курс «Машинное обучение»

Оценки среднего, вероятности и плотности; весовые схемы

Александр Дьяконов

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

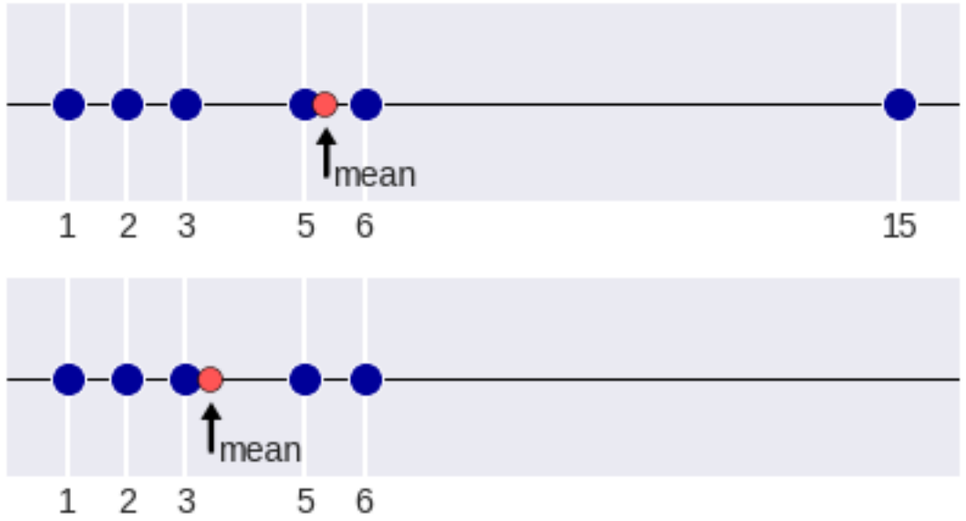
$$\text{mean}(X) = \frac{x_1 + \ldots + x_m}{m}$$

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!

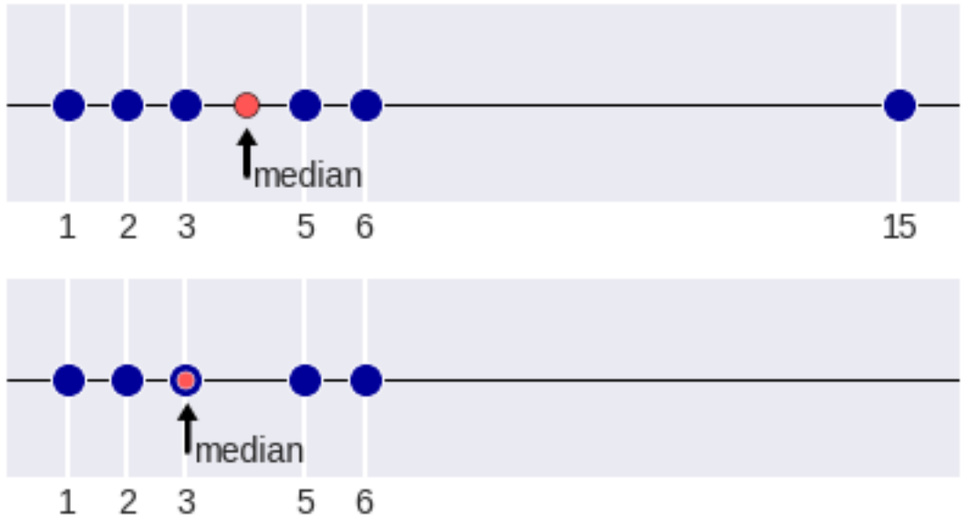
Что такое среднее?

Решение проблемы – медиана, для $x_1 \le x_2 \le \ldots \le x_m$ **:**

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

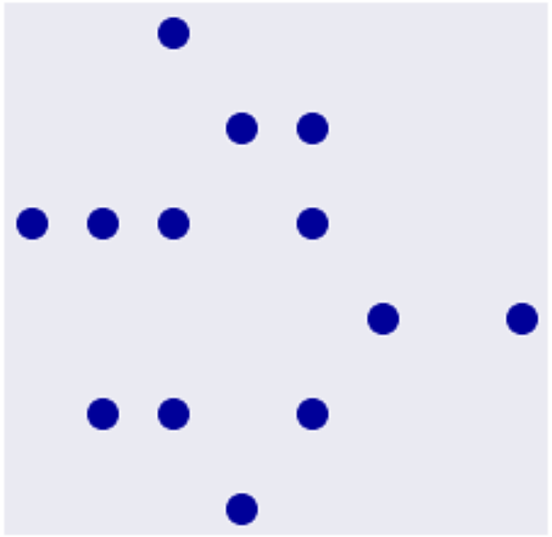
1) устойчива к выбросам

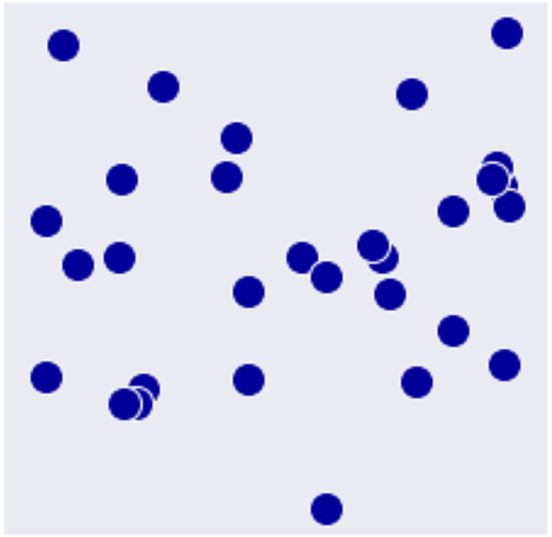
2) является (можно сделать!) точкой выборки



Проблема медианы

Что такое многомерная медиана?





Многомерная медиана

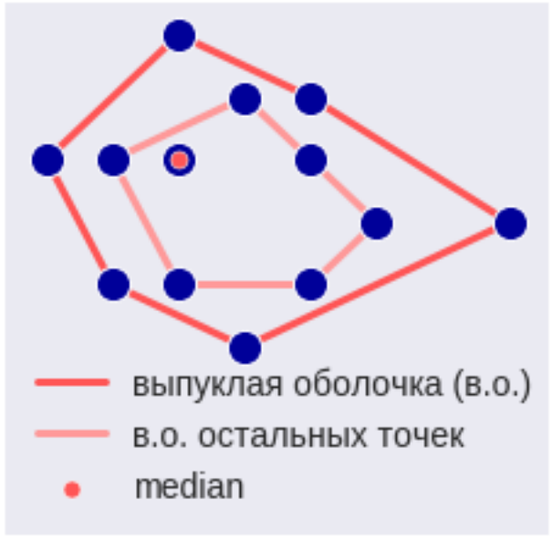
Хочется (может быть) инвариантность к

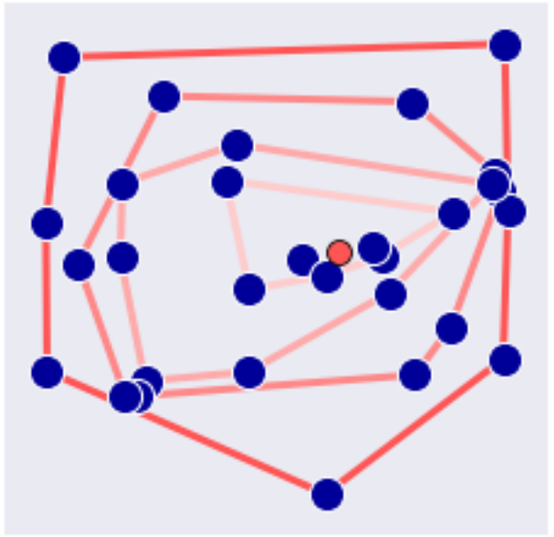
- **движениям**
  - **поворотам**
  - **сдвигам (параллельным переносам)**
- **сжатиям / растяжениям**

В одномерном случае должна совпадать с median!

Многомерная медиана как результат итерационного процесса

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

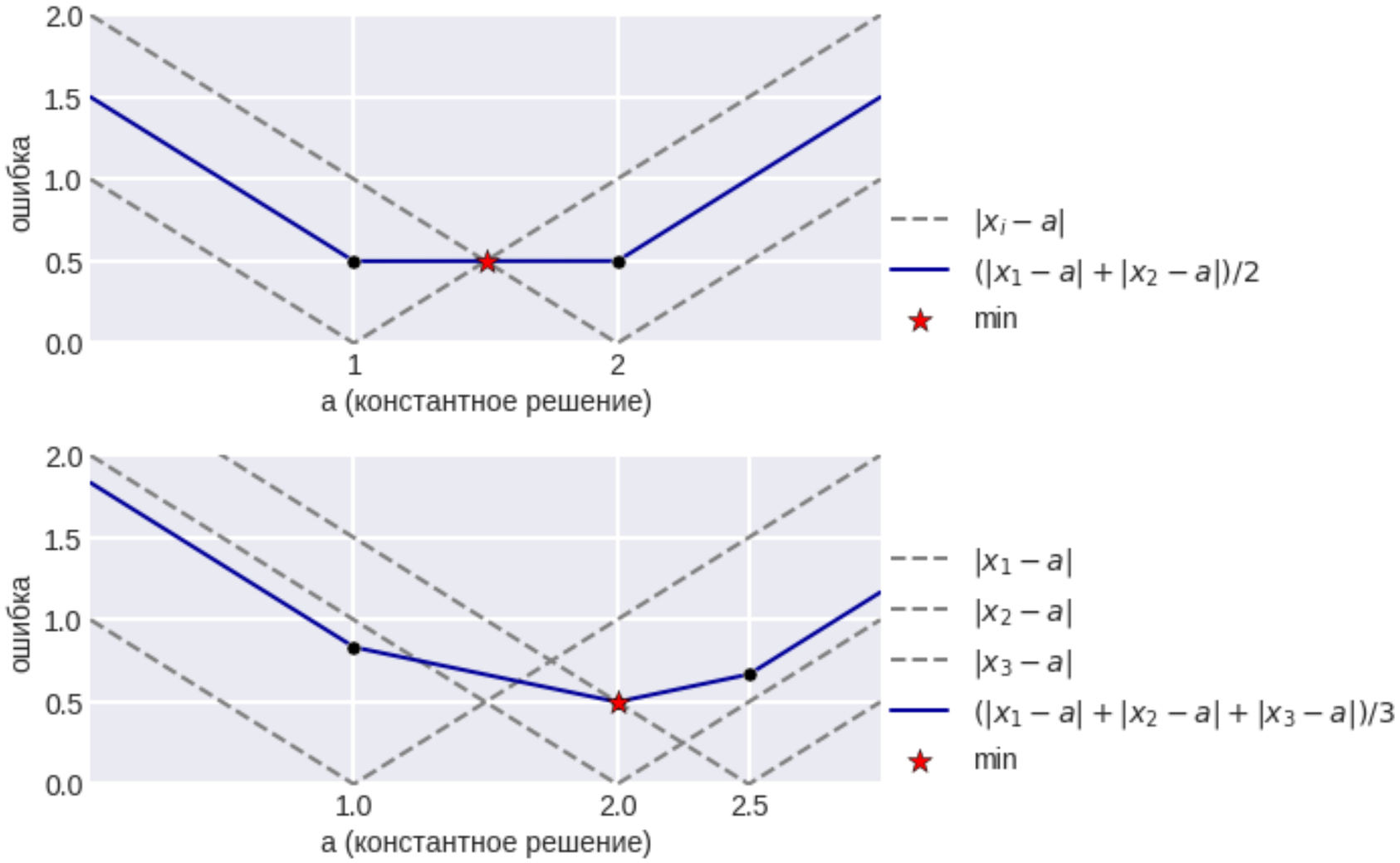
Среднее как решение оптимизационной задачи

Если суммарные затраты

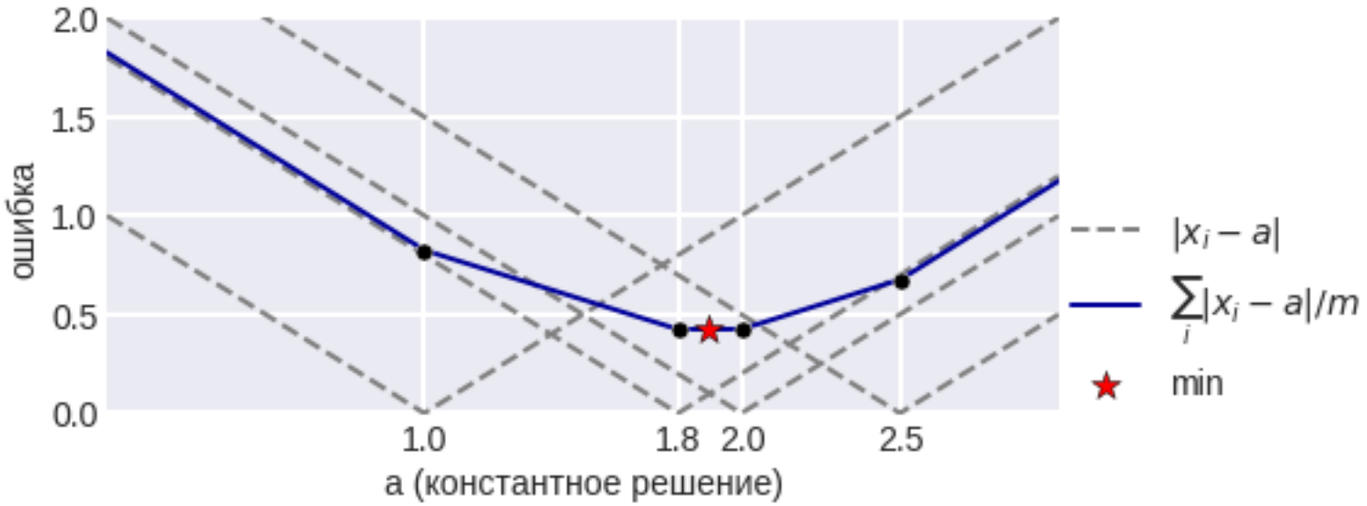
$$\sum_{i=1}^{m} | x_i - a | \to \min$$

то решение – медиана

Среднее как решение оптимизационной задачи



Среднее как решение оптимизационной задачи



Медиана в пространстве

2й способ формализации: аналогично минимизируем затраты

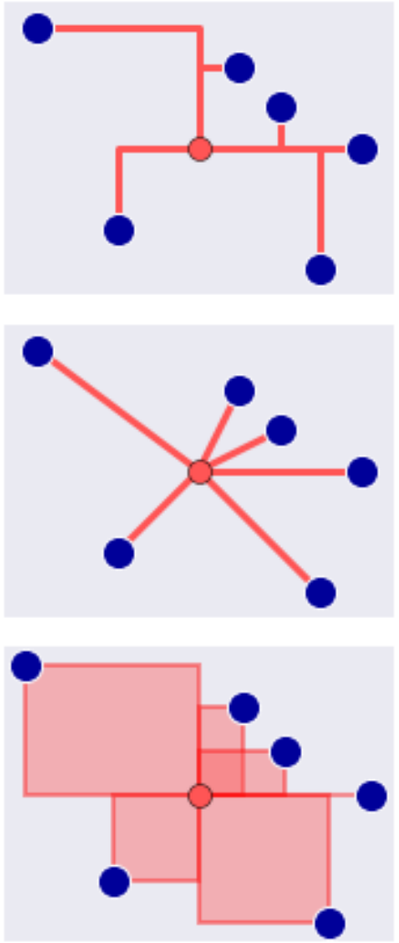
но тут может быть зависимость от координат!

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

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

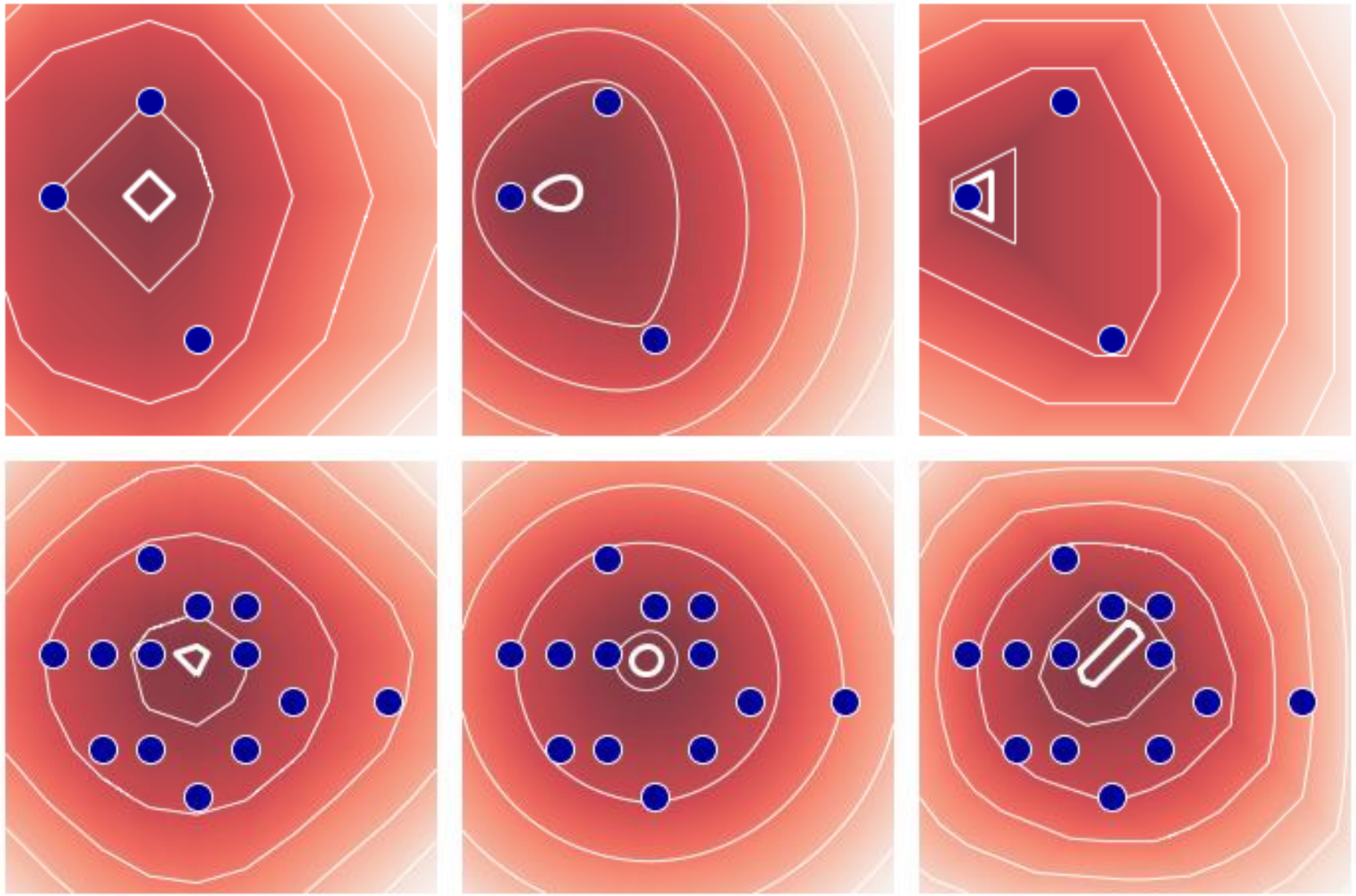
$$\sum_{i=1}^{m}\max[|x_i-a_1|,|y_i-a_2|]\to\min$$

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

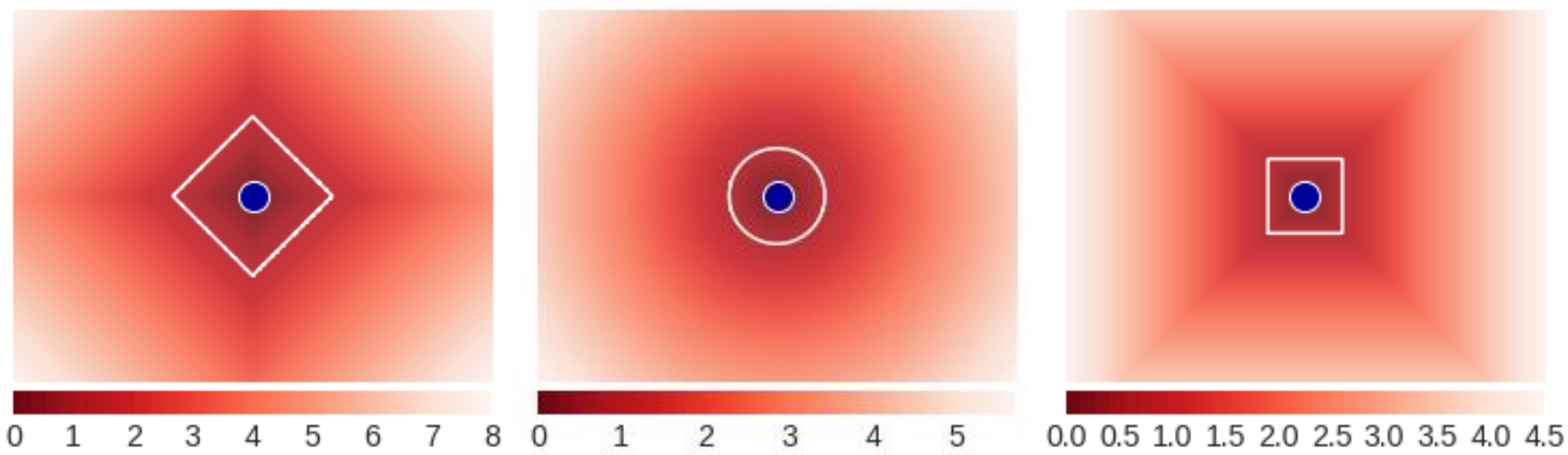


Решаем перебором по точкам выборки!!!

«Степень медианности» – какие функции представлены?



«Степень медианности»



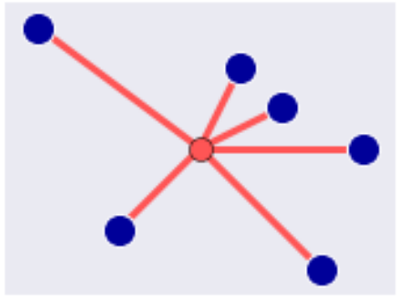
$$\sum_{i=1}^{m} |\, x_i - a_1 \,| + \sum_{i=1}^{m} |\, y_i - a_2 \,| \to \min$$

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

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

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 не коллинеарны

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

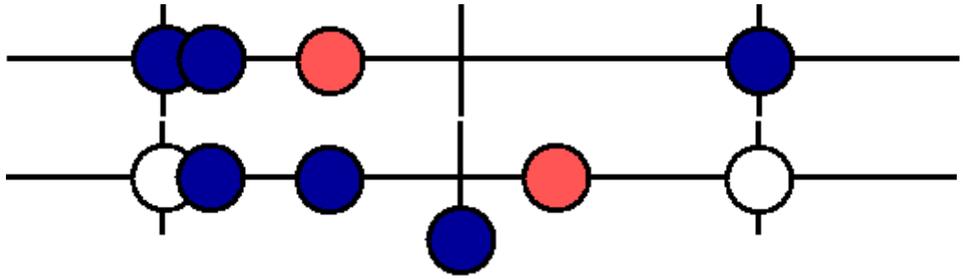
Но можно вычислить с произвольной точностью за почти линейное время

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»

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

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

Для минимизации можно выбрать «что угодно»

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

– 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?

Принстонский эксперимент 1972 года

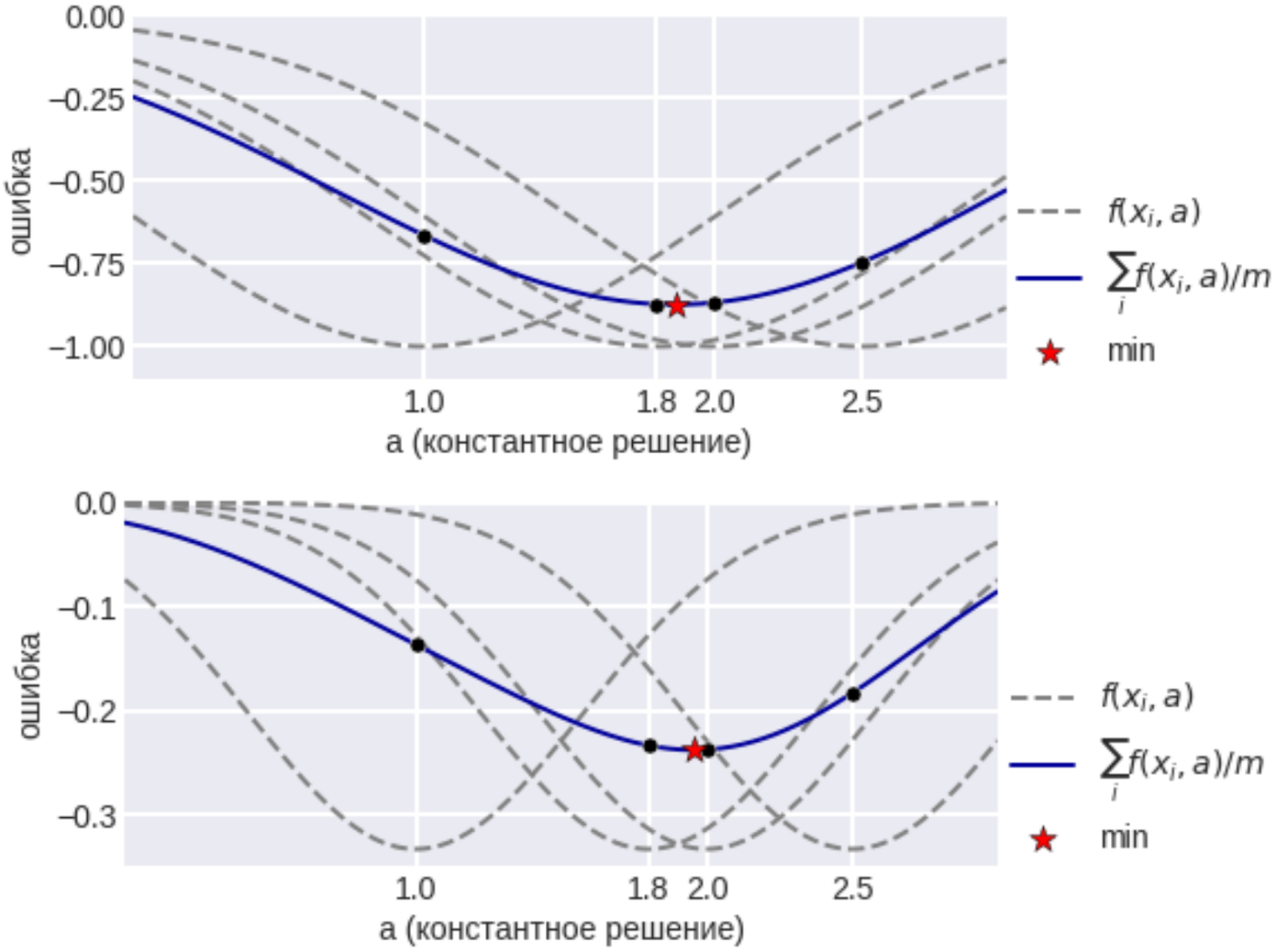
подбор различных функций

Мешалкин Л.Д. (1977) предлагал

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

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

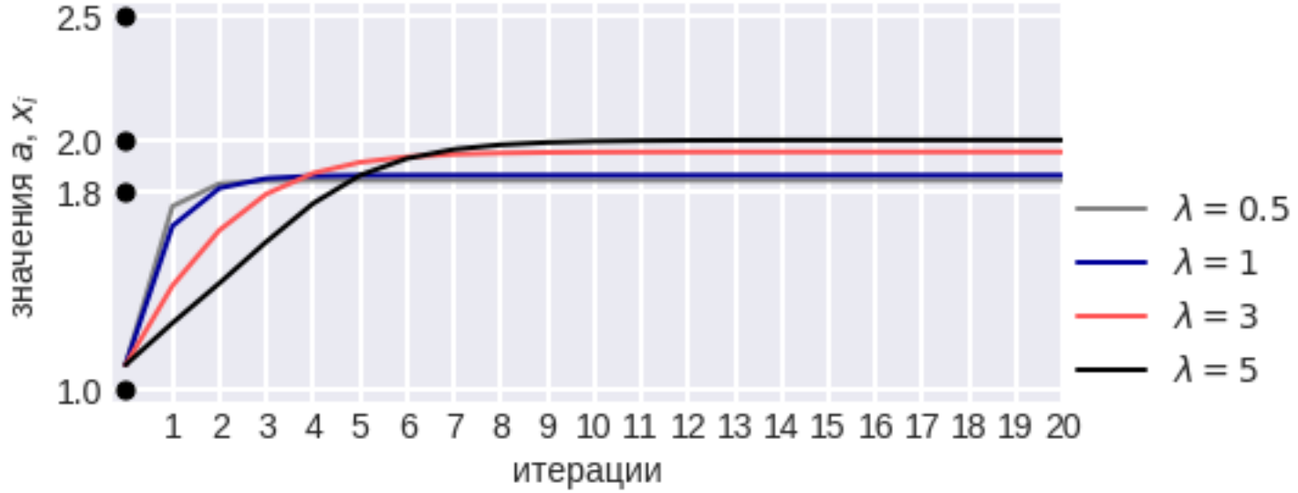
Чем отличаются рисунки?

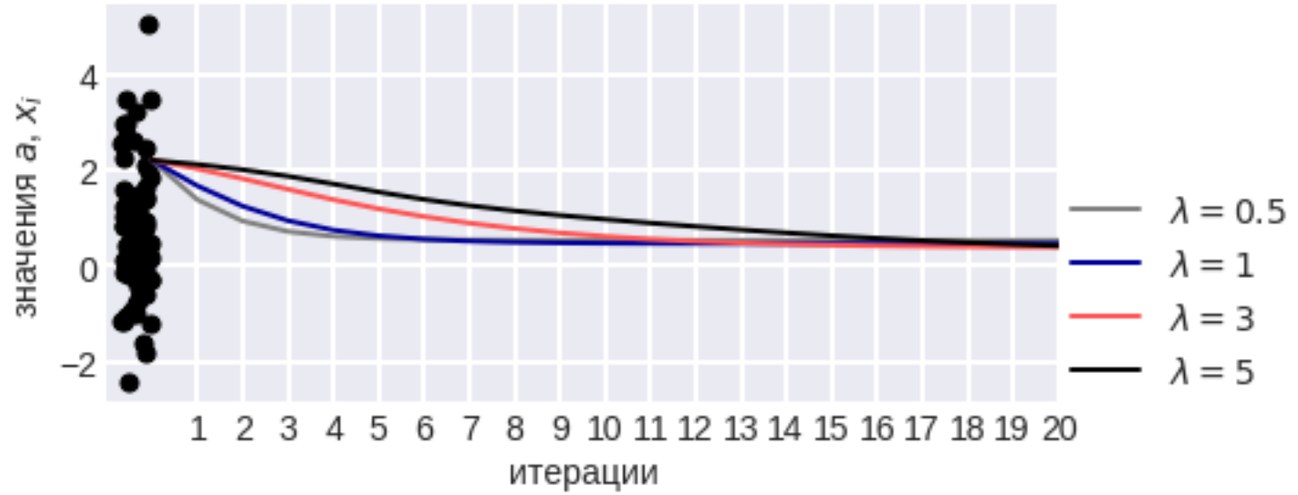


Чем отличаются рисунки?

$\lambda = 1$ $\lambda = 3$

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



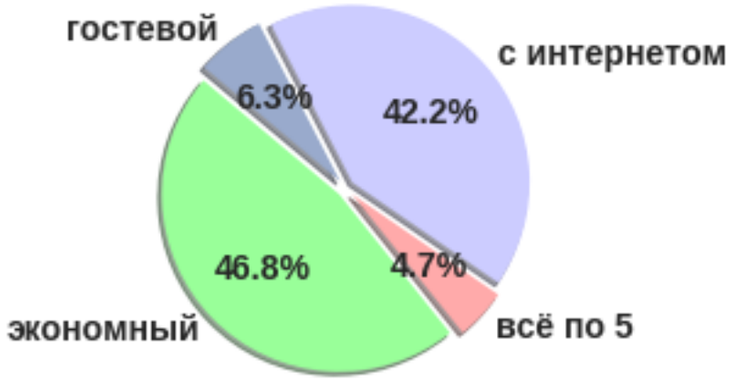


Что важно?

Начальное приближение

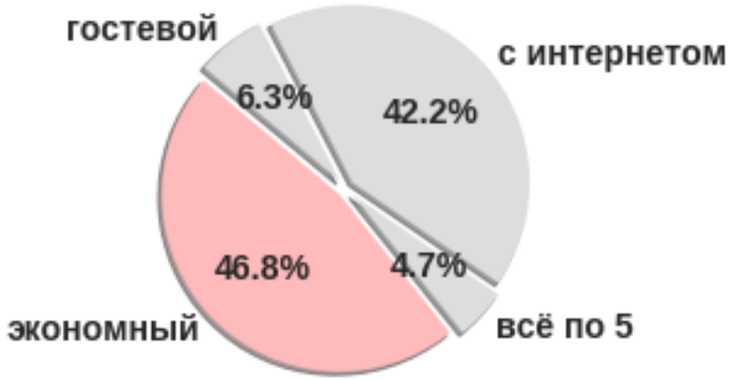
Масштаб

Что такое среднее для номинальных признаков?



Сколько клиентов выбрали определённой тариф сотовой связи

Что такое среднее для номинальных признаков?

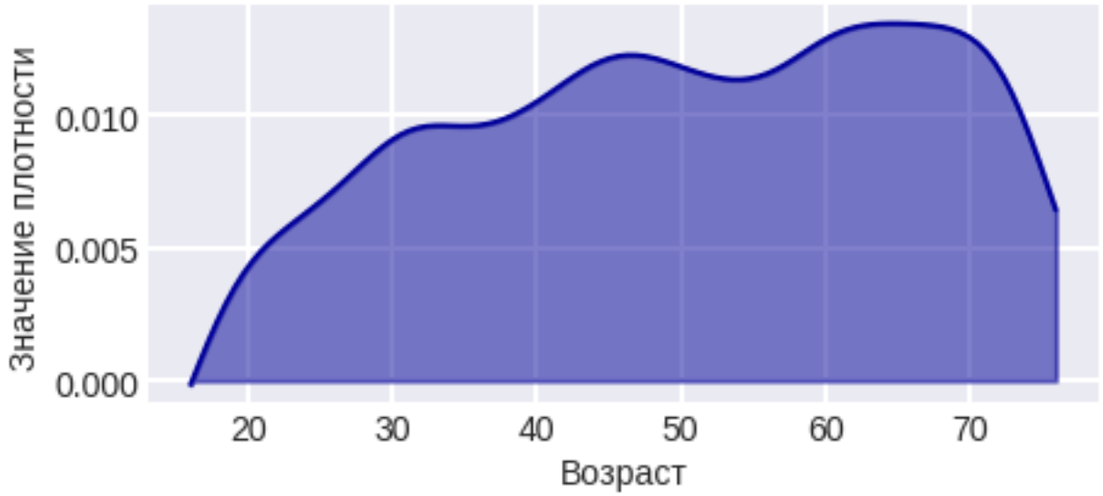


Мода – самое популярное значение

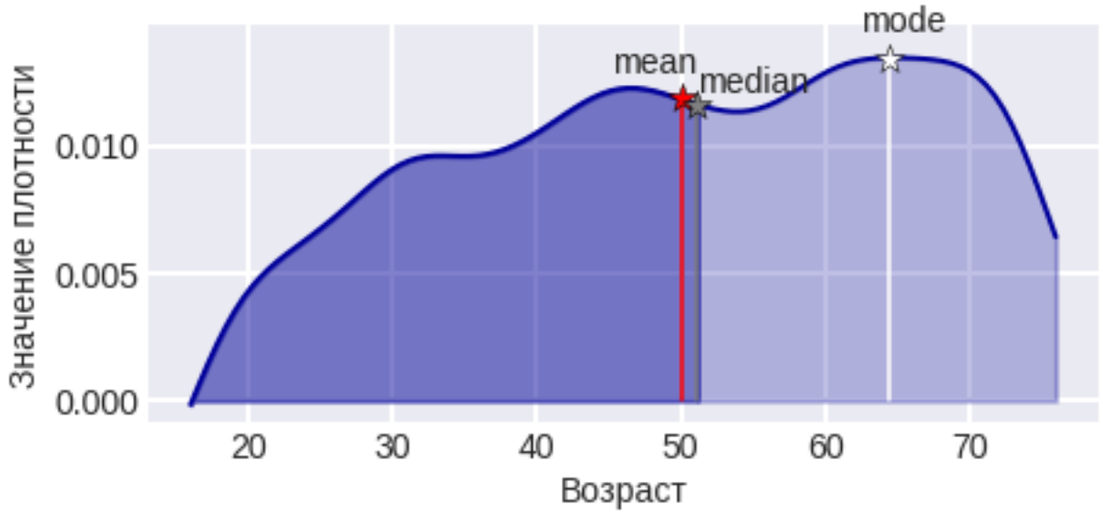
– самое вероятное значение

Что такое среднее для порядковых признаков?

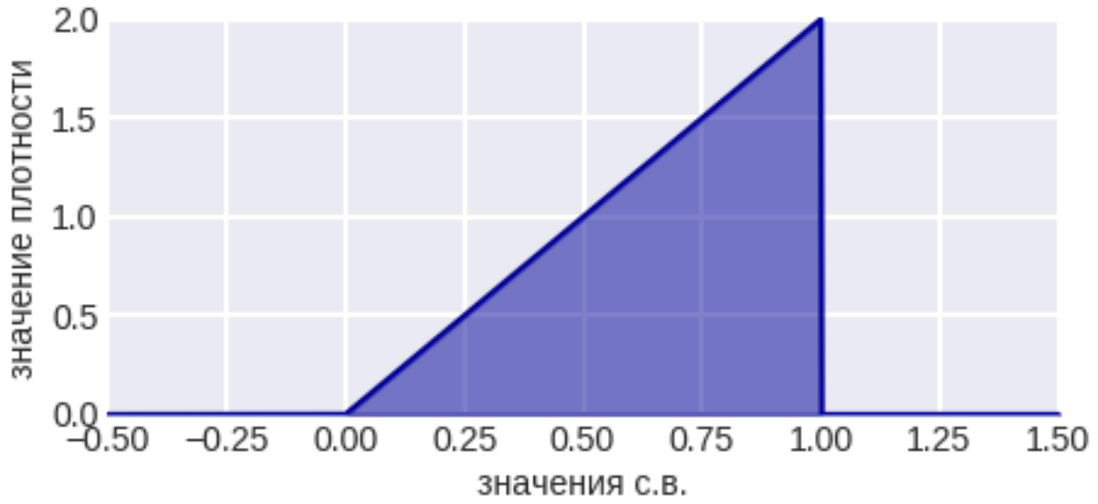
Где матожидание, медиана, мода?



Где матожидание, медиана, мода?

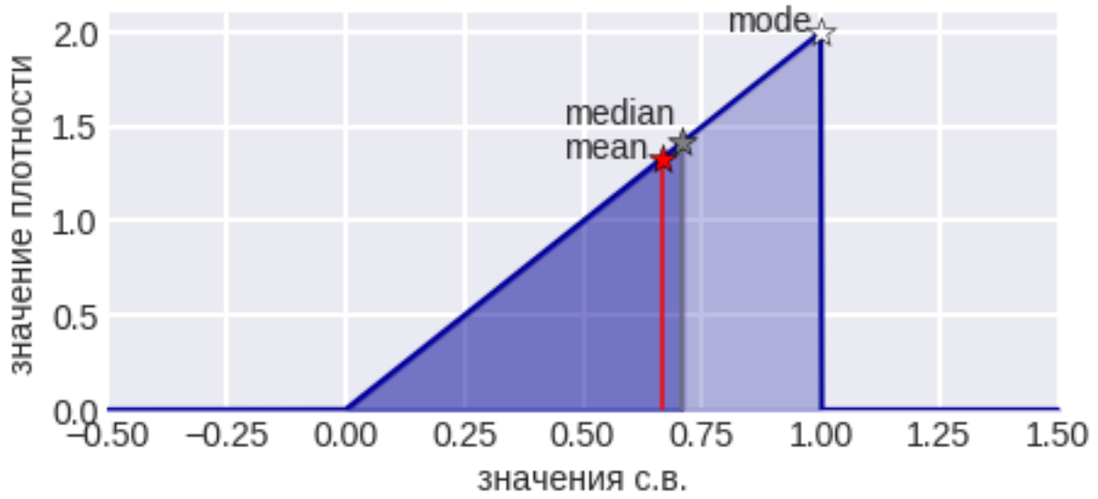


Как запомнить



Где мода, матожидание и медиана?

Как запомнить



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

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

Практика: придумывать не функционал, а среднее

Среднее по А.Н.Колмогорову

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

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?

Тропическое среднее

$$M_\beta(a,b) = \frac{1}{\beta}\ln\left(\frac{\exp(\beta a)+\exp(\beta b)}{2}\right)$$

кстати, с такой суммой и операций умножения «+» получаем ассоциативное кольцо

Крайние случаи – два естественных усреднения:

обычное

$$M_\beta(a,b) \xrightarrow[\beta\to 0]{} \frac{a+b}{2}$$

$$M_\beta(a,b) \xrightarrow[\beta\to +\infty]{} \max(a,b)$$

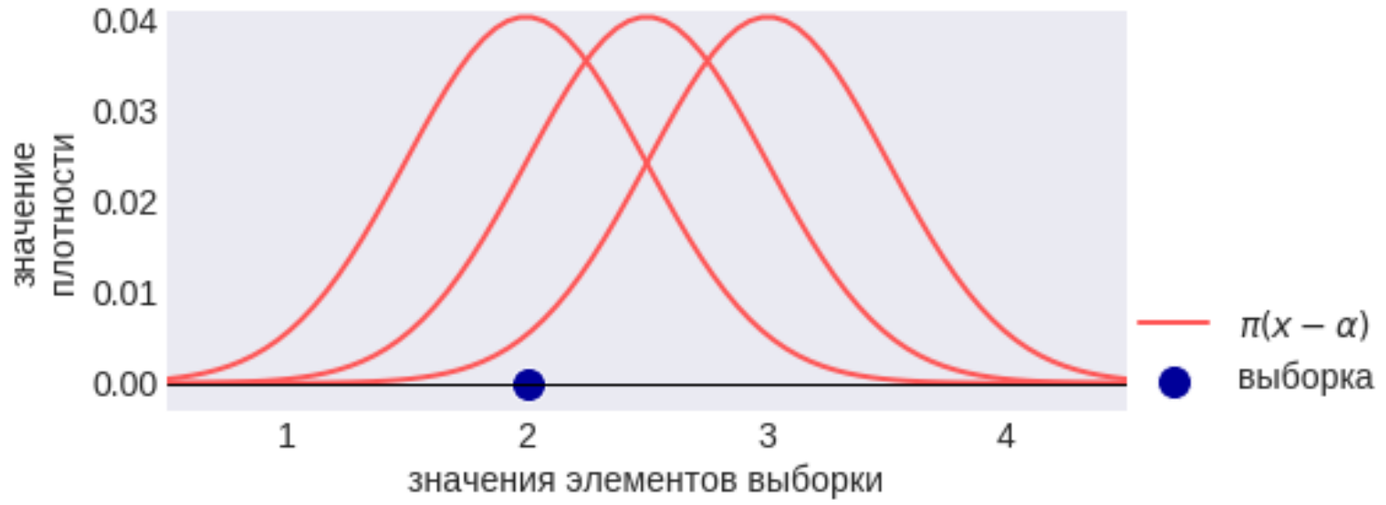
Оценивание вероятности

тоже, в некотором смысле, усреднение… сейчас объясним

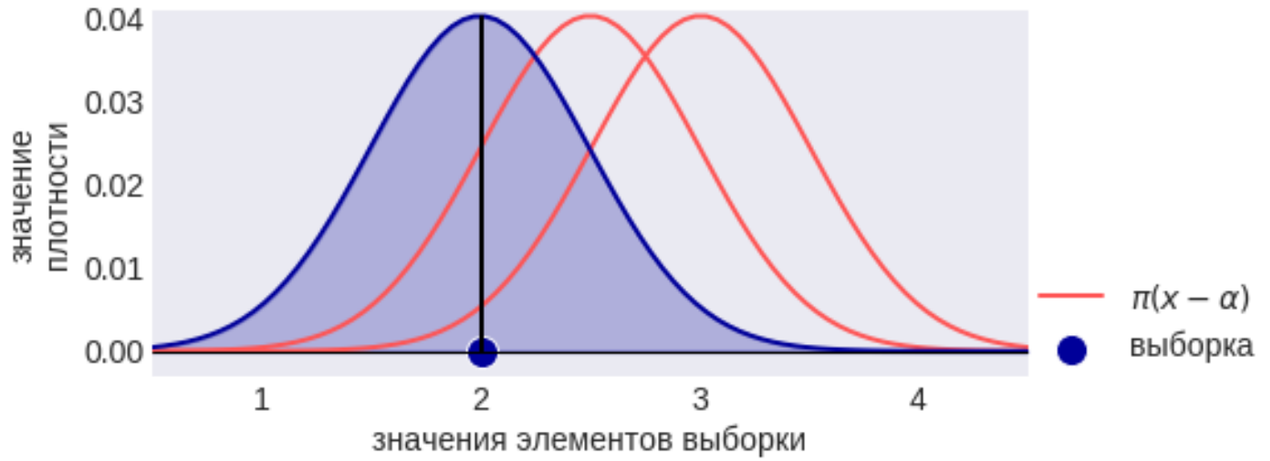
Метод максимального правдоподобия

Есть выборка $x_1, \ldots, x_m$ **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 $m = 2$

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

Почему произведение?

Общий случай:

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

Как максимизируют?

Случай биномиального распределения

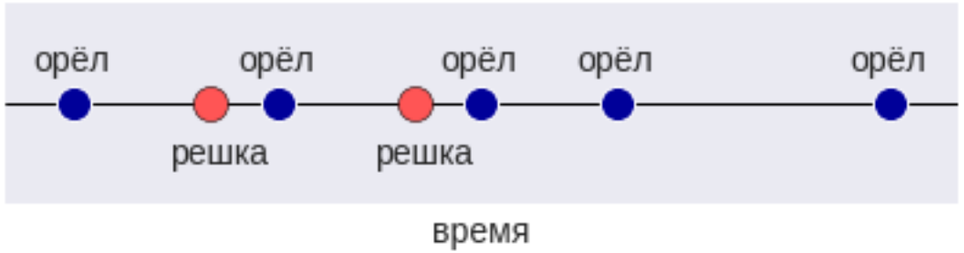
$$\pi_p(x) = \begin{cases} p, & x = 1, \\ 1-p, & x = 0. \end{cases}$$

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

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

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

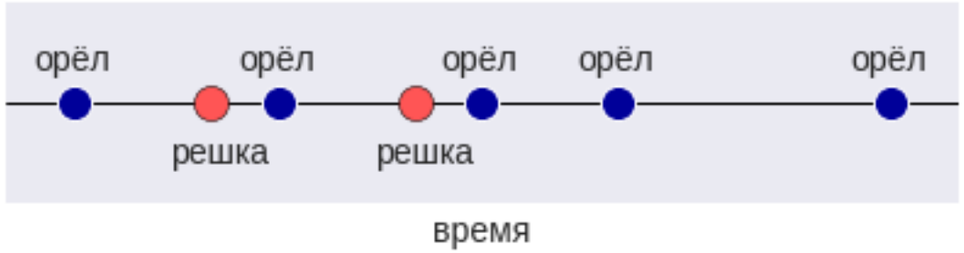
Самый очевидный ответ для оценки вероятности!



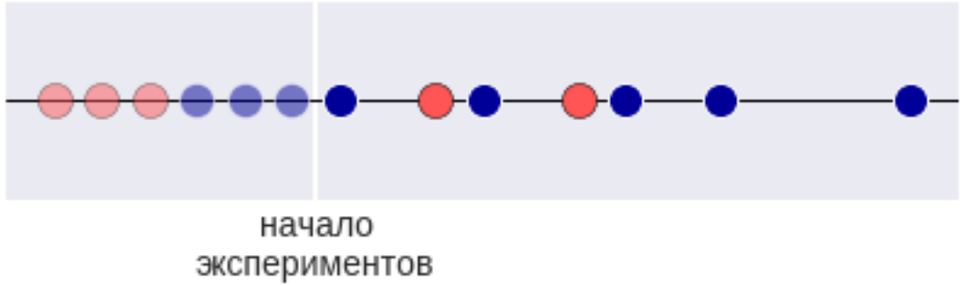
$$p = \frac{5}{5+2} = \frac{5}{7} \approx 0.71$$

Оценивание вероятности – сглаживание Лапласа

тоже, в некотором смысле, усреднение



на практике есть априорная вероятность



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

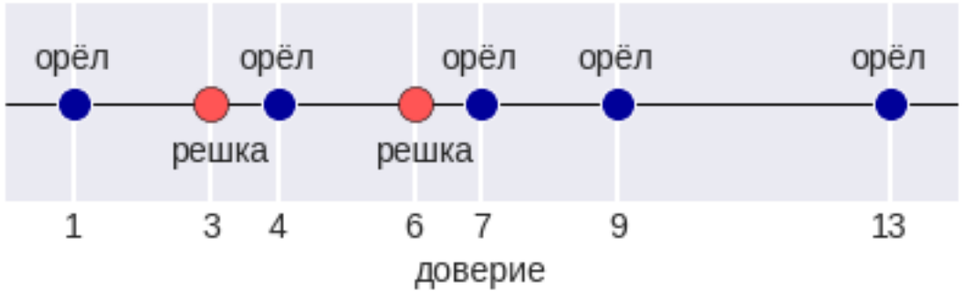
Есть разные эвристические методы

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

какую весовую функцию выбрать?

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)**

Международное соревнование «dunnhumby's Shopper Challenge»

Дано: статистика визитов

Предсказать: день первого визита + сумму покупки с точностью до 10 $

покупатель, дата визита, сумма
56, 2011-06-30, 35.01
56, 2011-06-08, 35.17
56, 2011-07-10, 24.12
56, 2011-07-12, 7.73
57, 2011-05-13, 29.38
57, 2011-05-19, 41.00
…

> 100000 клиентов customers

T = 1 год

http://www.kaggle.com/c/dunnhumbychallenge/

Международное соревнование «dunnhumby's Shopper Challenge»

Статистика визитов одного клиента:

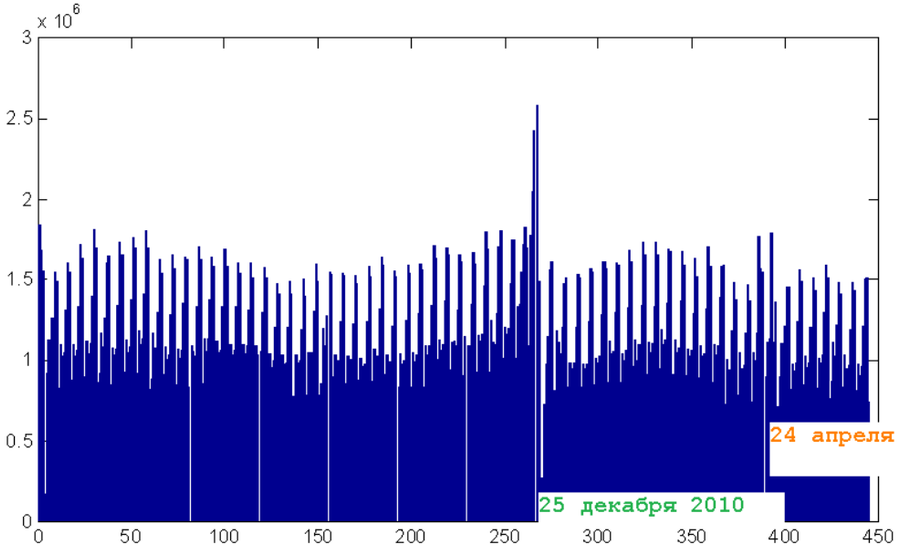
рт	Март 22	Март 23	Март 24	Март 25	Март 26	Март 27	Март 28	Март 29	Март 30	Март 31	Апрель 1	Апрель 2	Апр 3
5$		45$	5$				35$		60$		?	?	?

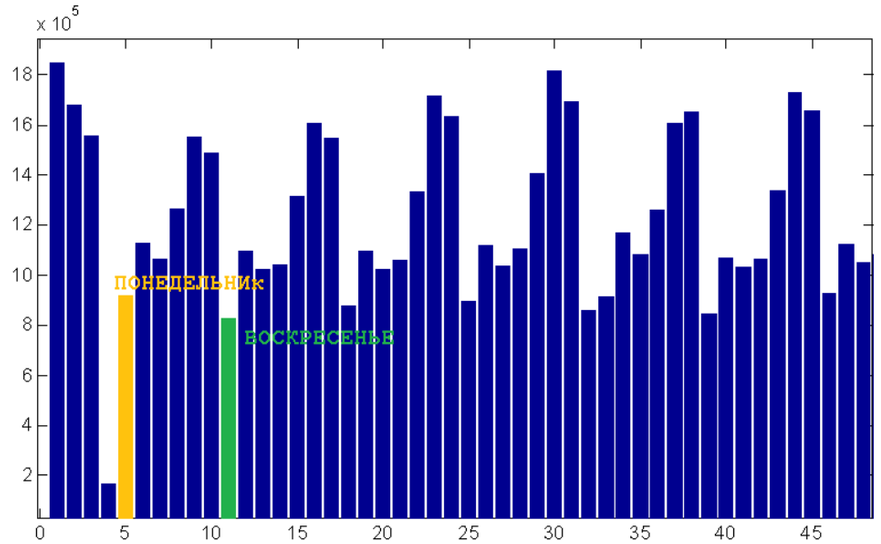
Опишем лучший алгоритм из 287

#	Team Name	$10,000 • 279 teams	Score	Entries
1	D'yakonov Alexander (MSU, Moscow, Russia) *		18.83	68
2	NSchneider *		18.67	20
3	Ben Hamner *		18.57	19
4	William Cukierski		18.44	75

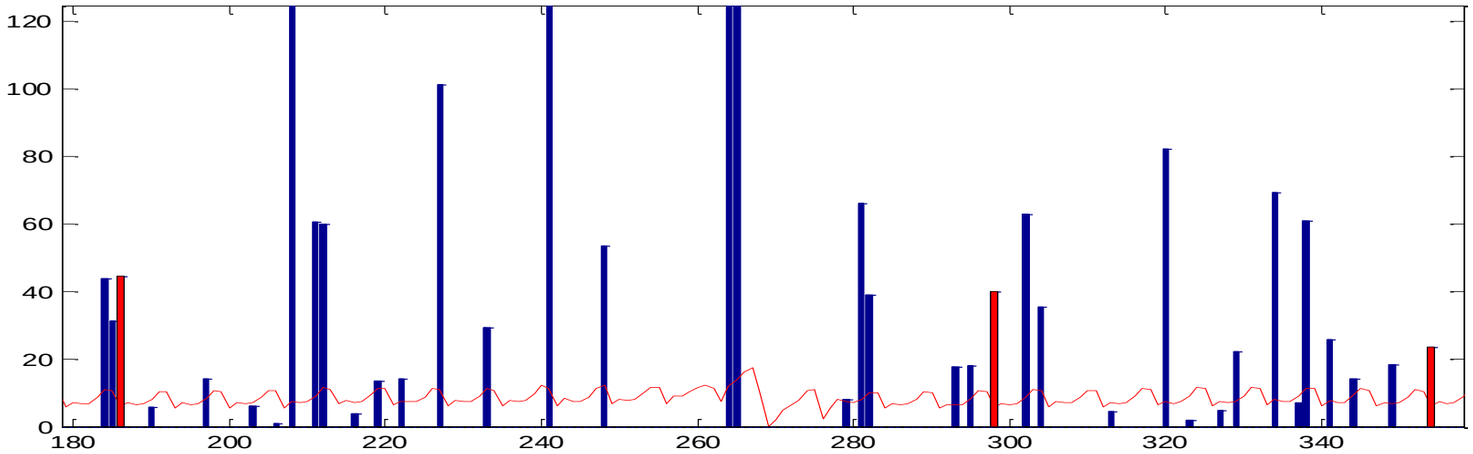
Агрегированная статистика всегда лучше

Суммы покупок всех клиентов





Покупки одного клиента

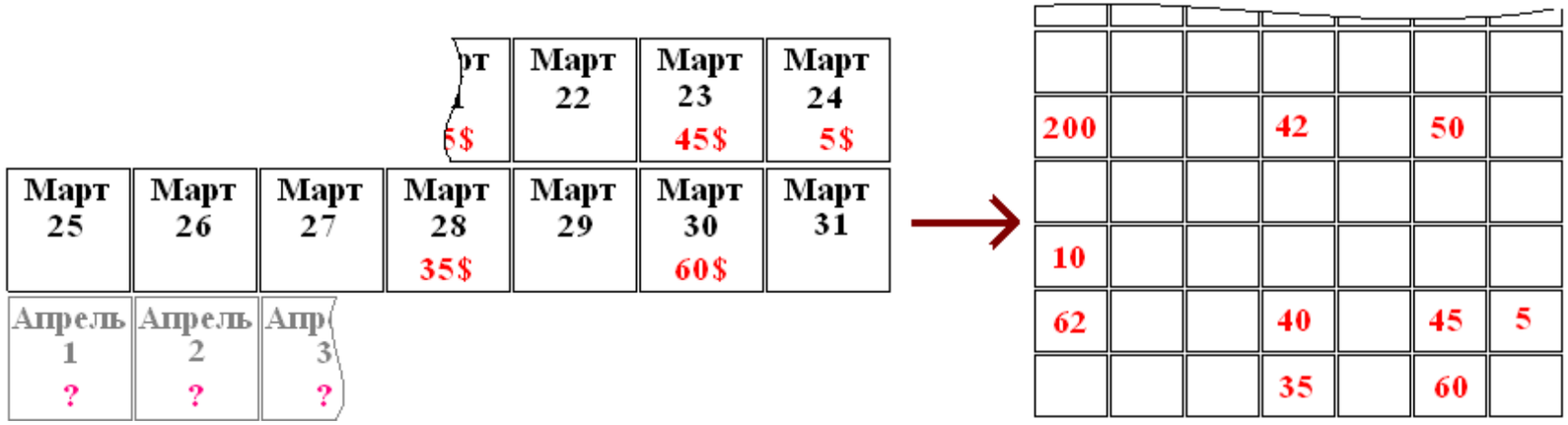


Предположения

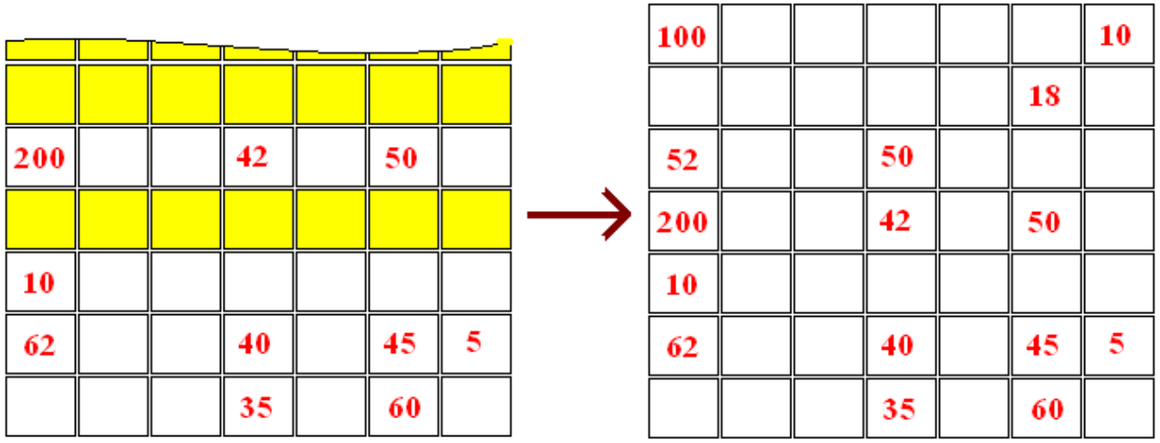
Все клиенты независимы

Будем анализировать каждого клиента отдельно

Разбиение на недели



Матрица разбивки по неделям



Сработало устранение пустых недель…

Вероятностная модель поведения клиента

Матрица затрат: $S = \| s_{ij} \|_{d \times 7}$

Матрица визитов: $V = \| v_{ij} \|_{d \times 7}$, $v_{ij} = 1 \quad \Leftrightarrow \quad s_{ij} > 0.$

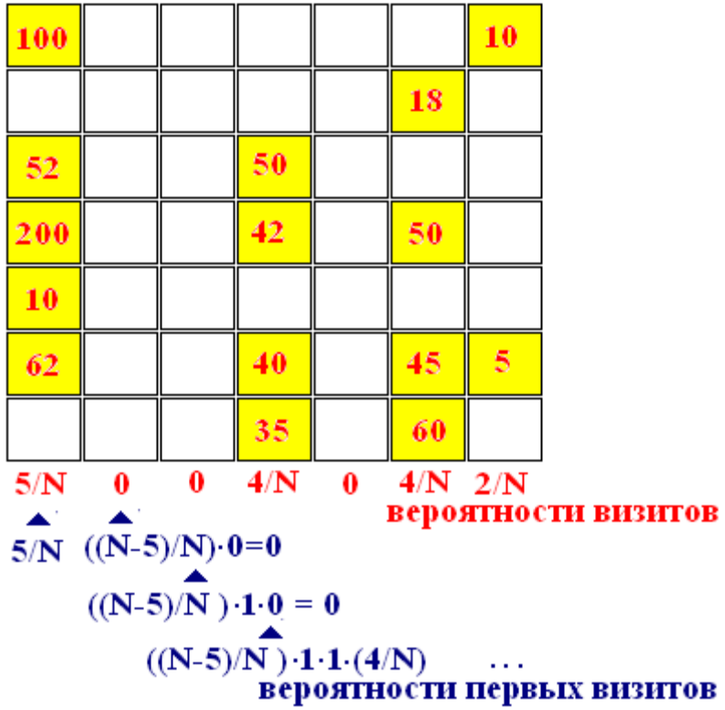
Вероятности визитов

оценки вероятностей…

$$p_1$$
$$p_2$$
$$\dots$$
$$p_7$$

первых визитов

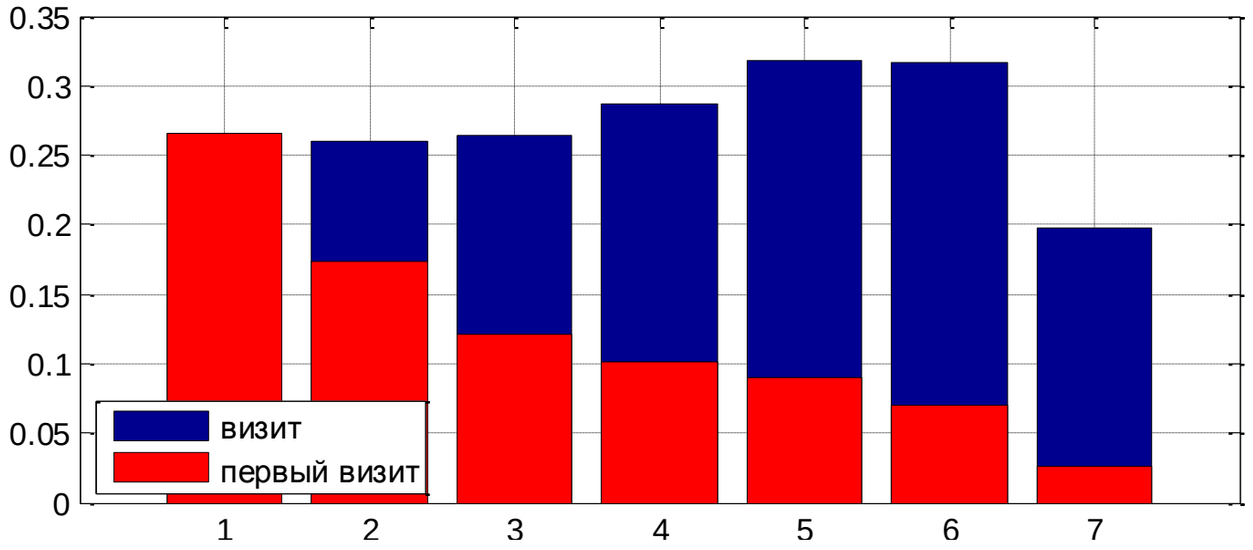
$$\tilde{p}_1 = p_1$$
$$\tilde{p}_2 = (1 - p_1)p_2$$
$$\dots$$
$$\tilde{p}_7 = \prod_{i=1}^{6}(1 - p_i)p_7$$



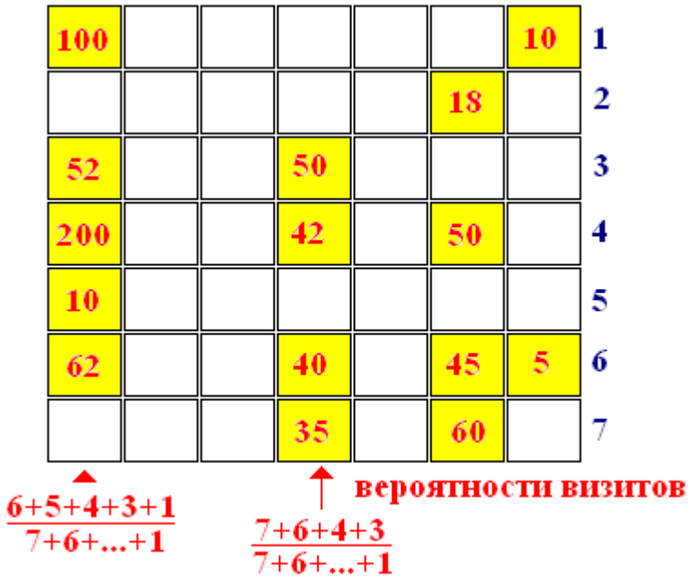
Находим максимум вероятностей!

Предположение: **Каждый клиент обязательно посетит магазин в течение следующей недели.**

Процент визитов и первых визитов на неделе



«Более свежие» данные о клиенте важнее устаревших



Весовые схемы!

Взвешенная схема оценки вероятности

$$p_j = \sum_{i=1}^{d} w_i v_{ij},$$

$$w_1 \geq w_2 \geq \ldots \geq w_d \geq 0, \sum_{i=1}^{d} w_i = 1.$$

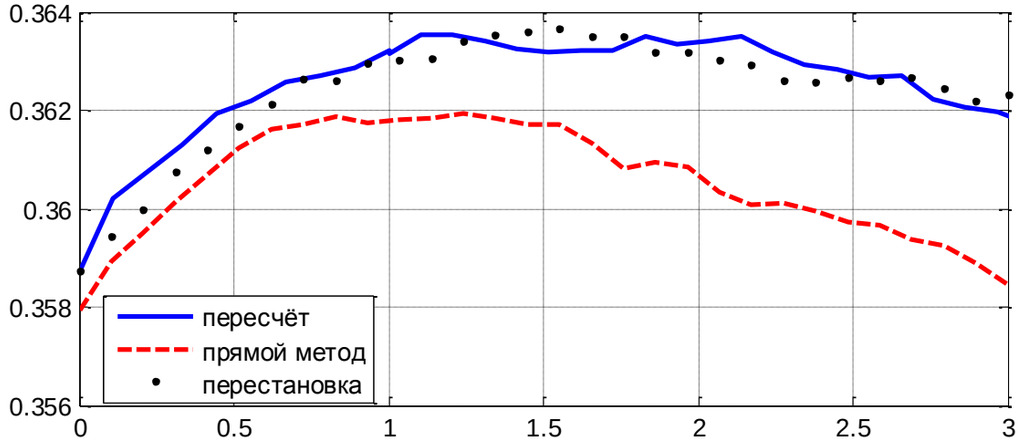
Способы

$$w_i^{\mathrm{N}} = \left(\frac{d-i+1}{d} \right)^{\delta}, i \in \{1,2,\ldots,d\},$$

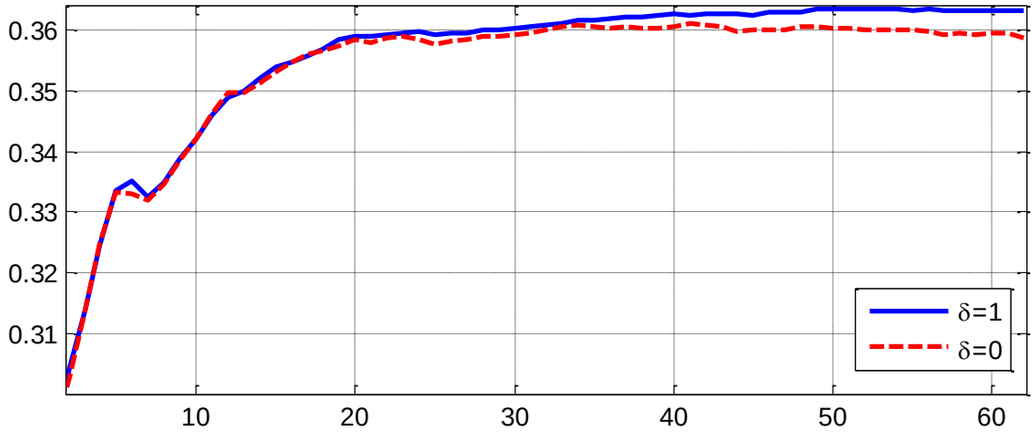
$$w_i = \frac{w_i^{\mathrm{N}}}{\sum_{i=1}^{d} w_i^{\mathrm{N}}}, i \in \{1,2,\ldots,d\}.$$ [просто нормировка]

Параметр $\delta \in [0,+\infty)$.

Веса – от равномерных к «агрессивным»

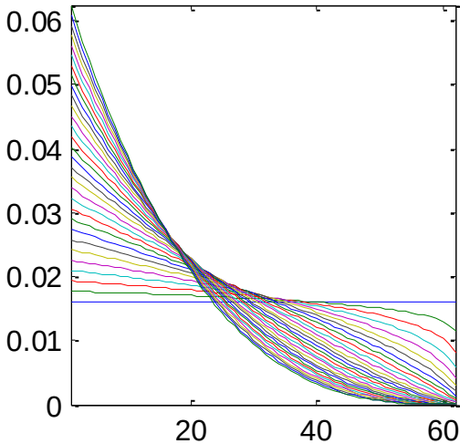


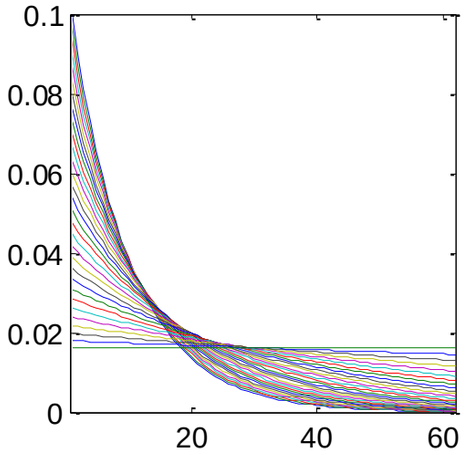
Зависимость качества прогноза от степени δ

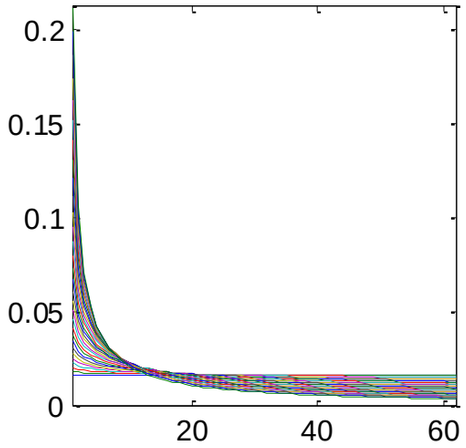


Зависимость качества прогноза от числа учитываемых недель

Три разные весовые схемы







вес недели в зависимости от её номера

$$w_i^{\mathrm{N}} = \left(\frac{d-i+1}{d}\right)^{\delta}$$

$$\delta \in [0,+\infty)$$

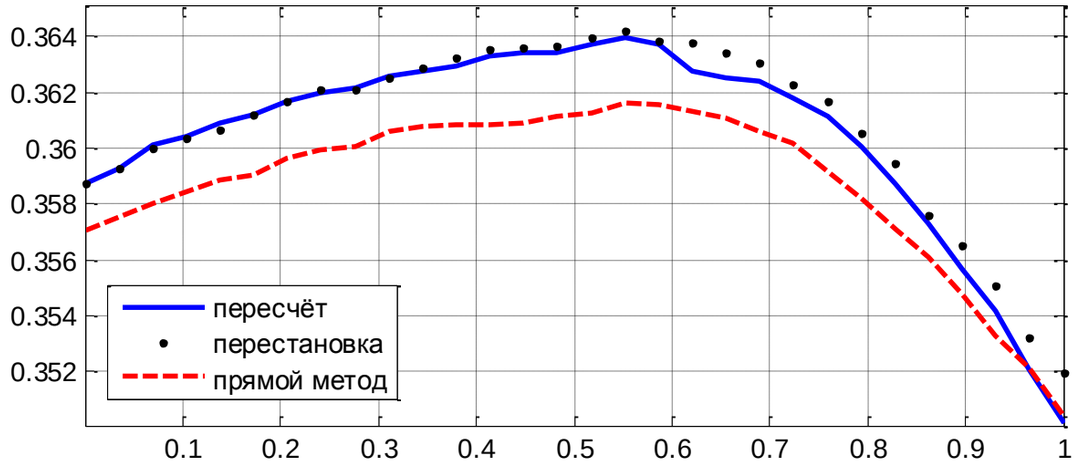
$$w_i^{\mathrm{N}} = \lambda^i$$

$$\lambda \in (0,1]$$

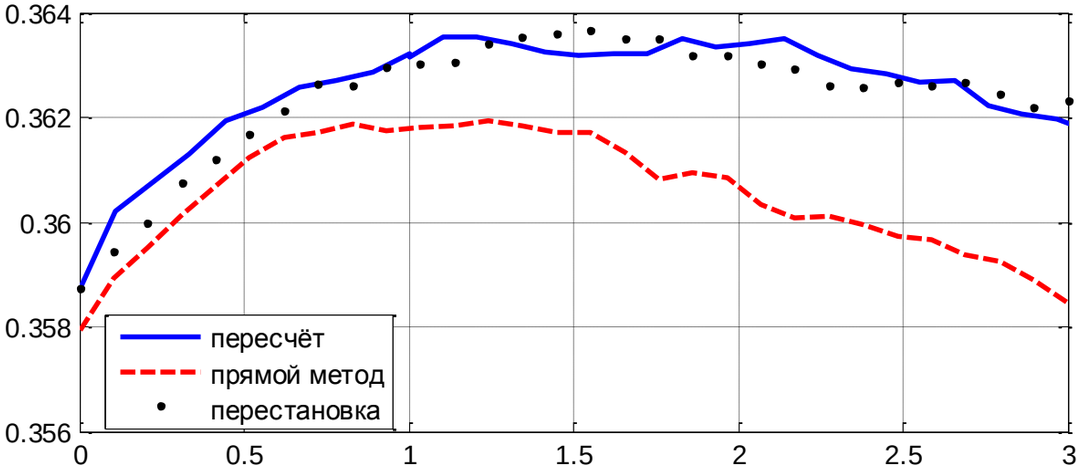
$$w_i^{\mathrm{N}} = \frac{1}{i^{\gamma}},$$

$$\gamma \in [0,+\infty)$$

Принципиально всё одинаково…



Третья весовая схема



Первая весовая схема

Два способа оценки вероятности первого визита

Прямой метод

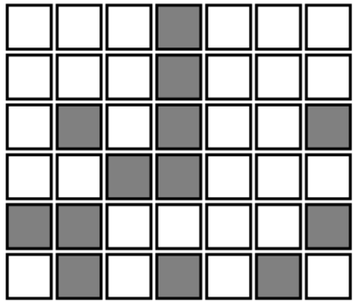
$$\tilde{p}_j^2 = \frac{1}{d} \left| \{ i \in \{1,2,\ldots,d\} \ : \ v_{i1} = \ldots = v_{i,j-1} = 0, v_{ij} = 1 \} \right|$$

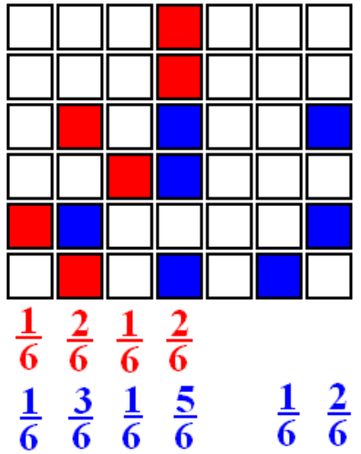
Более естественный, но хуже!

матрица первых визитов

$$V' = \| v'_{ij} \|_{d \times 7}$$

$$\tilde{p}_j^2 = \sum_{i=1}^{d} w_i v'_{ij}$$





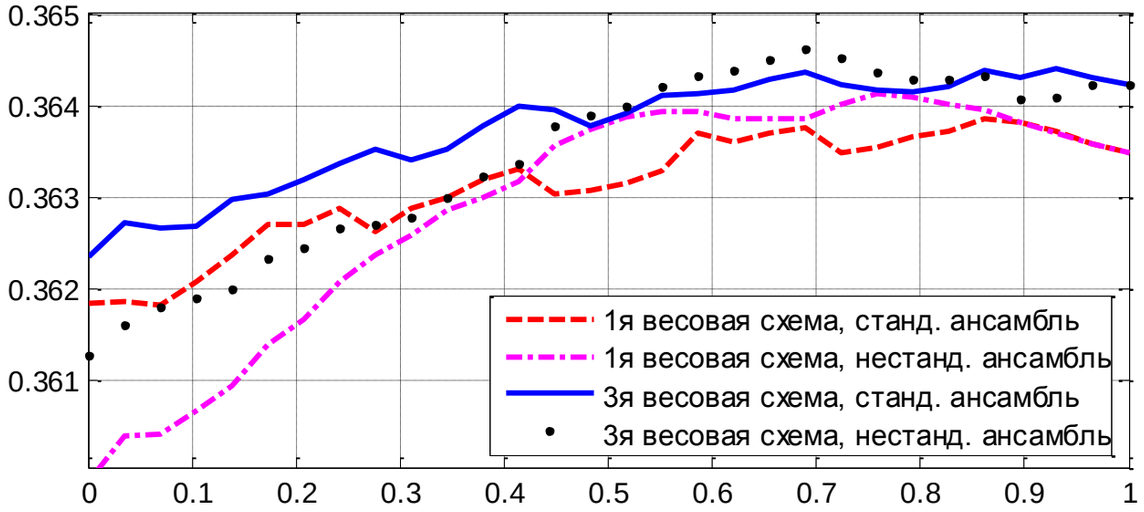
Ансамблирование

«Стандартный ансамбль» – взять выпуклую комбинацию:

$$\tilde{p}_j = \alpha \tilde{p}_j^1 + (1-\alpha)\tilde{p}_j^2, \quad \alpha \in [0,\ 1].$$

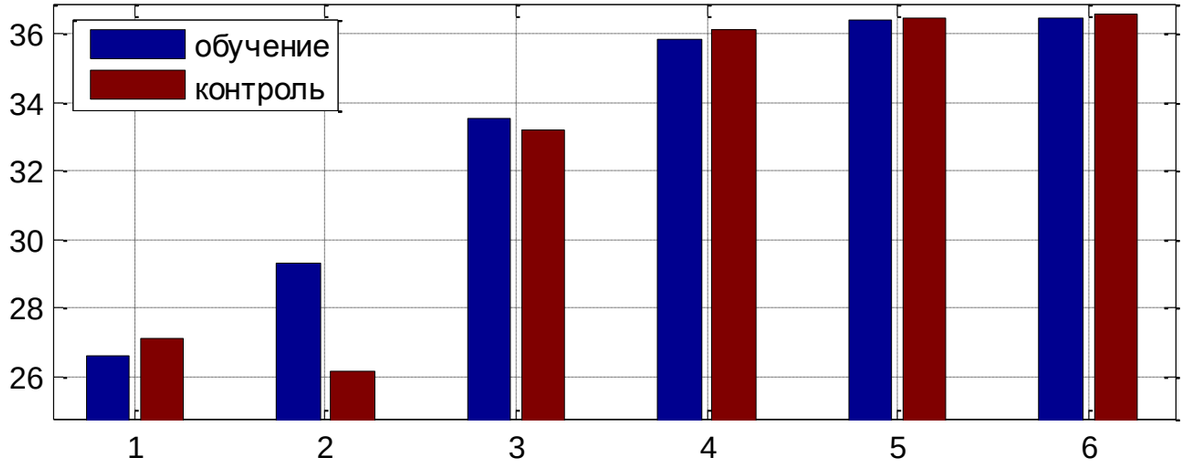
«Нестандартный ансамбль»

$$\alpha p_j + (1-\alpha)\tilde{p}_j^2 = \alpha \sum_{i=1}^{d} w_i v_{ij} + (1-\alpha)\sum_{i=1}^{d} w_i v'_{ij} = \sum_{i=1}^{d} w_i (\alpha v_{ij} + (1-\alpha) v'_{ij})$$



Качество ансамблирования от параметра $\alpha \in [0,\ 1]$

Про переобучение



Качество на обучении и отложенном контроле для шести алгоритмов

1. Константный («клиент придёт на следующий день»),
2. Визит клиента как на прошлой неделе,
3. Вероятности (*) оценены по последним 5 неделям,
4. Вероятности оценены по всем неделям,
5. Оптимальные значения весов,
6. Оптимальное нестандартное ансамблирование.

Не усложнение, а сглаживание!

Итог

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

среднее

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

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

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