}{2}\right)$$

и в многомерном случае

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))
```

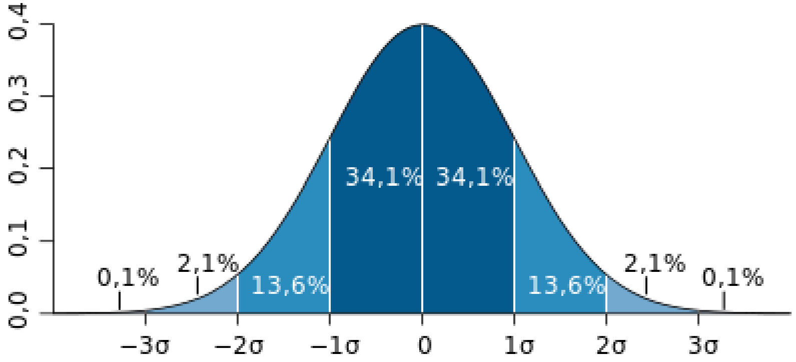
Парзеновский подход

Есть теорема о сходимости

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



«около центра данных очень много»

Нет так полезна на практике, как в теории…

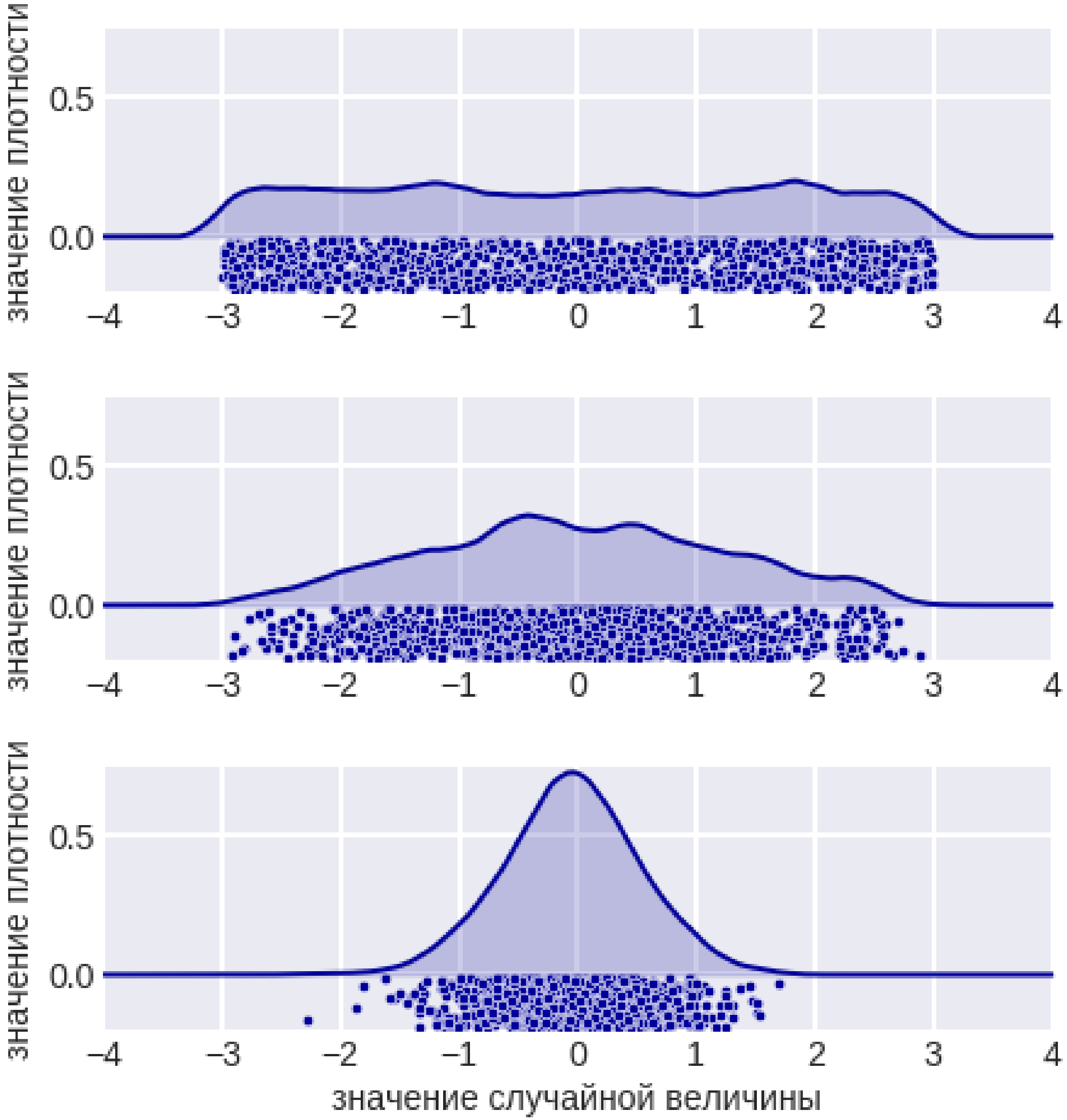
https://ru.wikipedia.org/wiki/%D0%A1%D1%80%D0%B5%D0%B4%D0%BD%D0%B5%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%BE%D0%B5_%D0%BE%D1%82%D0%BA%D0%BB%D0%BE%D0%BD%D0%B5%D0%BD%D0%B8%D0%B5

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**

Иллюстрация ЦПТ



Плотности оценены по Парзену

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

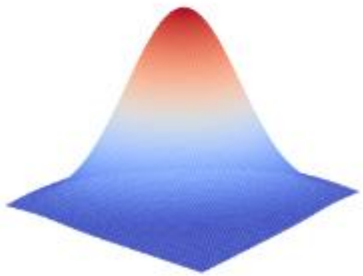
$$\frac{\xi_1 + \xi_2}{2}$$

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

Многомерное нормальное (гауссовское) распределение

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

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



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

Теория информации

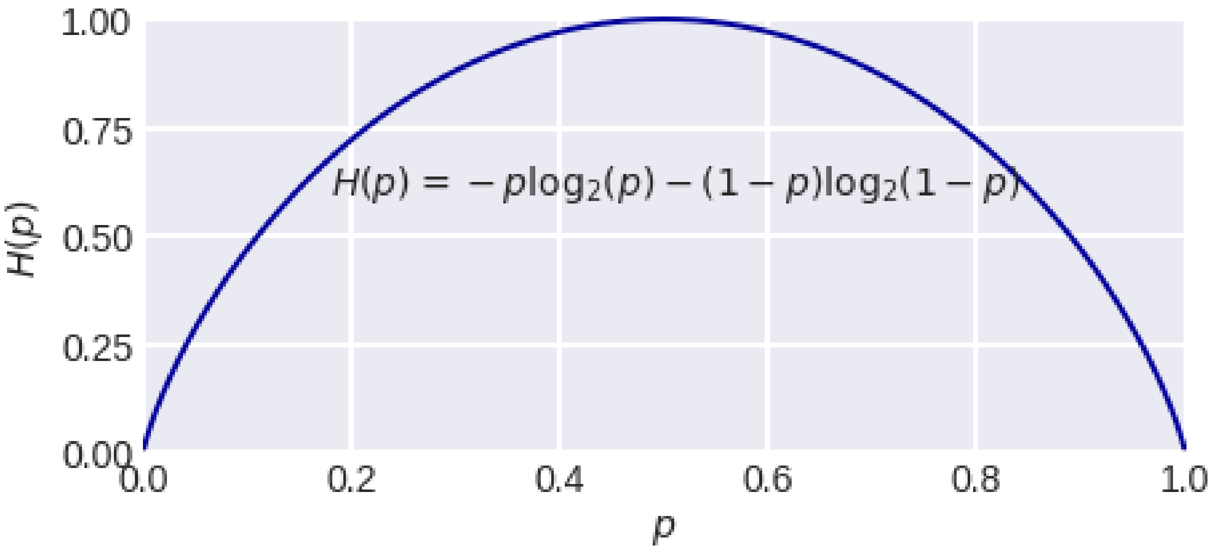
Информационная энтропия (Entropy) – мера неопределённости некоторой системы

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

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

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

Что зависит от основания логарифма?



Проклятие размерности

Объём шара радиуса r в $\mathbb{R}^n$

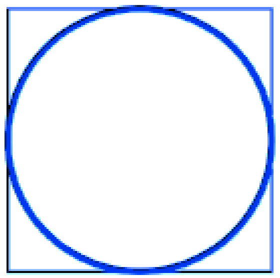
$$\mathrm{vol}(r) = \frac{\pi^{n/2}}{\Gamma(n/2+1)} r^n$$

$$n = 1$$

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

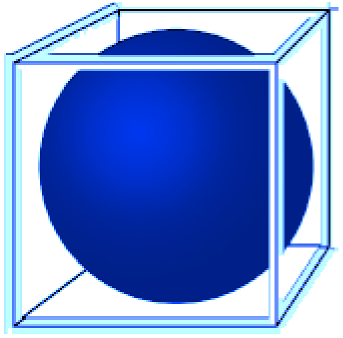
$$n = 2$$

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



$$n = 3$$

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



ДЗ доказать

Проклятие размерности (The Curse of Dimensionality)

Весь объём сосредоточен «на краю» шара

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

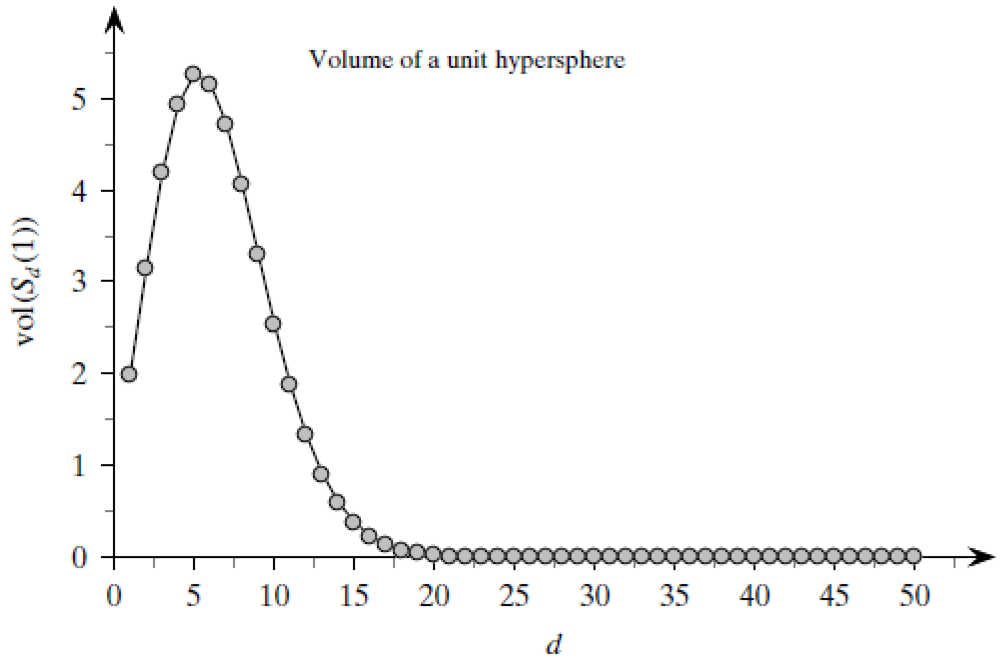
100% → 79% → 52% → … → 0

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



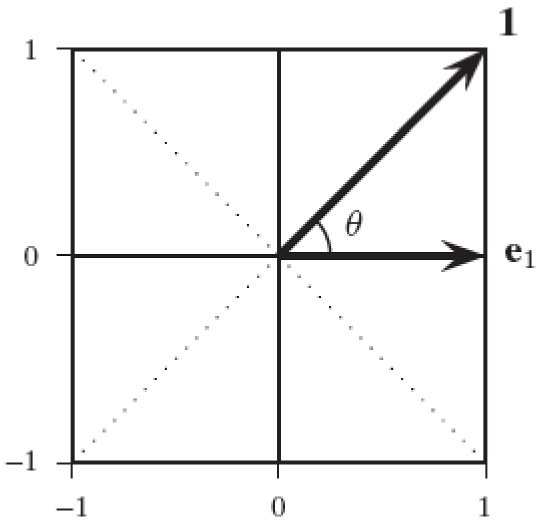
нет соответствия интуиции

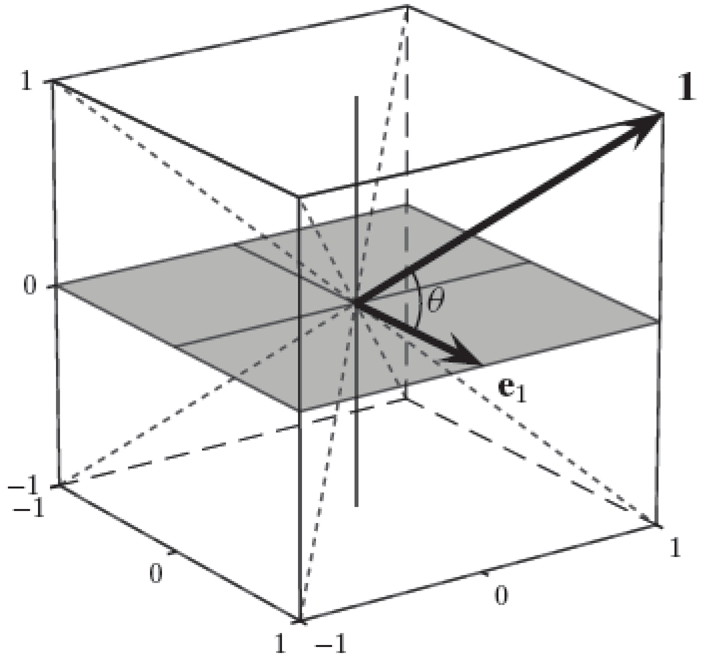
Проклятие размерности (The Curse of Dimensionality)

Угол между диагональю и первым базисным вектором

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

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





Проклятие размерности (The Curse of Dimensionality)

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

Д3 обосновать

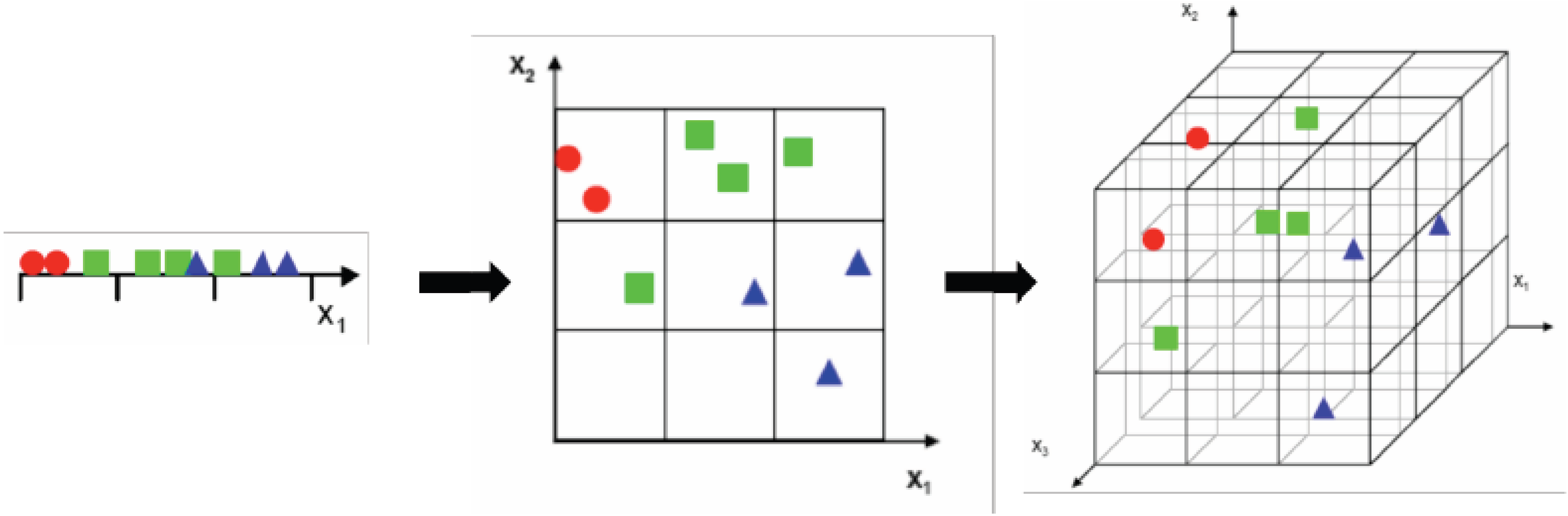
Как следствие из результата о нормах, метрики в конечномерном пространстве эквивалентны

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

но при росте размерности «по сути» они отличаются

Проклятие размерности (The Curse of Dimensionality)

необходимость больших данных с ростом числа признаков



Сингулярное разложение матрицы (SVD)

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

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

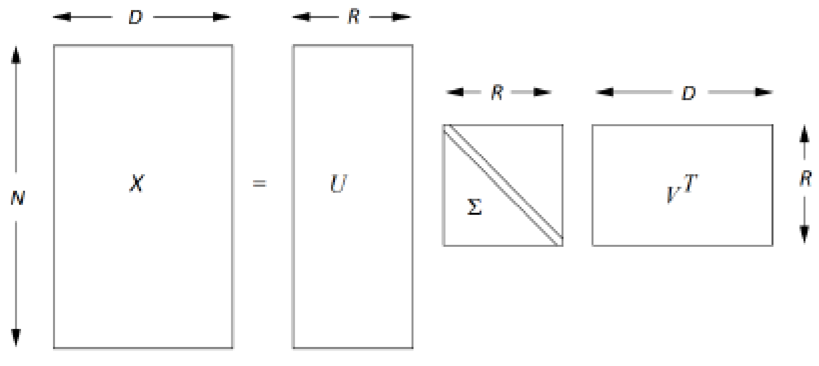
где $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$

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

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

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

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



Сингулярное разложение матрицы (SVD)

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

поэтому

$$X^{\mathrm{т}}XV = V\Lambda^2$$

и матрица V **состоит из с.в. матрицы** $X^{\mathrm{т}}X$**,**
которым соответствуют с.з. $\lambda_1^2 \geq \ldots \geq \lambda_k^2 > 0$
– сингулярные числа

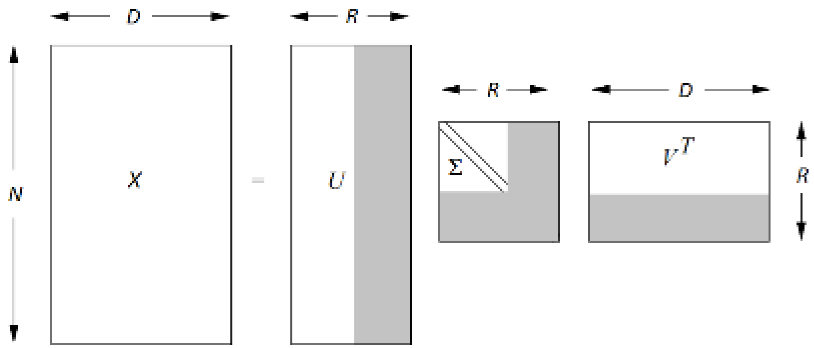
аналогично матрица U **состоит из с.в. матрицы** $XX^{\mathrm{т}}$
с теми же с.з.

Усечённое сингулярное разложение матрицы (Truncated SVD)

что будет если

$$X = U\Lambda V^{\mathrm{T}}$$
$$\Lambda' = \mathrm{diag}(\lambda_1, \ldots, \lambda_r, 0, \ldots, 0)$$

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



$$X' = \underset{H:\mathrm{rank}\, H = r}{\arg\min} \| X - H \|_2^2 = \sum_{i=r+1}^{k} \lambda_i^2$$

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 300×451 = 135 300

300×50 + 50×451 = 37 550

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

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

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