«Прикладные задачи анализа данных»

Оценки среднего, вероятности и плотности; весовые схемы

Александр Дьяконов

(ВМК МГУ имени М.В. Ломоносова)

7-8 ноября 2019 года

План лекции

Понятие «среднее»

- разные формализации
- полюсы / минусы
- практика

Оценка вероятности как среднего

case: некорректности при вычислении вероятности

Что такое среднее?

средний, типичный, среднестатистический…

Естественная формализация – среднее арифметическое

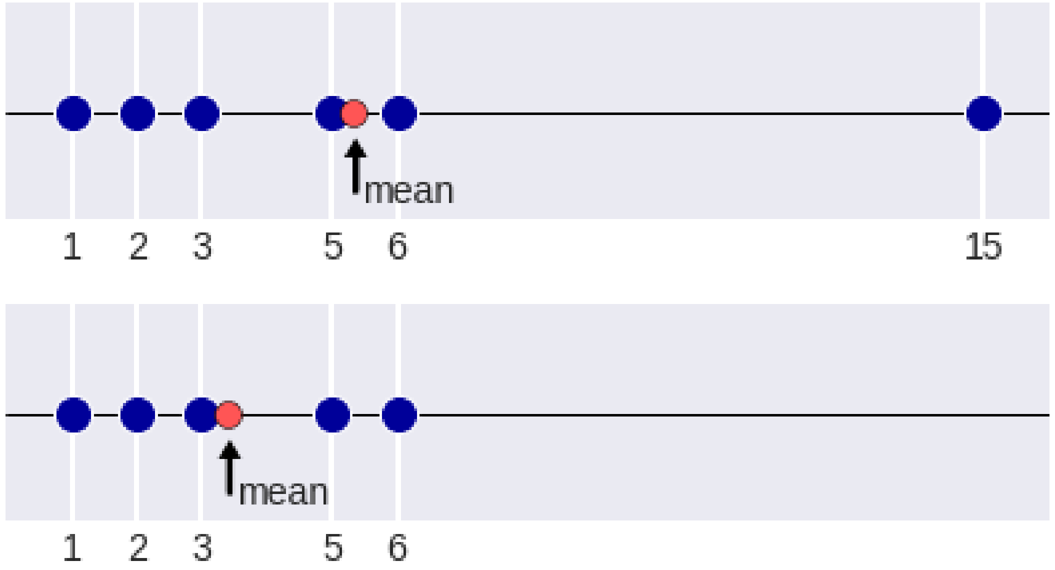
$$\mathrm{mean}(X) = \frac{x_1 + \ldots + x_m}{m}$$

Какие плюсы и минусы?

Среднее арифметическое

Большой плюс – среднее можно вычислять в $\mathbb{R}^n$

1) Проблема выбросов



Среднее арифметическое

2) Проблема «виртуальных точек»

Признак «пол»: `[M, F, F, M, M, M, F, F, F, F]`

– Какой у нас среднестатистический клиент?
– Он на 40% мужчина?
– Хочется конкретный пример!

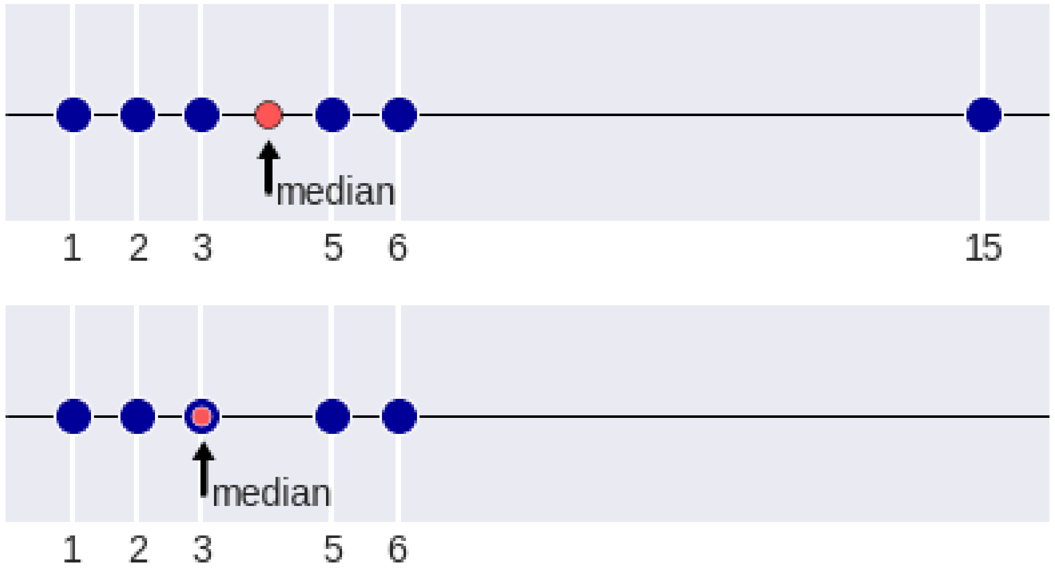
Что такое среднее?

Решение проблемы – медиана.

$$\mathrm{median}(X) = \frac{x_{\lfloor n/2 \rfloor} + x_{\lceil n/2 \rceil}}{2}$$

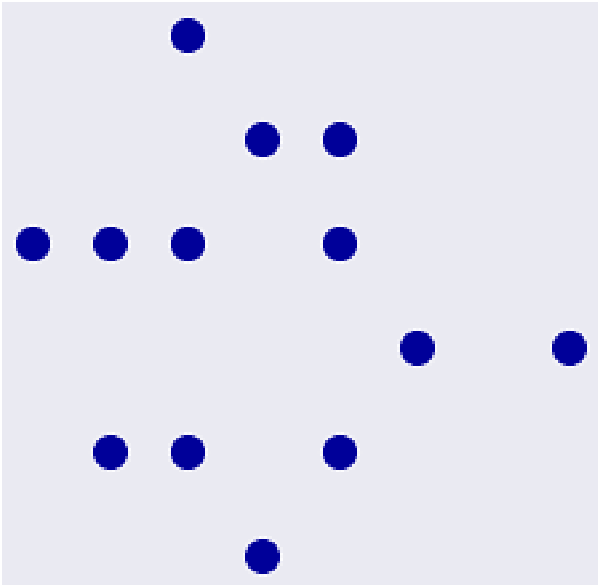
1) устойчива к выбросам

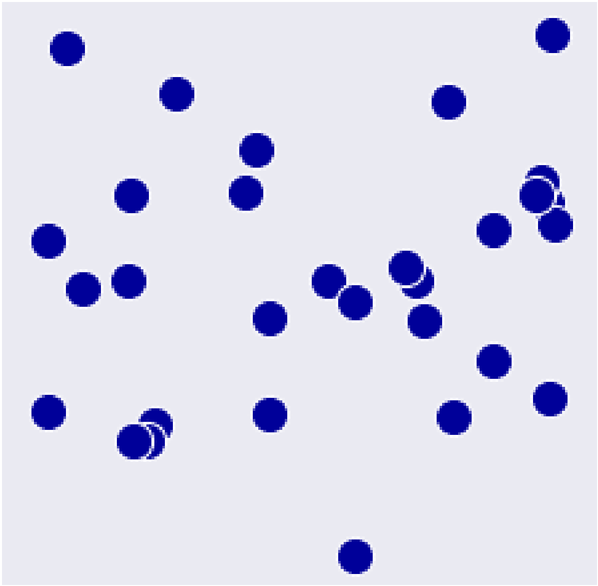
2) является (можно сделать!) точкой выборки



Проблема медианы

Что такое многомерная медиана?





Многомерная медиана

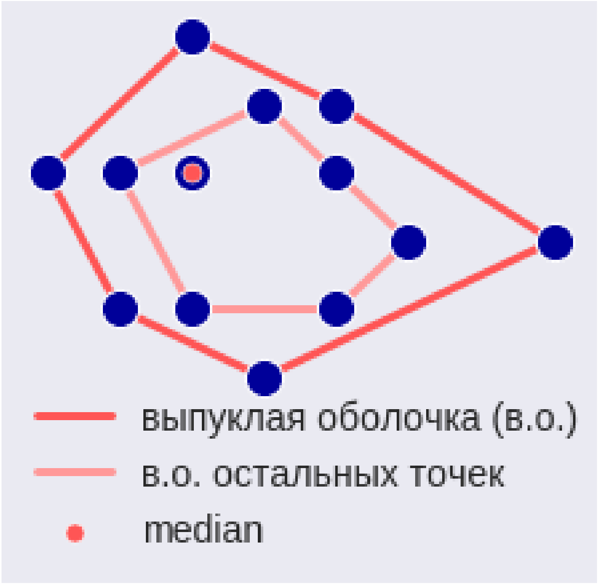
Хочется инвариантность к

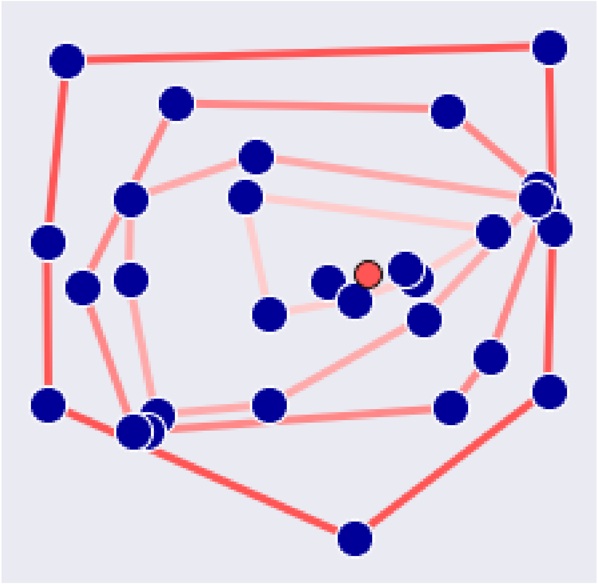
- **движениям**
  - **поворотам**
  - **сдвигам (параллельным переносам)**
- **сжатиям**

В одномерном случае должна совпадать с median!

Многомерная медиана

Что такое многомерная медиана?





Выход: сделать аналогичный процесс построения, как в одномерном случае

удаление крайних элементов!

Многомерная медиана

Если признаки разнородны, неравноценны и т.п.
(не нужно инвариантности к поворотам)

Всё равно можно применить подход
«отбрасывания крайних элементов».

Вопрос: как, где?

Среднее как решение оптимизационной задачи

- Живём в одномерном мире «на базе»
- Есть пункты интереса
- Есть функция затрат
- Надо минимизировать суммарные затраты

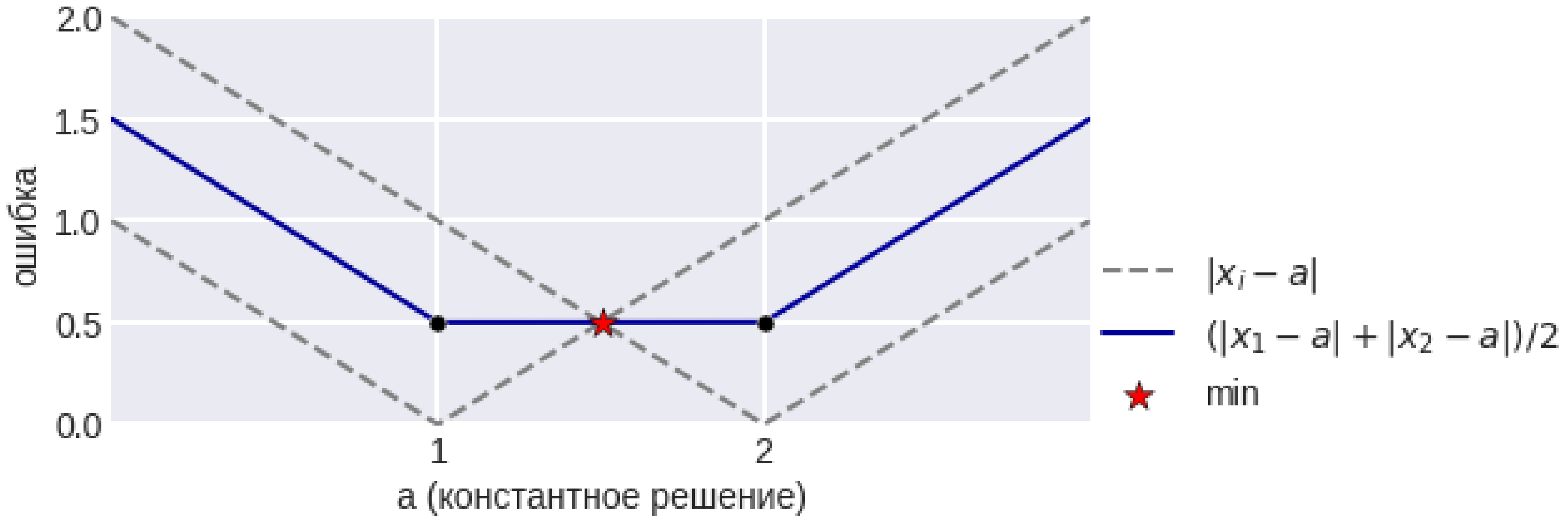
Среднее как решение оптимизационной задачи

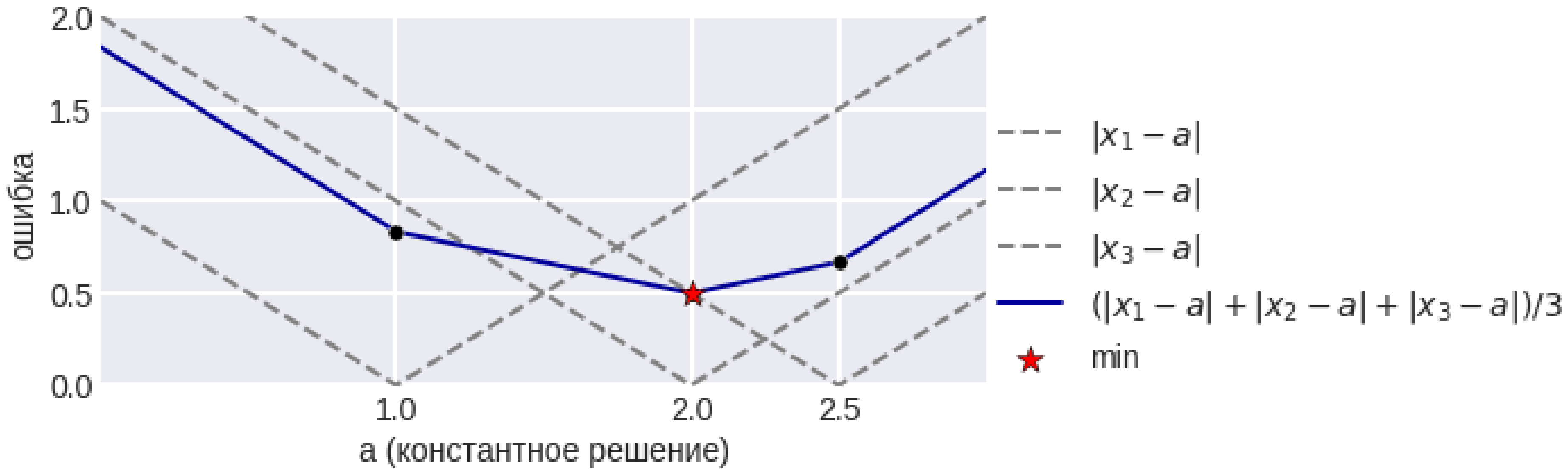
Если суммарные затраты

$$\sum_{i=1}^{m} |x_i - a| \to \min$$

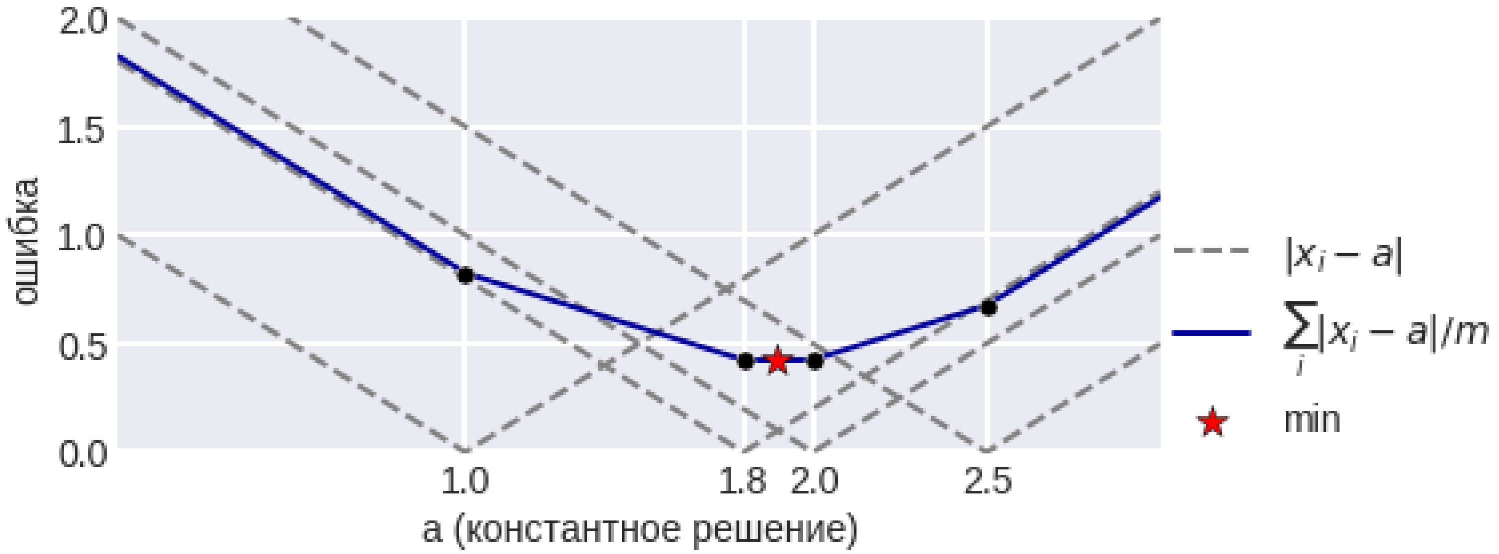
то решение – медиана

Среднее как решение оптимизационной задачи





Среднее как решение оптимизационной задачи



Медиана в пространстве

2й способ формализации: аналогично минимизируем затраты

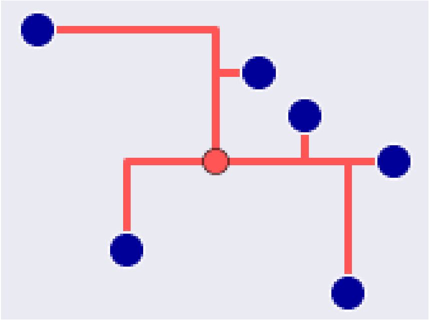
но тут может быть зависимость от координат!

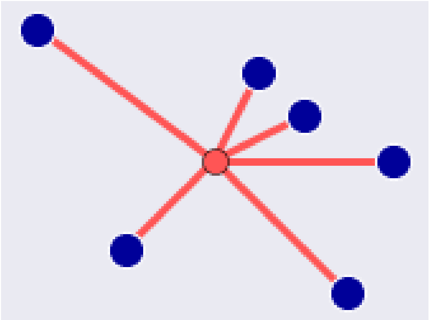
$$\sum_{i=1}^{m} \left(|x_i - a_1|^d + |y_i - a_2|^d \right)^{1/d} \to \min$$

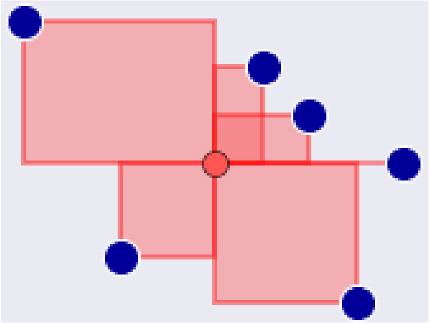
$$\sum_{i=1}^{m} |x_i - a_1| + \sum_{i=1}^{m} |y_i - a_2| \to \min$$

$$\sum_{i=1}^{m} \max[|x_i - \mu_1|, |y_i - \mu_2|] \to \min$$

$$\sum_{i=1}^{m} |x_i - a_1| \cdot |y_i - a_2| \to \min$$



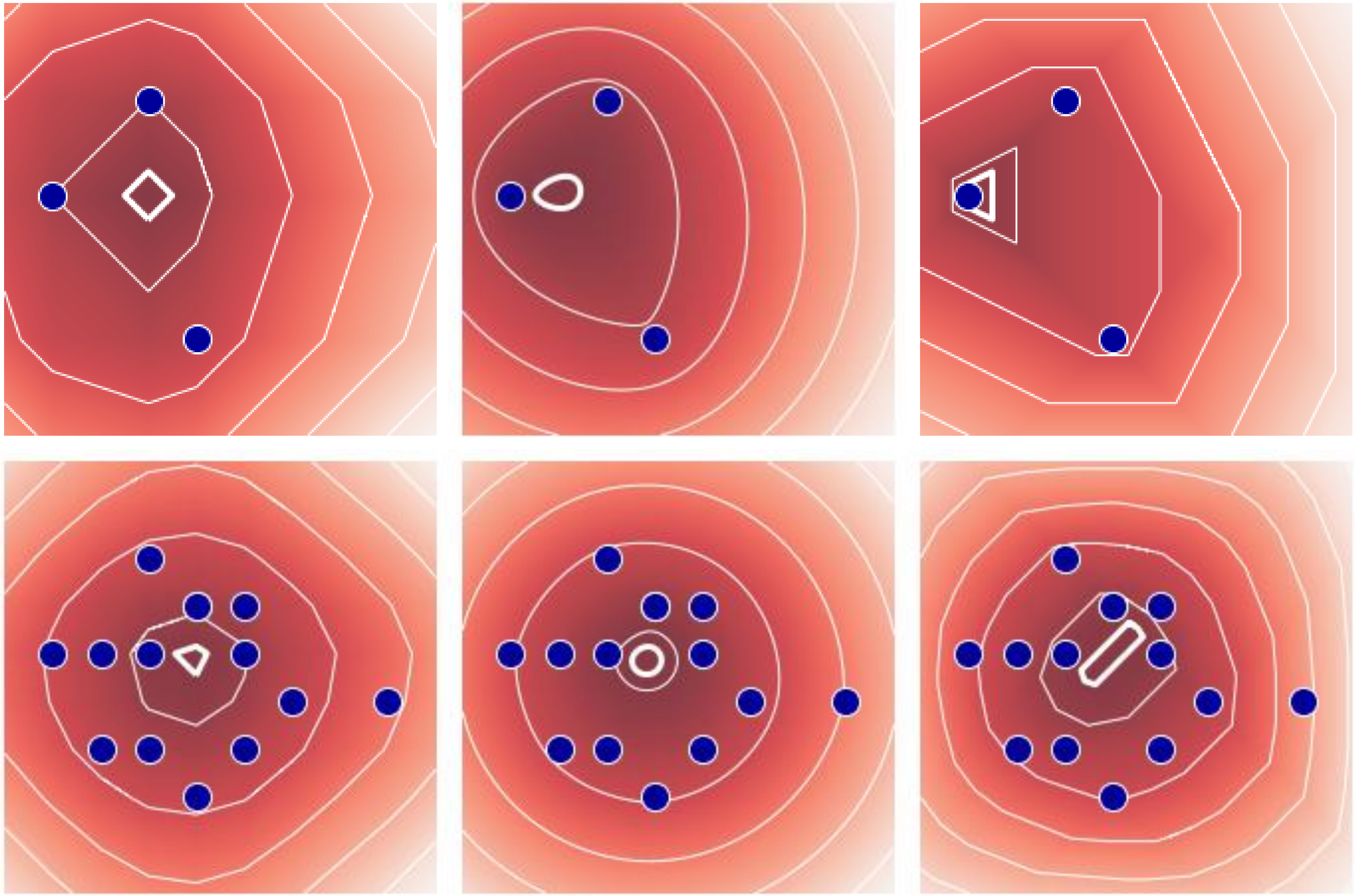




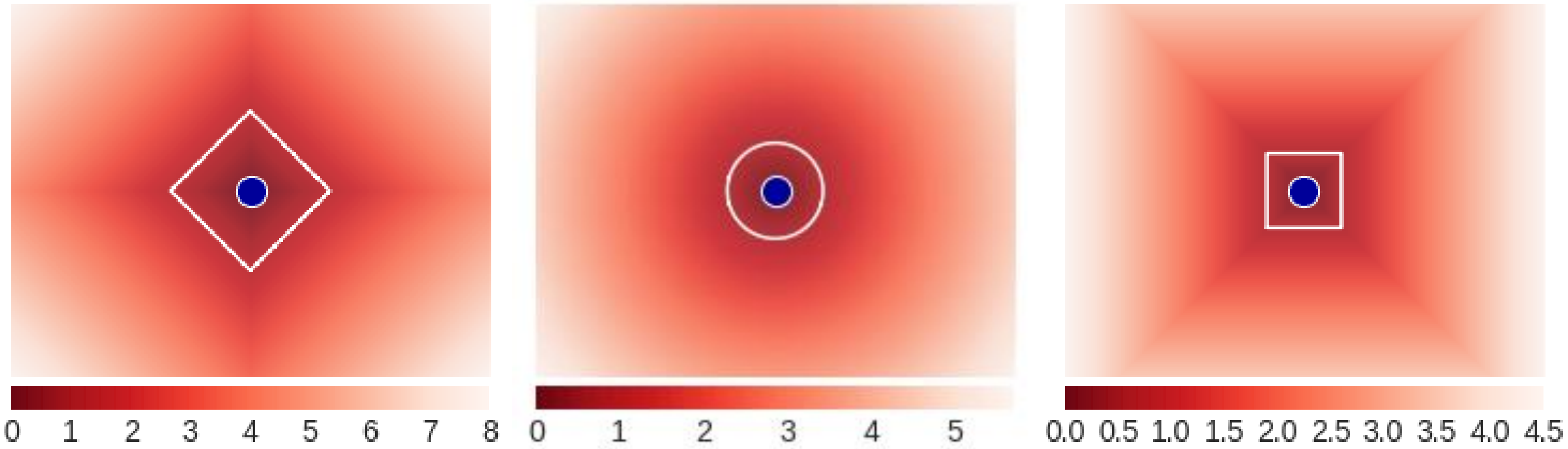
Решаем перебором по точкам выборки!!!

Д3 Оптимум на точках выборки?

«Степень медианности» – какие функции представлены?



«Степень медианности»



$$\sum_{i=1}^{m} |x_i - a_1| + \sum_{i=1}^{m} |y_i - a_2| \to \min$$

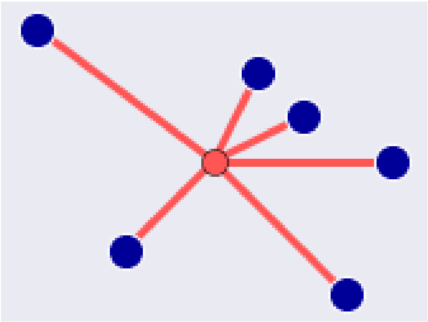
$$\sum_{i=1}^{m} (|x_i - a_1|^2 + |y_i - a_2|^2)^{1/2} \to \min$$

$$\sum_{i=1}^{m} \max[|x_i - a_1|, |y_i - a_2|] \to \min$$

ДЗ: есть ли в обобщениях желанные свойства медианы?

Геометрический центр

также 1-медиана, пространственная медиана, или точка Торричелли



$$\sum_{i=1}^{m} \left(| x_i - a_1 |^2 + | y_i - a_2 |^2 \right)^{1/2} \to \min$$

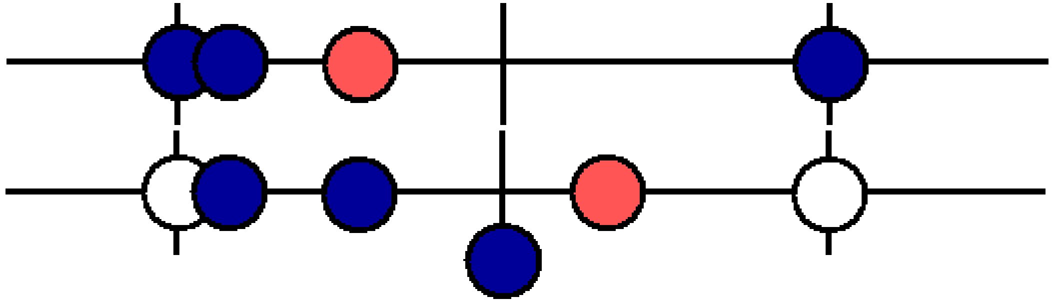
Геометрический центр единственный, когда точки неколлинеарны

Доказано: не существует ни явной формулы, ни точного алгоритма, использующего только арифметические операции и операции извлечения корней

Но можно вычислить с произвольной точностью за почти линейное время
дальше алгоритм Вайсфельда (но у него недостатки)
https://ru.m.wikipedia.org/wiki/Геометрический_центр

Эвристический способ борьбы с выбросами

$$a = \frac{1}{m}\sum_{i=1}^{m} x_i$$



Алгоритм Шурыгина

1. **Если $m \le 2$, то пользуемся формулой (*). Выход.**
2. **Пусть $x_1 \le \ldots \le x_m$ (без ограничения общности).**
3. **Если $\frac{x_1 + x_m}{2} \le x_2$, то удаляем из выборки x^1. Переходим к п.1 (с соответствующей перенумерацией объектов).**
4. **Если $\frac{x_1 + x_m}{2} \ge x_{m-1}$, то удаляем из выборки x_m. Переходим к п.1 (с соответствующей перенумерацией объектов).**
5. **Исключаем из выборки x_1, x_m, но добавляем в неё $\frac{x_1 + x_m}{2}$.**

Борьба с выбросами

В чём недостаток алгоритма Шурыгина?

Практика: часто забываем о выбросах

Что минимизирует «среднее»

$$\mathrm{median}(X) = \arg\min \sum_{i=1}^{m} | x_i - a |$$

$$\mathrm{mean}(X) = \arg\min \sum_{i=1}^{m} | x_i - a |^2$$

Для минимизации можно выбрать «что угодно»

$$\mathrm{mid}(X) = \arg\min \sum_{i=1}^{m} f(x_i, a)$$

– оценка минимального контраста

… другие формализации понятия «среднее»

Оценка минимального контраста

Если после дифференцирования
(здесь рассматриваем одномерный случай)

$$\sum_{i=1}^{m}\psi(x_i-a)=\sum_{i=1}^{m}(x_i-a)\xi(x_i-a)=0,$$

для некоторых функций ψ (оценочная функция) и ξ (весовая функция),
то часто успешно применяется итеративный способ
вычисления параметра a по формуле

$$a=\frac{\sum_{i=1}^{m}x_i\xi(x_i-a)}{\sum_{i=1}^{m}\xi(x_i-a)}.$$

Откуда взялась формула?

Д/З Проверить применимость формулы

Принстонский эксперимент 1972 года

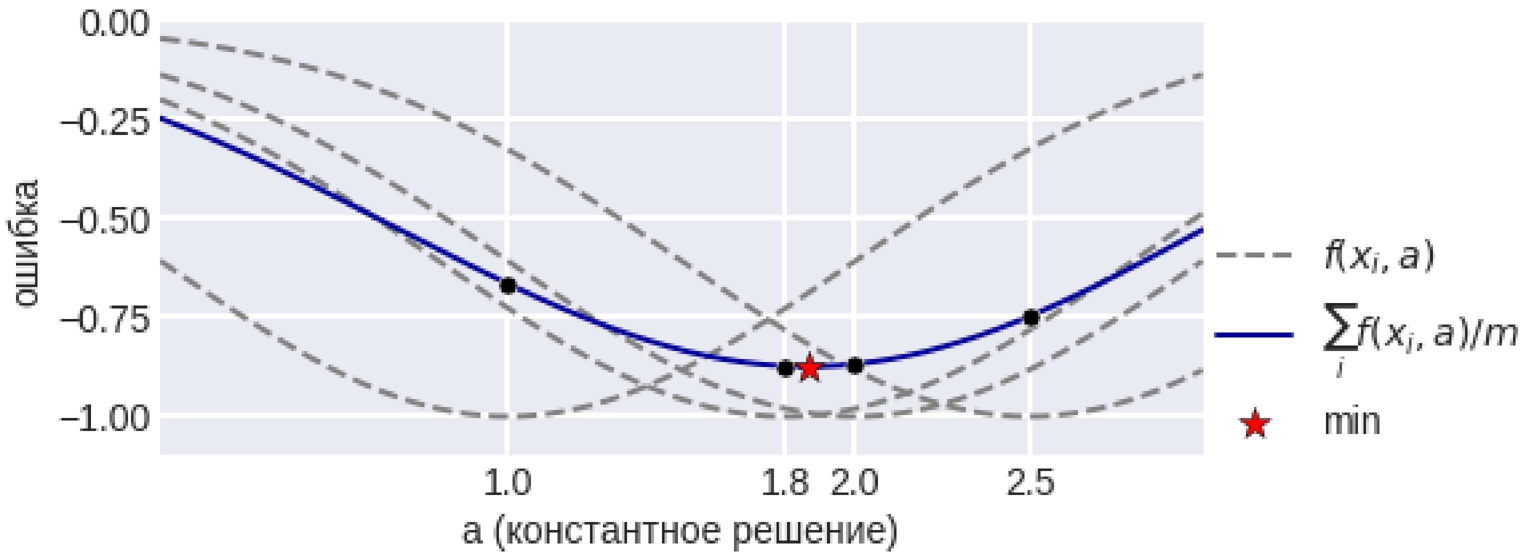
подбор различных функций

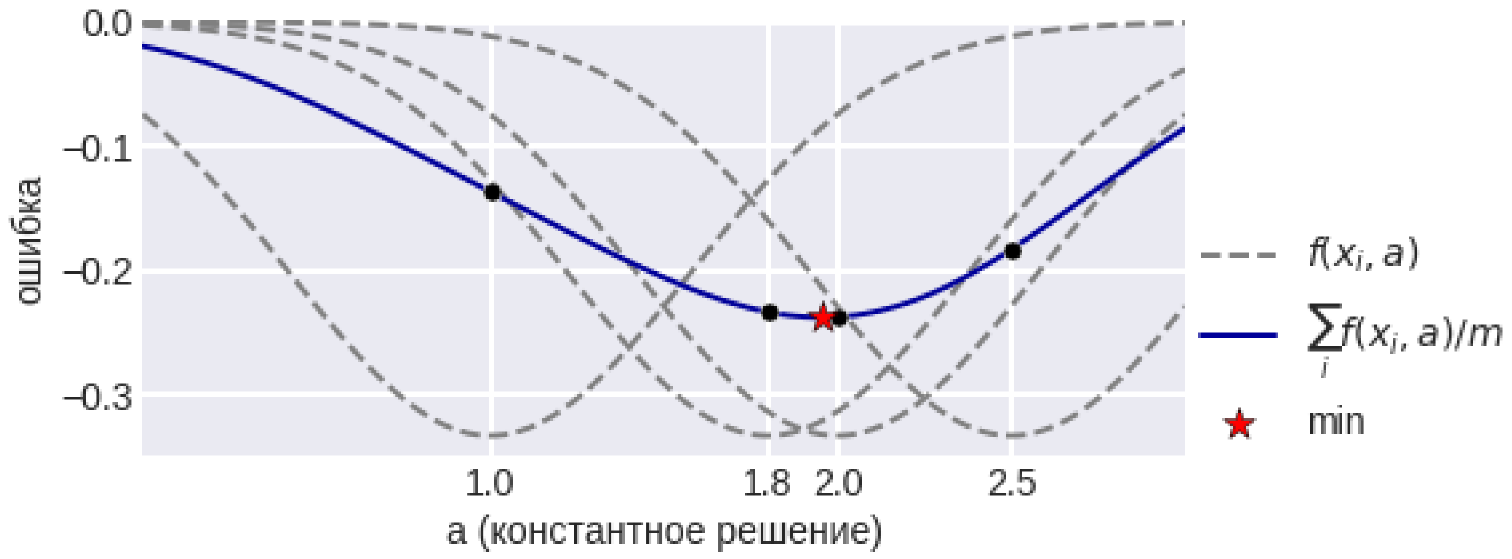
Мешалкин Л.Д. (1977) предлагал

$$f(x,a)=-\frac{1}{\lambda}e^{-\frac{\lambda(x-a)^2}{2}}$$

$$\psi(z)=ze^{-\lambda z^2/2},\ \xi(z)=e^{-\lambda z^2/2}.$$

Чем отличаются рисунки?

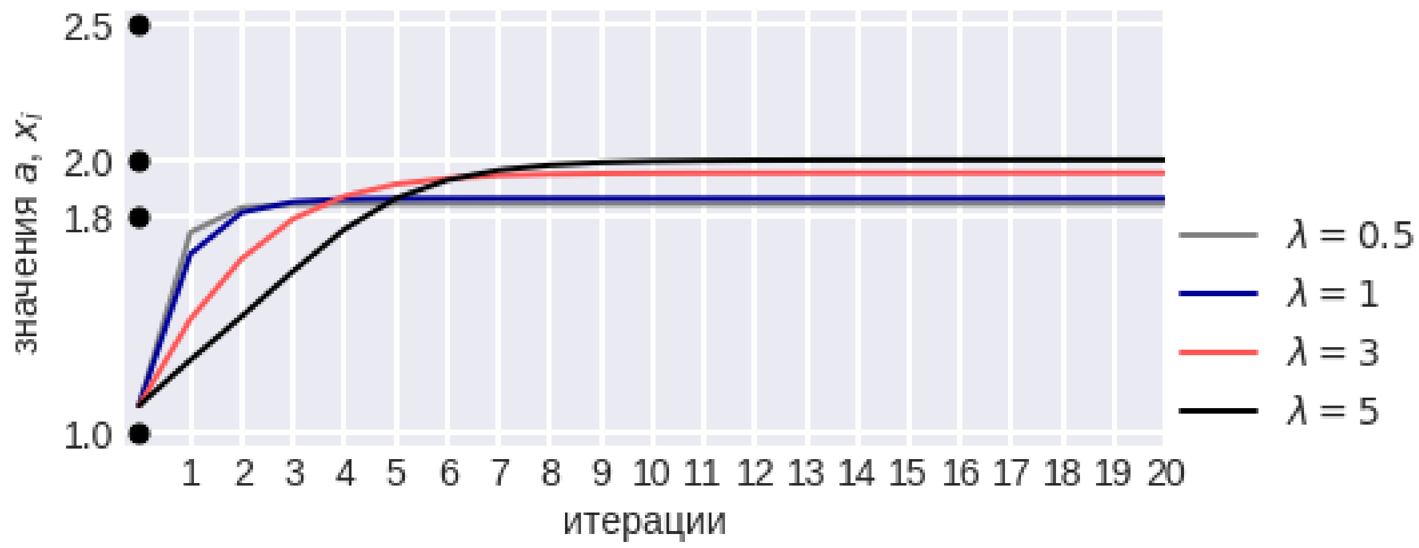


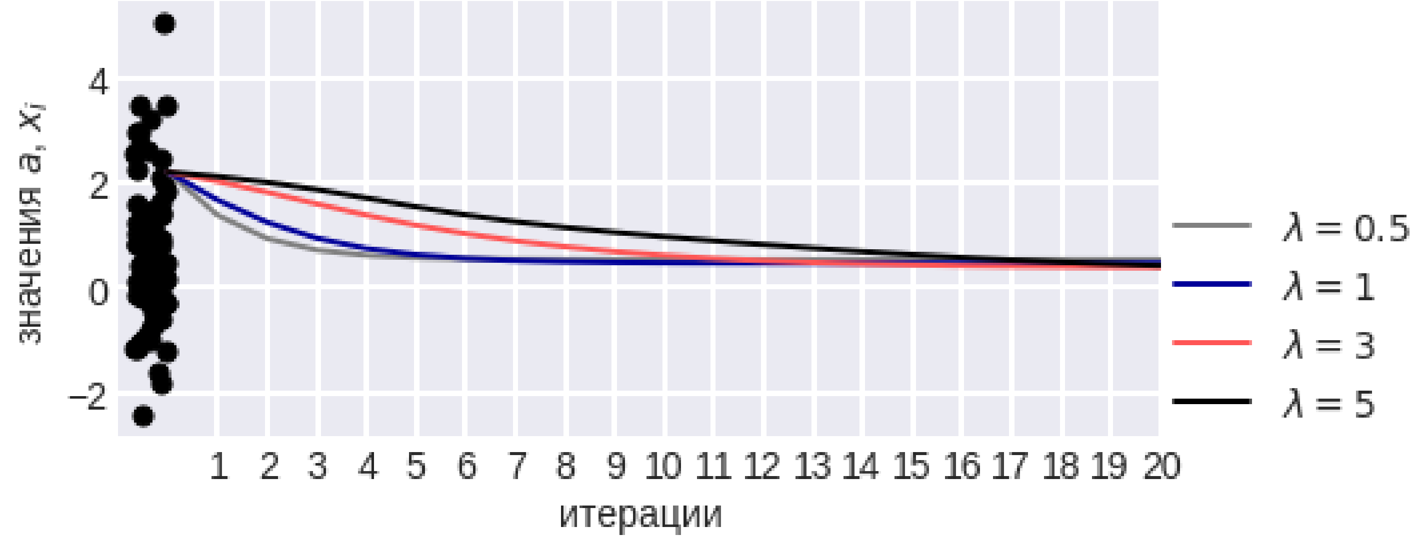


Чем отличаются рисунки?

$\lambda = 1 \ \lambda = 3$

Результаты пересчёта: что важно, как в любой задача оптимизации?



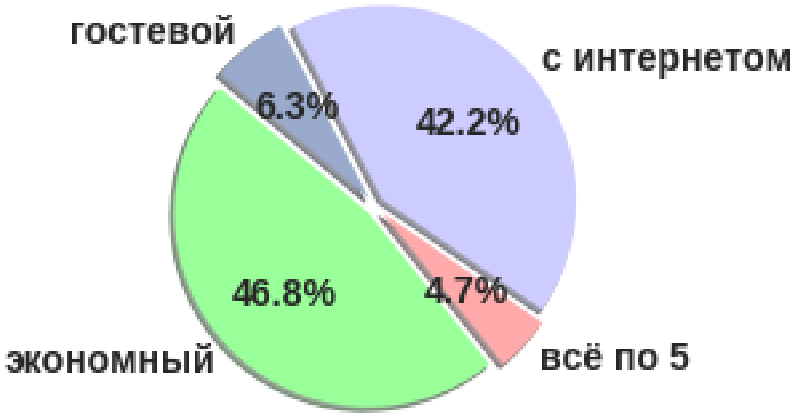


Что важно?

Начальное приближение

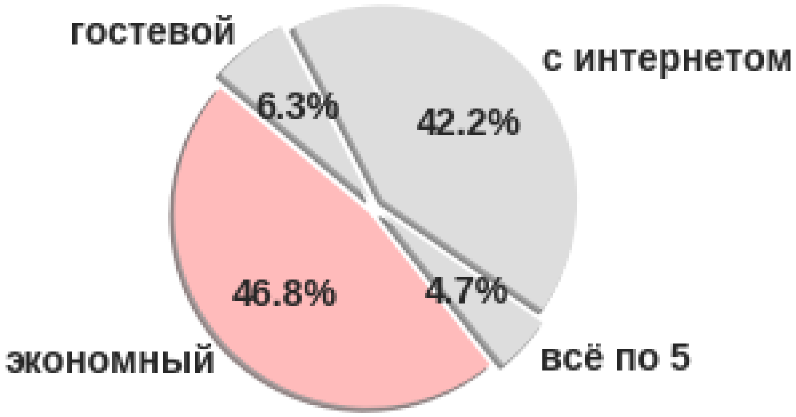
Масштаб

Что такое среднее для номинальных признаков?



Сколько клиентов выбрали определённой тариф сотовой связи

Что такое среднее для номинальных признаков?

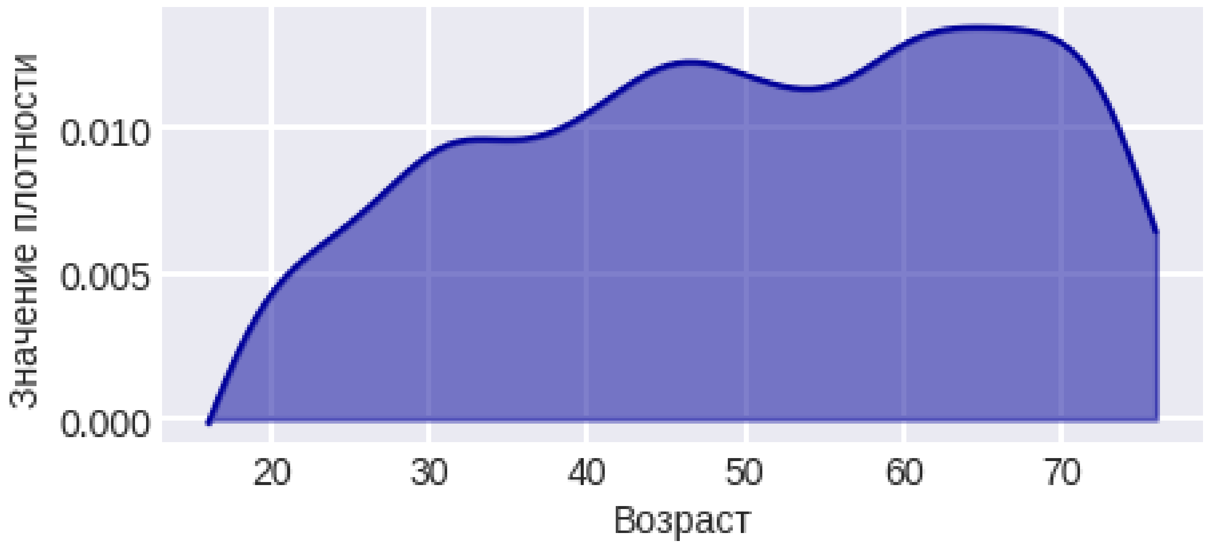


Мода – самое популярное значение

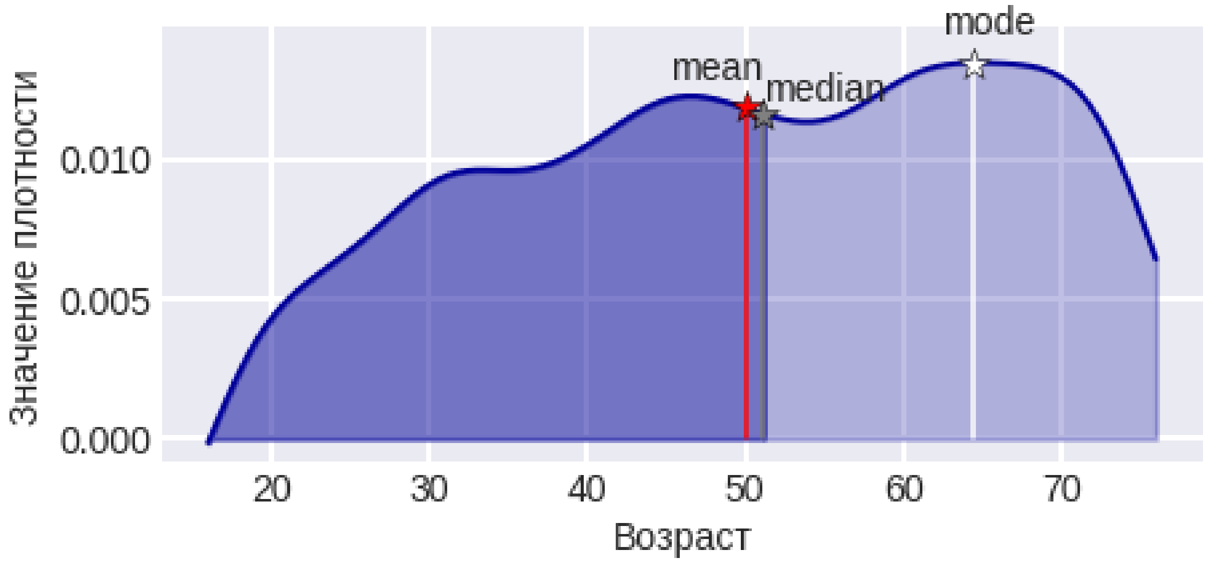
– самое вероятное значение

Что такое среднее для порядковых признаков?

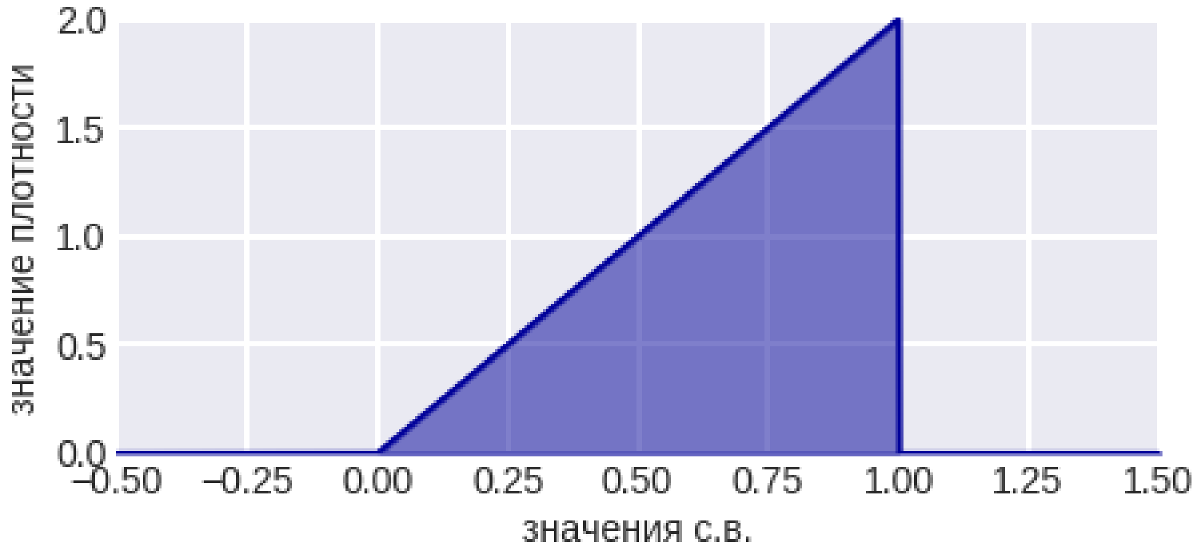
Где матожидание, медиана, мода?



Где матожидание, медиана, мода?

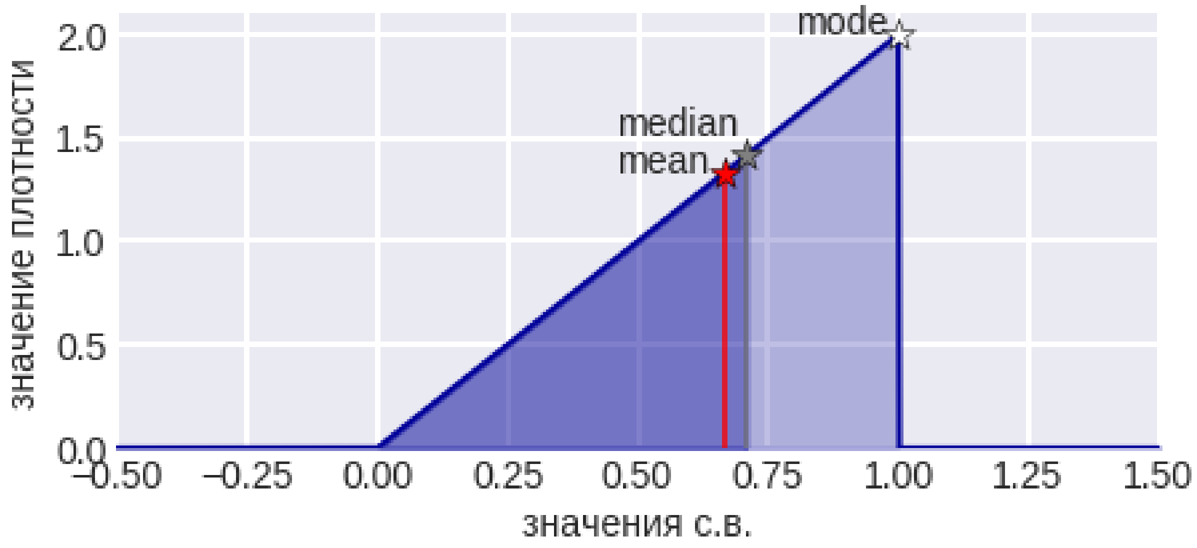


Как запомнить



Где мода, матожидание и медиана?

Как запомнить



$$\mathbf{E}x = \int_0^1 x 2x \partial x = \frac{2}{3} x^3 \Big|_0^1 = \frac{2}{3} \approx 0.(6)$$

$$\int_0^{\text{median}} 2x \partial x = \text{median}^2 = \frac{1}{2} \Rightarrow \text{median} = \frac{\sqrt{2}}{2} \approx 0.71$$

ДЗ может ли быть другой порядок?

Практика: придумывать не функционал, а среднее

Среднее по А.Н.Колмогорову

$$\varphi^{-1}\left(\frac{\varphi(x_1)+\ldots+\varphi(x_n)}{n}\right)$$

среднее арифметическое $\varphi(x)=x$

среднее геометрическое $\varphi(x)=\log x$

среднее гармоническое $\varphi(x)=x^{-1}$

среднее квадратическое $\varphi(x)=x^{2}$

где медиана и мода?

что такое среднее по Коши?

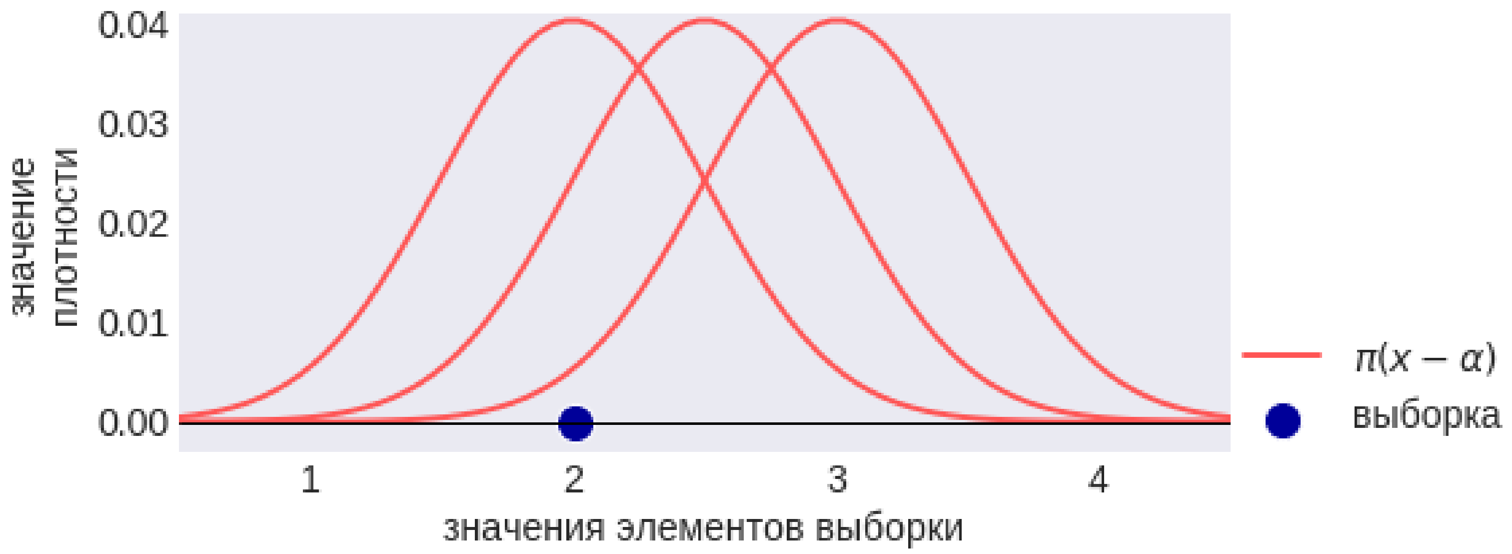
Оценивание вероятности

тоже, в некотором смысле, усреднение… сейчас объясним

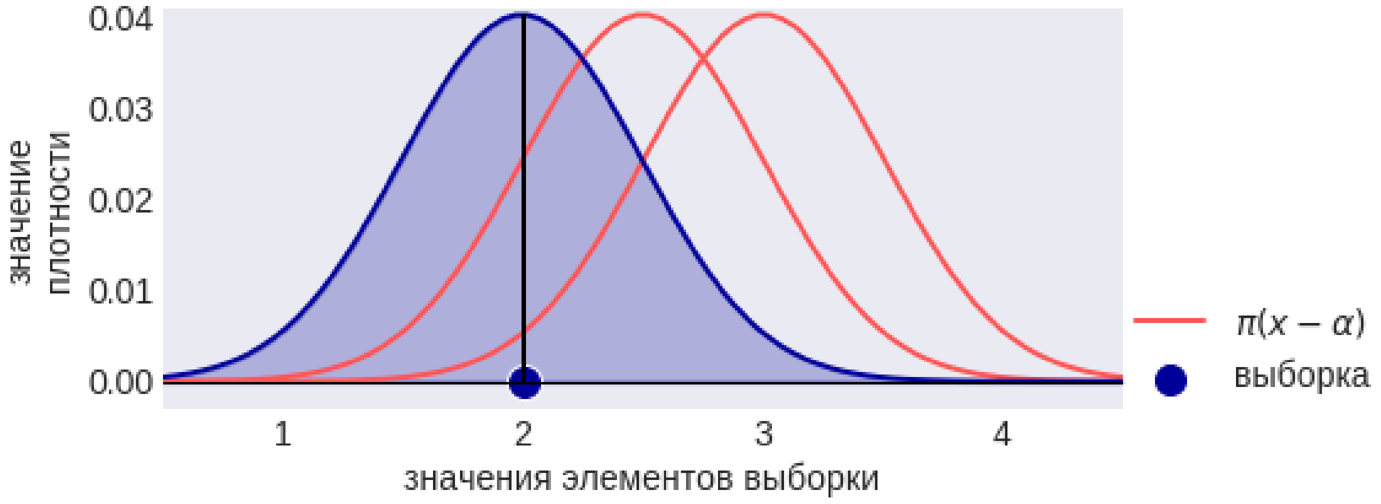
Метод максимального правдоподобия

Есть выборка $x_1,\ldots,x_n$ **какое распределение** $\pi_\alpha(x)$**?**

Пусть $m=1$**,** $\pi_\alpha(x)=\pi(x-\alpha)$ **какое распределение выбрать?**



Метод максимального правдоподобия



$$\pi_\alpha(x_1) \to \max_\alpha$$

Пусть $m = 2$

Метод максимального правдоподобия

Пусть $m = 2$

$$\pi_\alpha(x_1)\cdot\pi_\alpha(x_2)\to\max_\alpha$$

Общий случай:

$$\prod_{i=1}^{m}\pi_\alpha(x_i)\to\max_\alpha$$

Как максимизируют?

Случай биномиального распределения

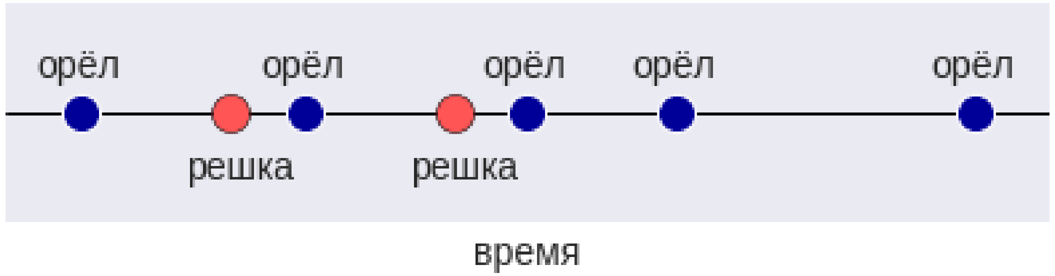
$$\pi_p(x) = \begin{cases} p, & x = 1, \\ 1 - p, & x = 0. \end{cases}$$

$$\Pi = \prod_{i=1}^{n} \pi_p(x_i) = p^m (1-p)^{n-m} \sim m \log p + (n-m) \log(1-p)$$

$$(\log \Pi)' = \frac{m}{p} - \frac{(n-m)}{1-p} = 0$$

$$p = \frac{m}{n}$$

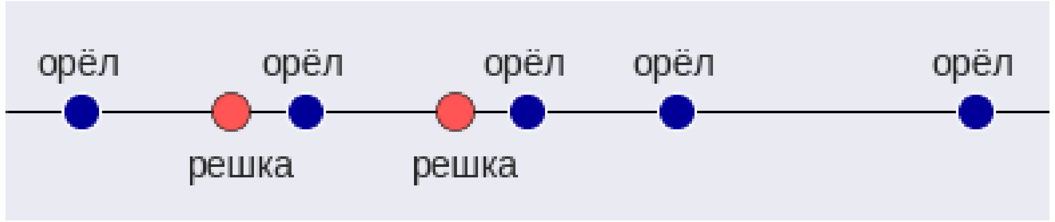
Самый очевидный ответ для оценки вероятности!



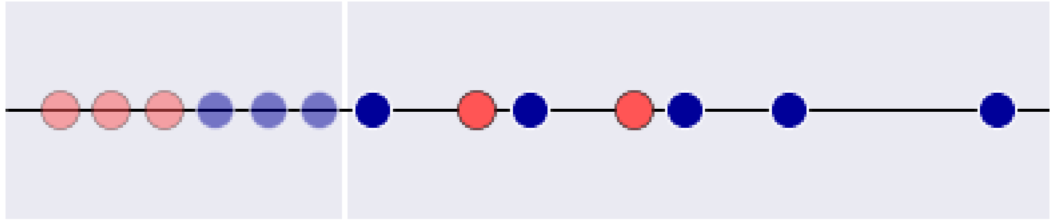
$$p = \frac{5}{5+2} = \frac{5}{7} \approx 0.71$$

Оценивание вероятности – сглаживание Лапласа

тоже, в некотором смысле, усреднение



на практике есть априорная вероятность



$$\frac{m+\lambda\cdot p}{n+\lambda}=\frac{5+6\cdot 0.5}{5+2+6}\approx 0.62$$

Есть разные эвристические методы

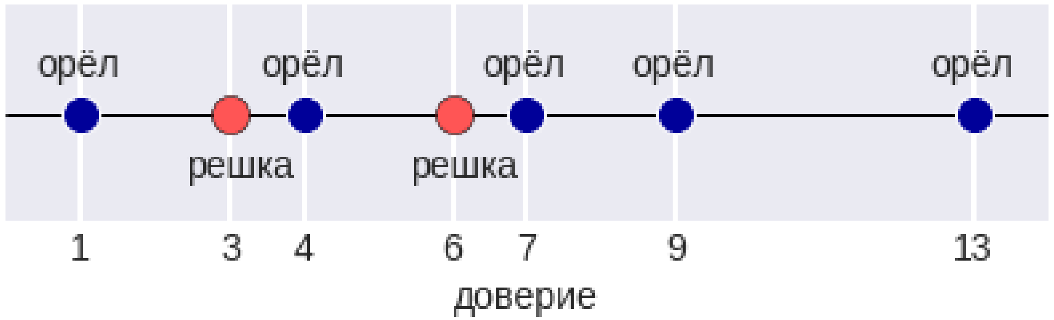
$$\sigma(n)\frac{m}{n}+(1-\sigma(n))\,p$$

какую весовую функцию выбрать?

ДЗ Придумать и обосновать подобные функции.

Вторая особенность практики

Не все эксперименты равнозначны!



$$\frac{1+4+7+9+13}{1+3+4+6+7+9+13}=0.79$$

Весовая схема

$$\frac{w_{i_1}+\ldots+w_{i_m}}{w_1+\ldots+w_n}$$

Веса (доверие) возникают даже там, где нет эксперта

- **есть временная ось**
- **есть «такие же условия»**
- **есть кластеры (и схожесть вообще)**

Зодиакальный скоринг

Знак зодиака	Сколько представителей знака допускают хотя бы одну просрочку
Овен 	35.3%
Дева 	35%
Рыбы 	34.2%

где ошибка?

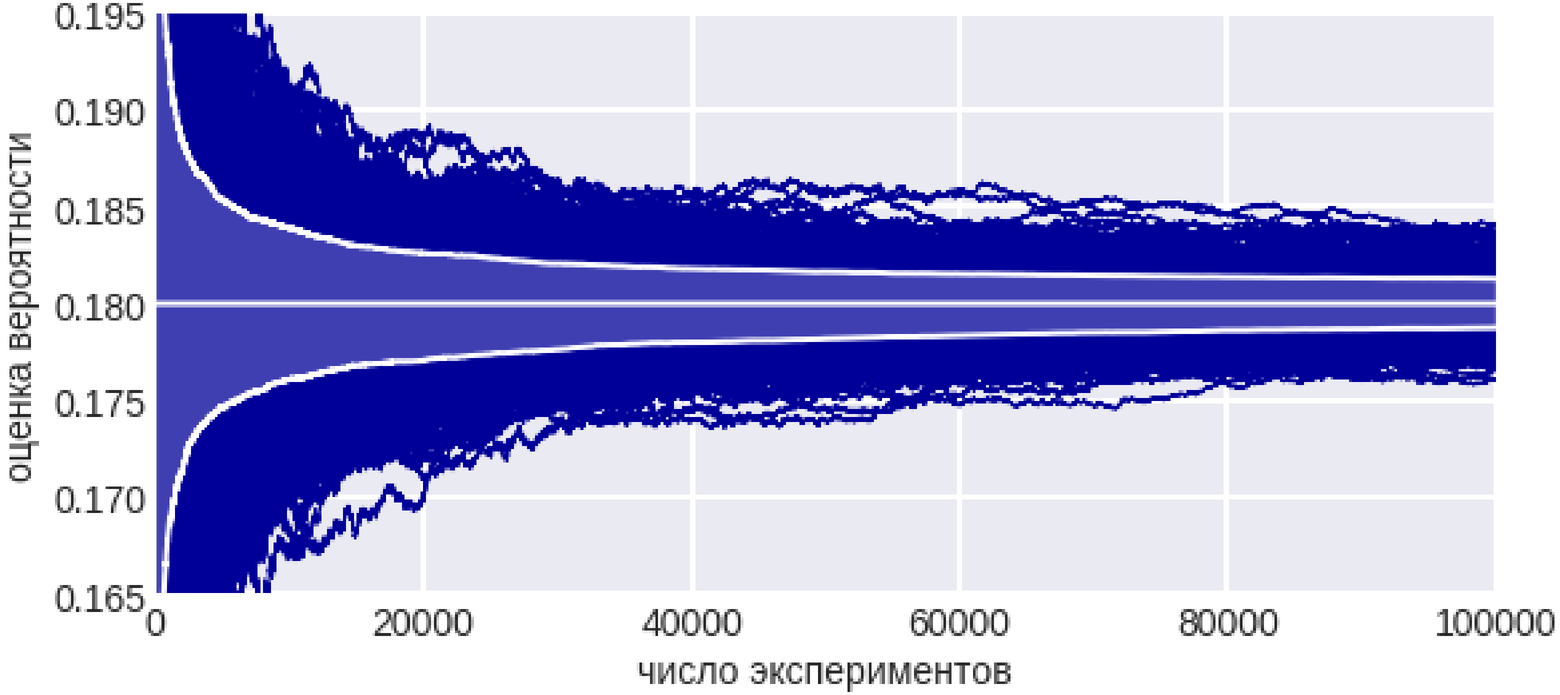
http://www.banki.ru/news/daytheme/?id=7408493

http://moneyman.ru/articles/goroskop-moneyman

Что ещё нужно знать про вероятности

Объёмы выборок

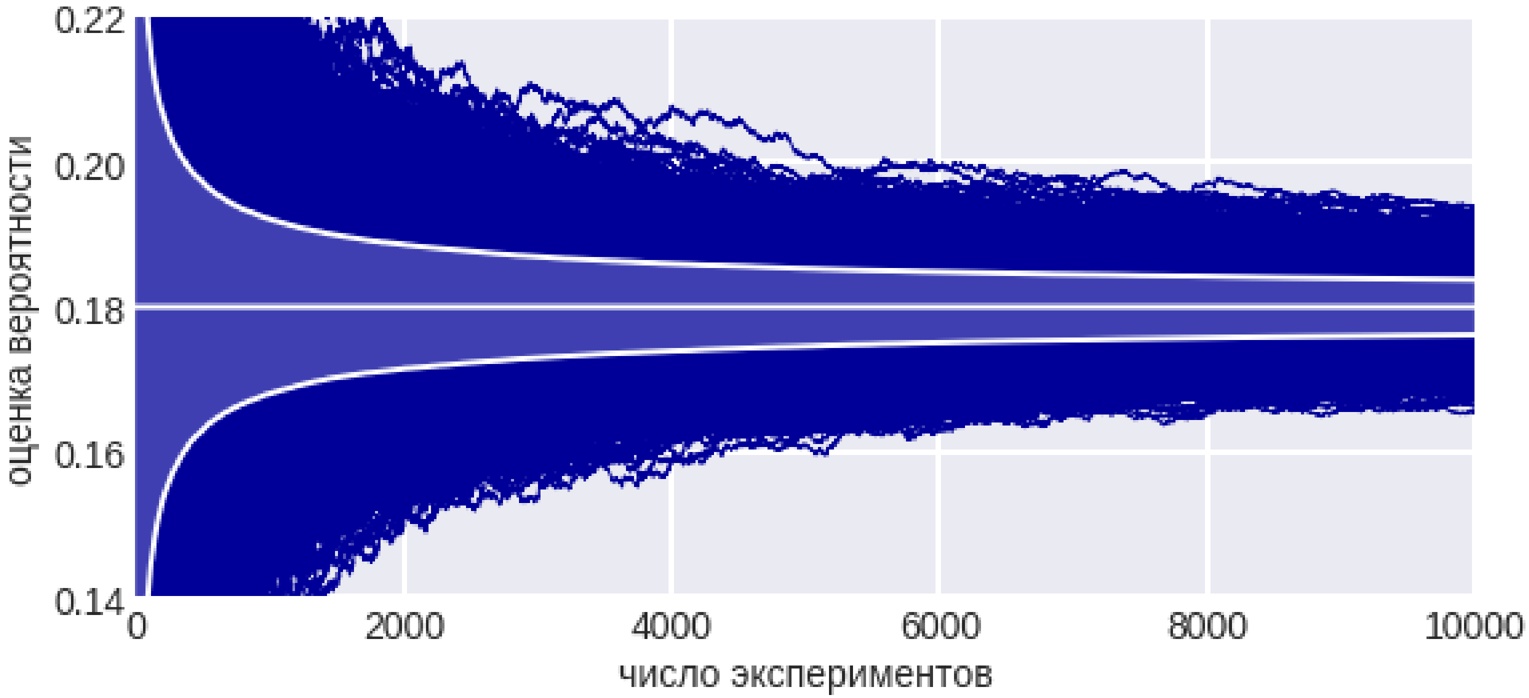
Оцениваем вероятность в схеме Бернулли (неизвестная p=0.18)



1000 экспериментов

Что ещё нужно знать про вероятности

Объёмы выборок



Выборки 10000 достаточно, но это чтобы оценить с точность ±0.01

с точностью 99%

Д/З так ли это?

Что ещё нужно знать про вероятности

Классика статистики: есть точность,
а есть вероятность того, что мы оценили с этой точностью

Д/З сколько нужно опросить перед выборами людей,
чтобы получить достоверную оценку общественного мнения?
что здесь такое «достоверная»?

Зодиакальный скоринг

- **достаточно ли велика выборка**

более 250000 + <10% каждого знака + 10% получили микрозаймы

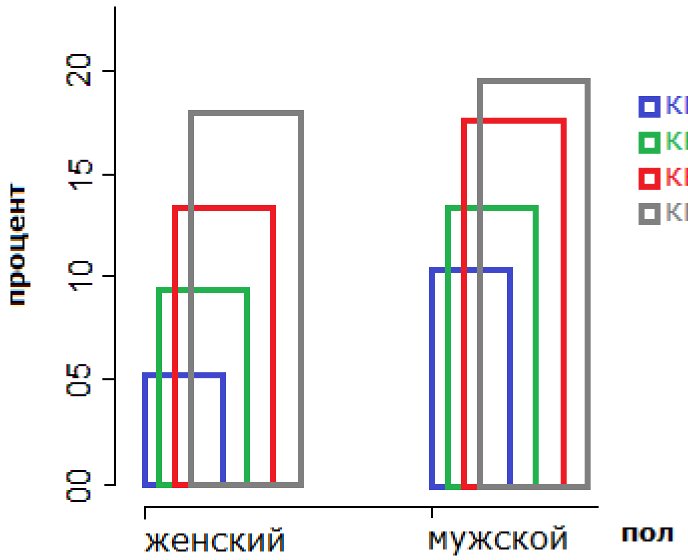
- **значимы ли отклонения в процентах**
- **насколько закономерности устойчивы**

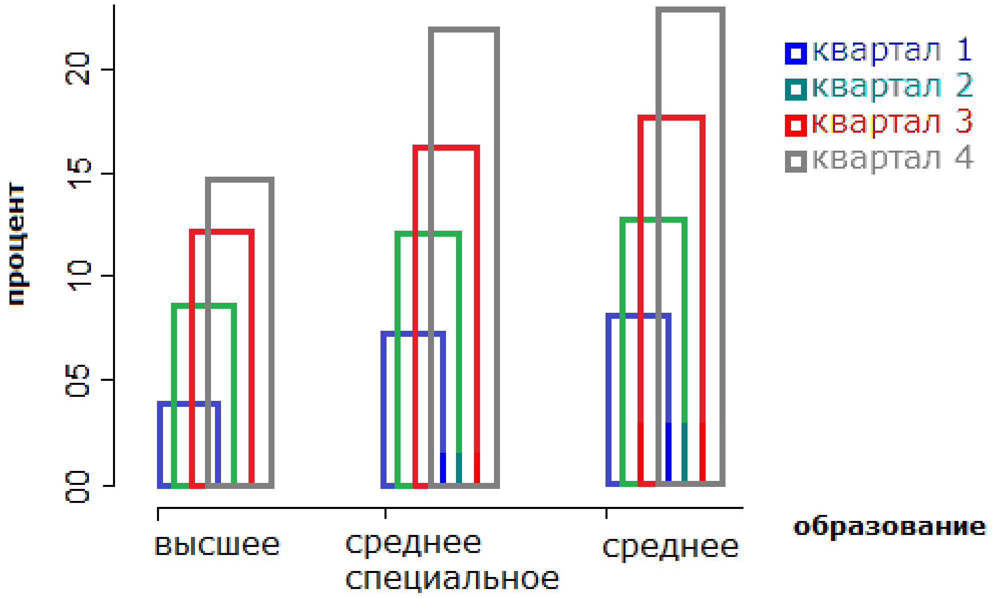
(ex: не зависят от времени)

Эксперименты с банковскими данными

300000 клиентов

классические скоринговые признаки

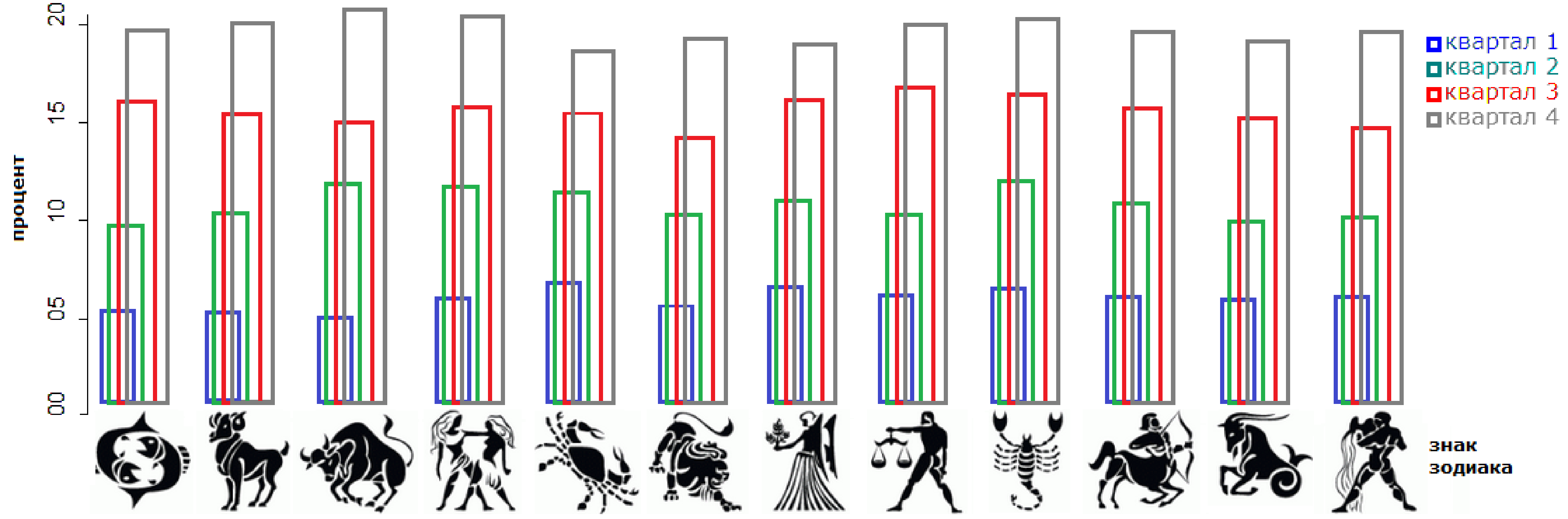




Есть устойчивость по кварталам!

Эксперименты с банковскими данными

Неклассические скоринговые признаки



Нет устойчивости по кварталам!

Логическая закономерность тогда является таковой, когда с её помощью можно что-то предсказать!

Эксперименты с банковскими данными

Д/З в чём слабость наших аргументов?

Итог

формализаций средних много
(по Колмогорову + медиана, мода, …)

среднее
- формула
- решение задачи оптимизации
- ответ некоторого алгоритма
- есть ещё подход…

важны априорные знания (сглаживание Лапласа)!
Не все объекты равноценны (весовые схемы)

Объём выборки для правильных выводов

Д/З другие способы обобщения медианы…

Ещё подход к формализации среднего…

среднее арифметическое – оценка ММП центра нормального распределения

медиана – оценка ММП центра распределения Лапласа

Поэтому можно формализовать с помощью распределения!

вспомним, когда будем говорить про оценку качества регрессоров

Литература

- **Шурыгин А.М. Математические методы прогнозирования // М., Горячая линия — Телеком, 2009, 180 с.**

  нужные фрагменты есть в http://www.machinelearning.ru/wiki/images/7/7e/Dj2010up.pdf

- **Неправильные интерпретации и ложные закономерности в анализе данных**
  https://alexanderdyakonov.files.wordpress.com/2015/07/dyakonovfunnydm.pdf