Практические задачи анализа данных

Лекция 2. Оценки среднего, вероятности, весовые схемы

Московский авиационный институт
«МАИ»

15 сентября 2021 г.

Что такое среднее?

средний, типичный, среднестатистический...

Естественная формализация — среднее арифметическое

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

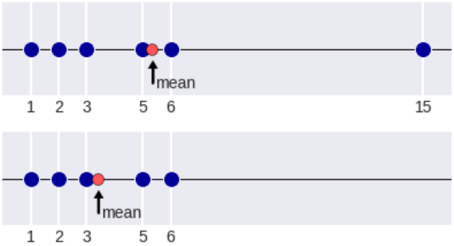
Какие плюсы и минусы?

Что такое среднее

Среднее арифметическое

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

1) Проблема выбросов



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?

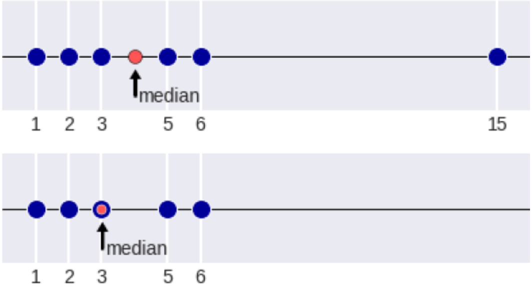
– Хочется конкретный пример!

Что такое среднее

Решение проблемы — медиана, для $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(m)}$:

$$\mathrm{median}(X) = \begin{cases} x_{\left(\frac{m+1}{2}\right)}, & \text{m — нечетное} \\ \frac{1}{2}\left[x_{\left(\frac{m}{2}\right)} + x_{\left(\frac{m+1}{2}\right)}\right], & \text{m — четное} \end{cases}$$

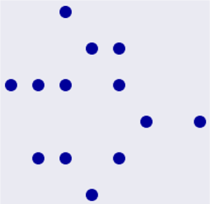
1. устойчива к выбросам
2. является (можно сделать) точкой выборки

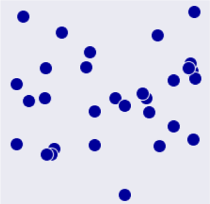


Что такое среднее

Проблема медианы

Что такое многомерная медиана?





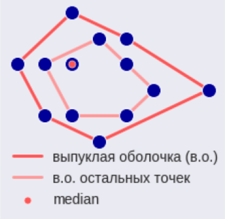
Хочется (может быть) инвариантность к

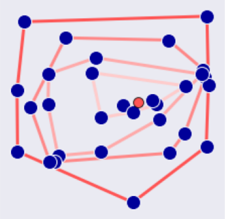
- движениям
  - поворотам
  - сдвигам (параллельным переносам)
- сжатиям / растяжениям

В одномерном случае должна совпадать с median

Что такое среднее

Многомерная медиана как результат итерационного процесса





Выход: сделать аналогичный процесс построения, как в одномерном случае

удаление крайних элементов!

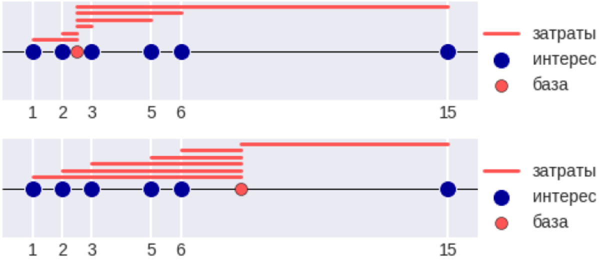
Многомерная медиана

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

Всё равно можно применить подход «отбрасывания крайних элементов»

Среднее как решение оптимизационной задачи

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



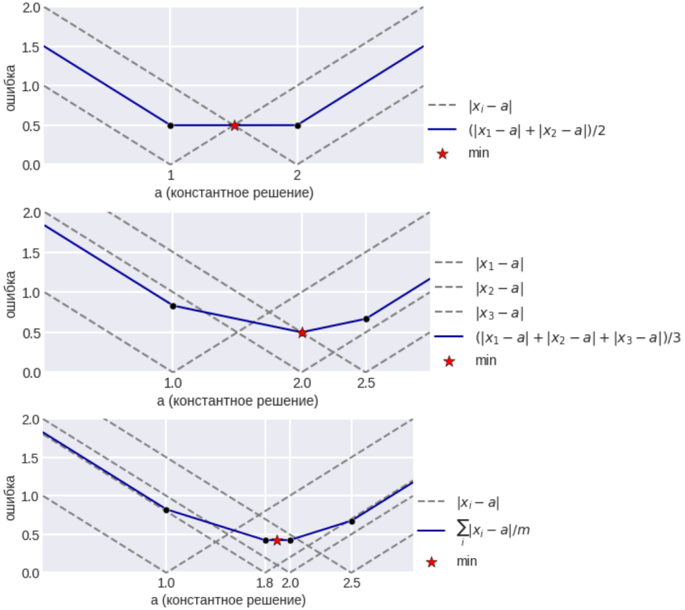
Что такое среднее

Если суммарные затраты

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

то решение — медиана

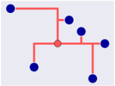
Что такое среднее



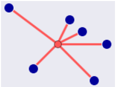
Медиана в пространстве

2-й способ формализации: аналогично минимизируем затраты, но тут может быть зависимость от координат

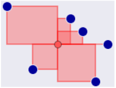
$$\sum_{i=1}^{m} \left(|x_i - a_1|^d + |y_i - a_2|^d\right)^{1/d} \to \min_{a_1, a_2}$$



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



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

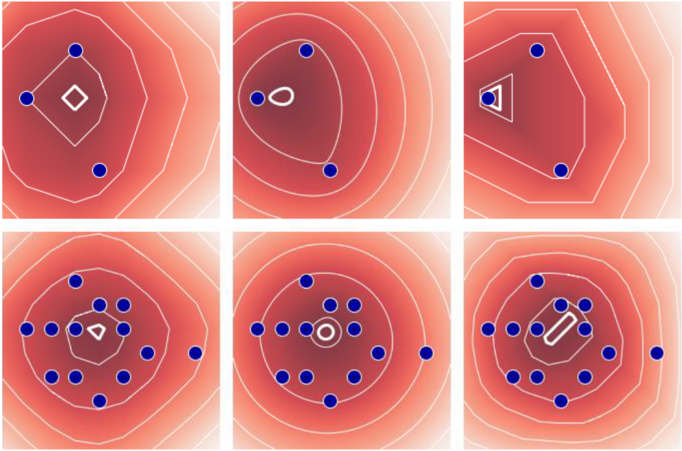


$$\sum_{i=1}^{m} |x_i - a_1| \cdot |y_i - a_2| \to \min_{a_1, a_2}$$

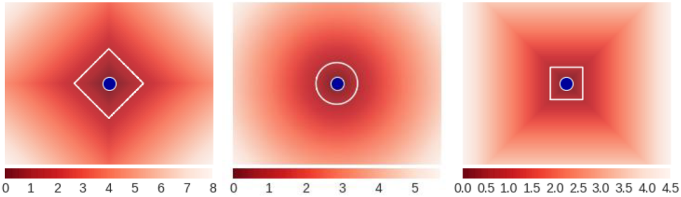
Решаем перебором по точкам выборки

Что такое среднее

«Степень медианности» — какие функции представлены?



«Степень медианности»



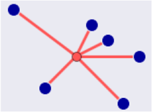
$$\sum_{i=1}^{m} |x_i - a_1| + \sum_{i=1}^{m} |y_i - a_2| \to \min_{a_1, a_2}$$

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

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

Геометрический центр

пространственная медиана или точка Торричелли



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

Геометрический центр единственный, когда точки не находятся на одной прямой

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

Можно вычислить с произвольной точностью за почти линейное время

Алгоритм Вайсфельда Геометрический центр

Эвристический способ борьбы с выбросами

$$a = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad (*)$$

Алгоритм Шурыгина

1. Если $m \leq 2$, то используем формулу $(*)$. Выход
2. Пусть $x_1 \leq x_2 \leq \ldots \leq x_m$ (без ограничения общности)
3. Если $\dfrac{x_1 + x_m}{2} \leq x_2$, то удаляем из выборки x_1. Переходим к п. 1 (с соответствующей перенумерацией объектов)
4. Если $\dfrac{x_1 + x_m}{2} \geq x_{m-1}$, то удаляем из выборки x_1. Переходим к п. 1 (с перенумерацией объектов)
5. Исключаем из выборки x_1, x_m, но добавляем в неё $\dfrac{x_1 + x_m}{2}$. Переходим к п. 1 с перенумерацией

Что минимизирует «среднее»

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

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

Для минимизации можно выбрать «что угодно»

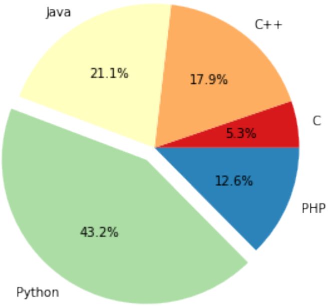
$$\mathrm{mid}(X) = \operatorname*{arg\,min}_a \sum_{i=1}^{m} f(x_i, a)$$

— оценка минимального контраста

см. *Шурыгин А.М. Прикладная стохастика: робастность, оценивание, прогноз.– М.: Финансы и статистика, 2000.*

Что такое среднее

Что такое среднее для номинальных признаков?



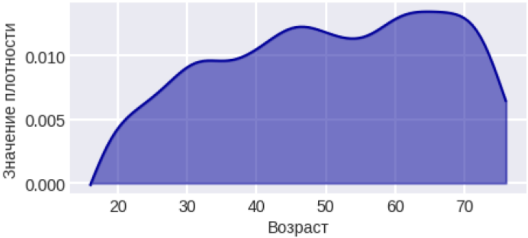
Мода — самое популярное значение
— самое вероятное значение

Что такое среднее для порядковых признаков?

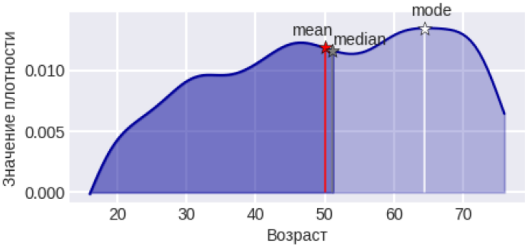
Кто здесь типичный представитель?

Что такое среднее

Где матожидание, медиана, мода?

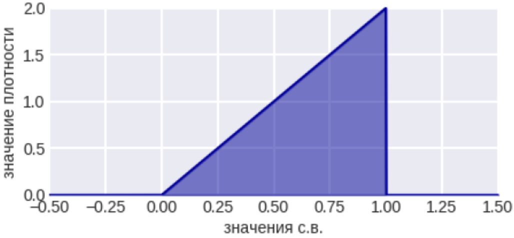


Где матожидание, медиана, мода?

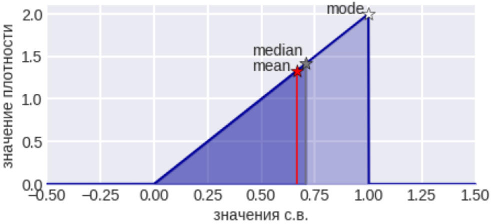


Что такое среднее

Где матожидание, медиана, мода?



Где матожидание, медиана, мода?



$$\mathrm{M}[X] = \int_0^1 x \cdot 2x dx = \frac{2}{3} x^3 \Big|_0^1 = \frac{2}{3}$$

$$\int_0^{\text{median}} 2x dx = (median)^2 = \frac{1}{2} \quad \Rightarrow \quad \text{median} = \frac{\sqrt{2}}{2} \approx 0.71$$

ДЗ: Какие порядки вообще могут быть?
насколько среднее и медиана могут отличаться?

Что такое 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?

Что такое среднее

Медиана и моду получить нельзя

среднее по Коши: любое значение из отрезка

$$\left[x_{(1)}, x_{(m)}\right],$$

$$x_{(1)} = \min\{x_1, \ldots, x_m\}, \; x_{(m)} = \max\{x_1, \ldots, x_m\}$$

Что такое среднее

Тропическое среднее

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

Крайние случаи — два естественных усреднения:

$$\lim_{\beta \to 0} M_\beta(a,b) = \frac{a+b}{2}$$

$$\lim_{\beta \to \infty} M_\beta(a,b) = \max(a,b)$$

Как обобщить на случай выборки?

Тропическая геометрия

Оценивание вероятности

тоже, в некотором смысле, усреднение
Требуется найти вероятность случайного события A по наблюдениям $\{X_1, \dots, X_m\}$ за СВ X:

$$X = \begin{cases} 1, & \text{если } A \\ 0, & \text{если } \overline{A} \end{cases}$$

т.е. $X \sim Be(p)$, где $p = \mathrm{P}(A)$.
Тогда из метода максимального правдоподобия:

$$L(p, X_1, \dots, X_n) = \prod_{i=1}^{m} p^{X_i}(1-p)^{1-X_i}$$

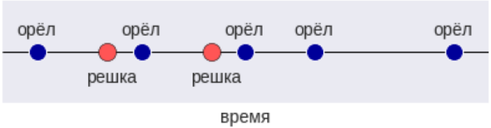
$$\ln L(\cdot) = \sum_{i=1}^{m} (X_i \ln p + (1 - X_i) \ln(1-p))$$

Оценка вероятности

$$\frac{\partial \ln L(\cdot)}{\partial p} = \frac{1}{p}\sum_{i=1}^{m} X_i + \frac{1}{1-p}\sum_{i=1}^{m}(1 - X_i) = 0$$

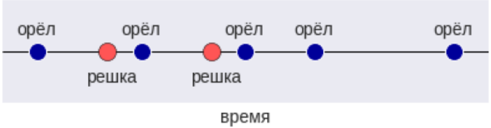
$$p^* = \frac{1}{m}\sum_{i=1}^{m} X_i$$

Если обозначить n — количество 1 среди $\{X_1, \ldots, X_m\}$, то получим $p^* = \dfrac{n}{m}$ — частота случайного события A

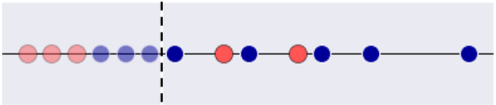


$$\widehat{P}(«\text{орёл}») = \frac{5}{5+2} \approx 0.71 \text{ — самый очевидный ответ}$$

Оценивание вероятности — сглаживание Лапласа



на практике есть априорная вероятность



$$p^* = \frac{n + \lambda \cdot p}{m + \lambda} = \frac{5 + 6 \cdot 0.5}{7 + 6} \approx 0.62$$

Оценка вероятности

Есть разные эвристические методы

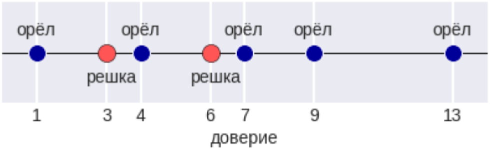
$$\sigma(m)\frac{n}{m} + (1 - \sigma(m))p$$

какую весовую функцию выбрать?

ДЗ: Придумать и обосновать подобные функции

Вторая особенность практики

Не все эксперименты равнозначны, например, недавние события важнее



Весовая схема

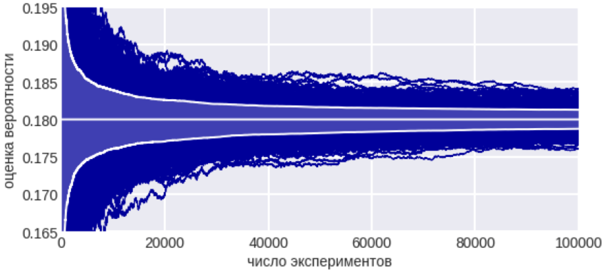
$$\frac{w_{i_1} + \ldots + w_{i_n}}{w_1 + \ldots + w_m} = \frac{1+4+7+9+13}{1+3+4+5+7+9+13} \approx 0.79$$

Веса (доверие) возникают даже там, где нет эксперта

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

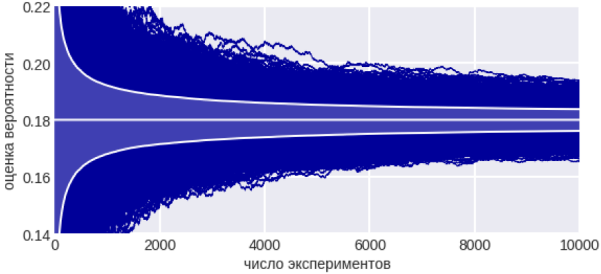
Объёмы выборок

Оцениваем вероятность в схеме Бернулли (неизвестная $p = 0.18$)



Оценка вероятности

Объёмы выборок



Выборки 10000 достаточно, чтобы оценить с точность ± 0.01?

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

Панков А.Р., Платонов Е.Н. Практикум по математической статистике: Учебное пособие. — М.: Изд-во МАИ, 2006.

Международное соревнование «dunnhumby's Shopper Challenge»

Дано: статистика визитов покупателей

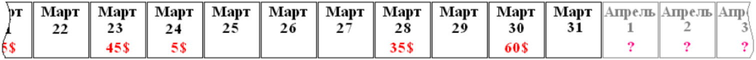
Предсказать: день первого визита и сумму покупки с точностью до 10$

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
...		

больше 100000 покупателей, время — 1 год

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

Статистика визитов одного клиента:



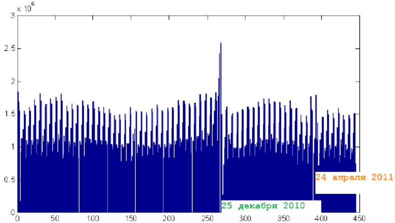
Опишем лучший алгоритм из 287

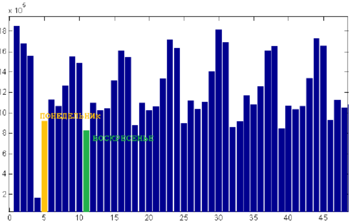
Дьяконов А.Г. Прогноз поведения клиентов супермаркетов с помощью весовых схем оценок вероятностей и плотностей // Бизнес-информатика. 2014. № 1 (27). С. 68–77

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

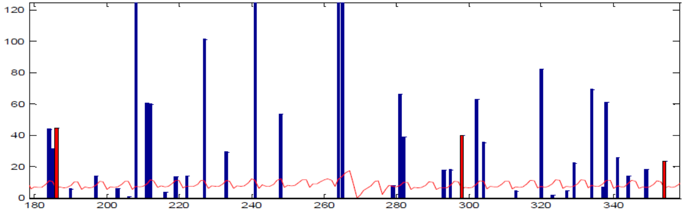
Агрегированная статистика всегда лучше

Суммы покупок всех клиентов





Покупки одного клиента



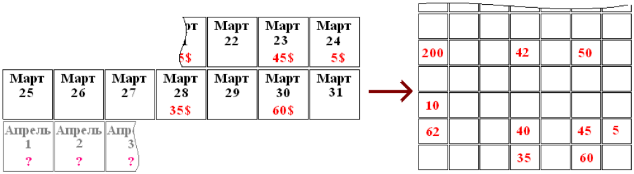
Shopper Challenge

Предположения

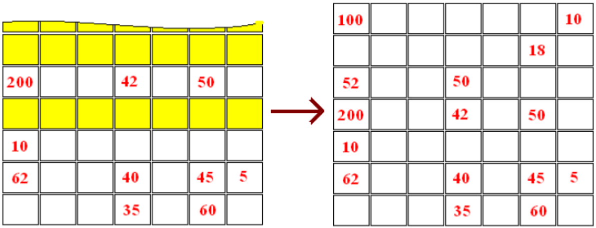
Все клиенты независимы

Будем анализировать каждого клиента отдельно

Разбиение на недели



Матрица разбивки по неделям



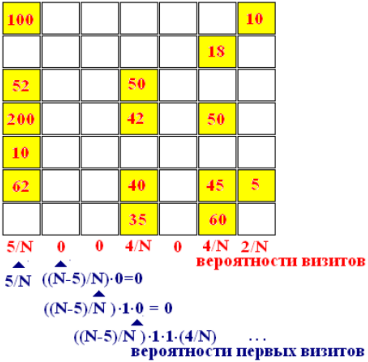
Сработало устранение пустых недель (на рисунках это будет точечный график «перестановки»

Вероятностная модель поведения клиента

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

Матрица визитов: $V = [v_{ij}]_{d\times 7}$, $v_{ij} = 0 \Leftrightarrow s_{ij} = 0$

Вероятности визитов



Shopper Challenge

Пусть $p_1, \ldots, p_7$ — оценки вероятности визита по дням недели

Оценка вероятностей **первого визита** для каждого дня недели:

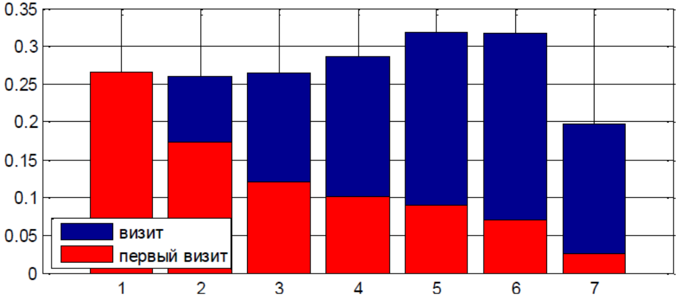
$$\begin{aligned}
\widehat{p}_1 &= p_1 \\
\widehat{p}_2 &= (1 - p_1) p_2 \\
\vdots \quad & \vdots \quad \vdots \\
\widehat{p}_7 &= \prod_{i=1}^{6} (1 - p_i) p_7
\end{aligned}$$

Находим максимум из вероятностей и получаем бейзлайн

Предположение: каждый клиент обязательно посетит магазин в течение следующей недели

Можно провести аналогичный расчёт на 14 дней, но лучше не будет

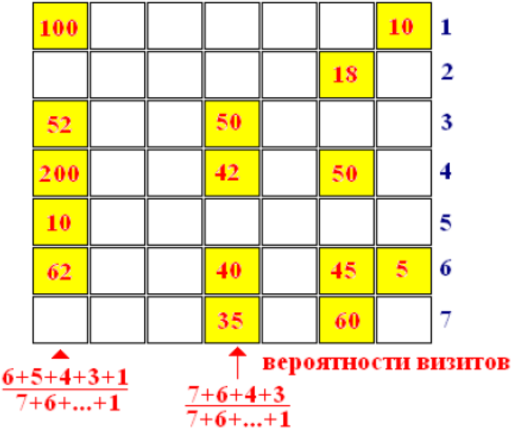
Процент визитов и первых визитов на неделе



Максимум вероятности первого визита по статистике приходится на понедельник

Shopper Challenge

«Более свежие» данные о клиенте важнее устаревших



Можно использовать весовые схемы

Взвешенная схема оценки вероятности

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

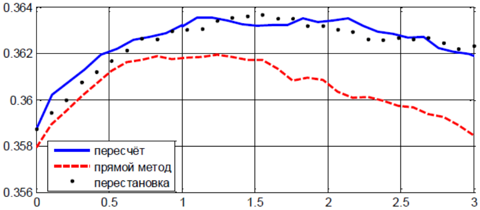
$$w_1 \geq w_2 \geq \ldots \geq w_d \geq 0, \quad \sum_{i=1}^{d} w_i = 1$$

Способы выбора весов (предыдущий слайд $\delta = 1$)

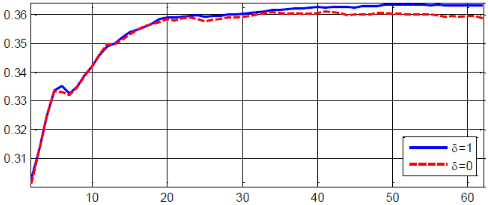
$$w_i^N = \left(\frac{d-i+1}{d}\right)^{\delta}, \ \delta \in [0,\infty), \ i \in \{1,2,\ldots,d\}$$

$$w_i = \frac{w_i^N}{\sum\limits_{i=1}^{d} w_i^N}, \ i \in \{1,2,\ldots,d\} \text{ (просто нормировка)}$$

Веса — от равномерных к «агрессивным»



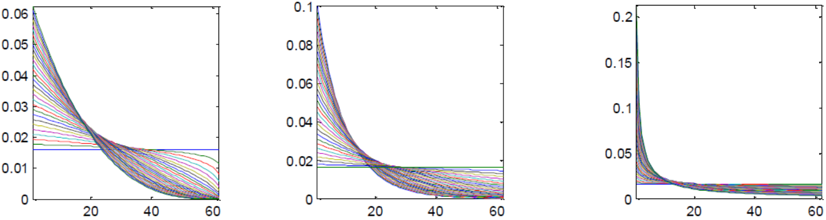
Зависимость качества прогноза от степени δ



Зависимость качества прогноза от числа учитываемых недель

Shopper Challenge

Три разные весовые схемы

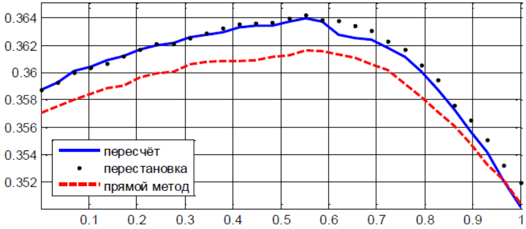


Вес недели в зависимости от её номера

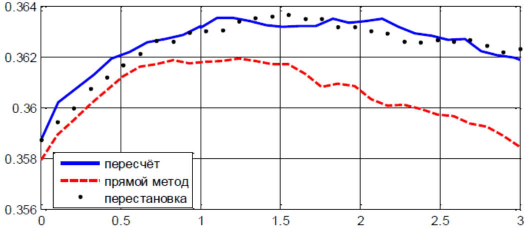
$$w_i^N = \left(\frac{d-i+1}{d}\right)^{\delta}; \qquad w_i^N = \lambda^i; \qquad w_i^N = \frac{1}{i^{\gamma}}$$

$\delta \in [0,\infty)$, $\lambda \in (0,1]$, $\gamma \in [0,\infty)$

Первая весовая схема



Третья весовая схема



Второй способ оценки вероятности первого визита

Прямой метод (можем просто посчитать сколько раз первый визит приходился на конкретный день недели)

$$\widehat{p}_j^{(2)} = \frac{1}{d} \cdot |\{i \in \{1,2,\ldots,d\} : v_{i1} = \cdots = v_{i\,j-1} = 0,\ v_{ij} = 1\}|$$

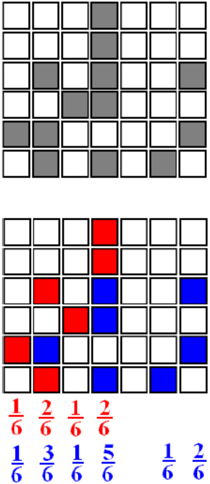
более естественный, но хуже

Матрица первых визитов

$$V' = \left[v'_{ij}\right]_{d\times 7}$$

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

На графиках это решение — красная линия и она хуже



Красным выделен первый визит

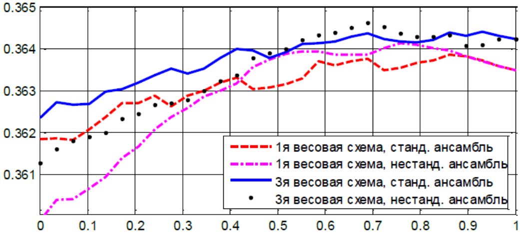
Ансамблирование (двух способов оценки вероятности)

«Стандартный ансамбль» — взять выпуклую комбинацию:

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

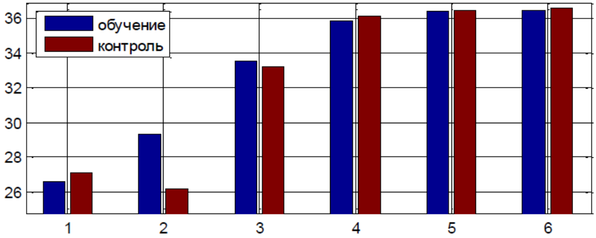
«Нестандартный ансамбль»

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



Качество ансамблирования от параметра α

Про переобучение

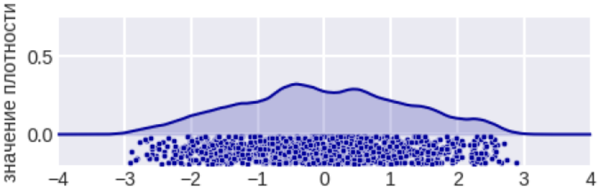


Качество на обучении и отложенном контроле для 6 алгоритмов

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

Оценка плотности распределения

Оценка плотности, какие методы знаете?



Оценка плотности распределения

- **Параметрические**
  Плотность известна с точностью до параметров
- **Непараметрические**
  Ядерные оценки
- **Восстановление смесей**
  Плотность как сумма плотностей

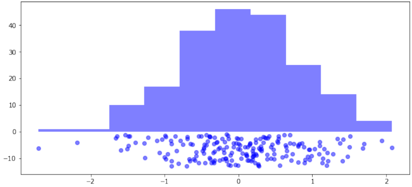
Непараметрические методы, парзеновский подход

Дана выборка $Z_m = \{X_1, \ldots, X_m\}^T$, порождённая $X \sim f(x)$

Гистограмма

Сначала строим интервалы: $[x_0 + kh, x_0 + (k+1)h)$,
m — целое, x_0 — начало, h — ширина интервала (окна)

$$\widehat{f}(x) = \frac{1}{m}\frac{\text{количество } X_i \text{ в одном интервале с } x}{\text{ширина интервала, содержащего } x} =$$
$$= \frac{1}{mh}\sum_{i=1}^{m} \mathrm{I}(X_i \text{ в одном интервале с } x)$$



Ядерная оценка плотности

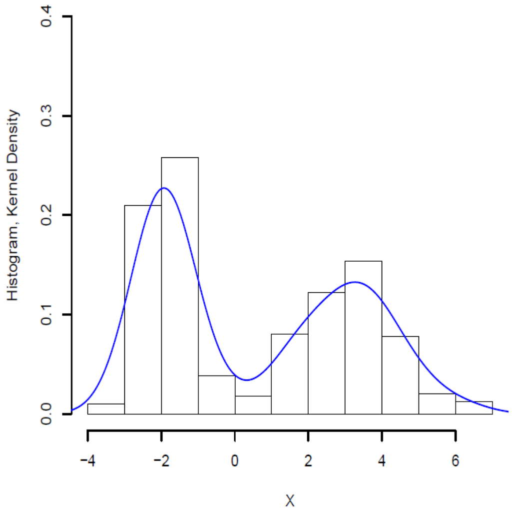
Заменим индикаторную функцию на симметричную взвешивающую функцию $K(z)$

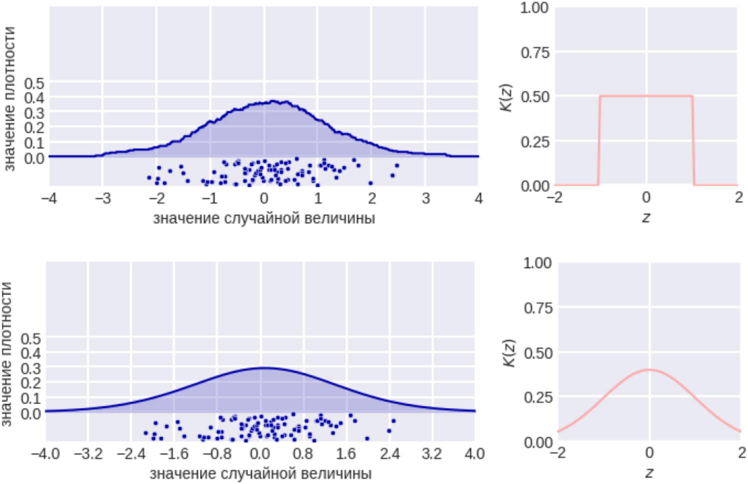
$$\widehat{f}(x) = \frac{1}{mh}\sum_{i=1}^{m} K\left(\frac{X_i - x}{h}\right)$$

Эту оценку часто называют оценкой Розенблатта-Парзена.

$K(z) \geq 0, \int\limits_{-\infty}^{+\infty} K(z)dz = 1$

Оценка плотности распределения



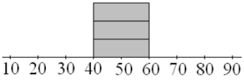


Ядро определяет степень гладкости $\widehat{f}(x)$

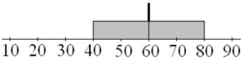
Предсказание суммы покупки (с точностью до 10$)
— непараметрическая оценка плотности

«Суммы ступенек» при покупках (с учётом ± 10)

50, 50, 50



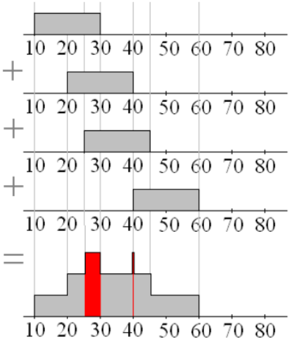
50, 70



Наилучшая стратегия предсказания суммы при условии, что пользователь ведёт себя как раньше, т.е. это оценка среднего

Прогноз с помощью моды

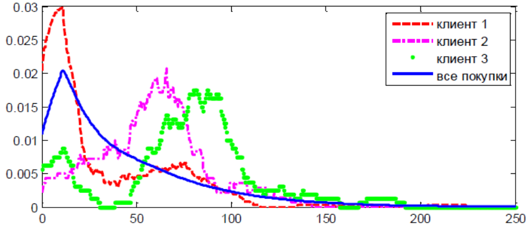
«Суммы ступенек» при покупках 20, 30, 35, 50



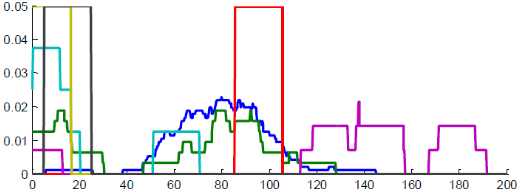
максимум достигается на отрезке $[25, 30]$ и в точке 40

Как выглядят плотности

Плотности распределения покупок



Плотности покупок одного клиента в разные дни недели



Предсказание суммы покупки

Пример весовой схемы

$$\widehat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} K\left(|s_i - x|\right)$$

s_i — сумма покупки

$$K\left(|s - x|\right) = \begin{cases} \dfrac{1}{2\varepsilon}, & |s - x| \leq \varepsilon \\ 0, & |s - x| > \varepsilon \end{cases}$$

$$\widehat{f}(x) = \sum_{i=1}^{m} w_i K\left(|s_i - x|\right)$$

Предсказание суммы покупки

Учёт времени, дня недели

Пусть $s_1, \dots, s_m$ — упорядоченные по времени покупки клиента $\tilde{s}_1, \dots, \tilde{s}_{\tilde{m}}$ — упорядоченные покупки, сделанные в этот день недели (мы его предсказали)

Оценим плотность для расширенного набора $\tilde{s}_1, \dots, \tilde{s}_{\tilde{m}}, s_1, \dots, s_m$
Веса:

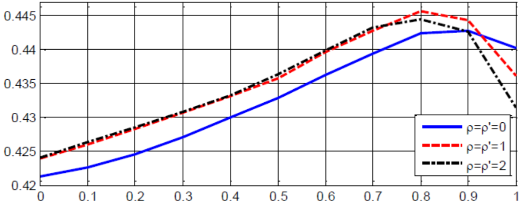
$$\tilde{s}_i \leftrightarrow \beta \frac{(\tilde{m} - i + 1)^{\tilde{\rho}}}{\sum\limits_{j=1}^{\tilde{m}} j^{\tilde{\rho}}} \qquad s_i \leftrightarrow (1 - \beta) \frac{(m - i + 1)^{\rho}}{\sum\limits_{j=1}^{m} j^{\rho}}$$

β регулирует насколько важна статистика за день недели

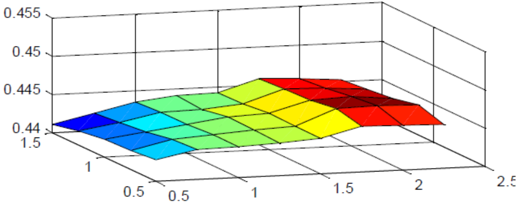
Параметры настраиваем по обучающей выборке

Весовая схема

Качество прогноза суммы покупок от параметра β



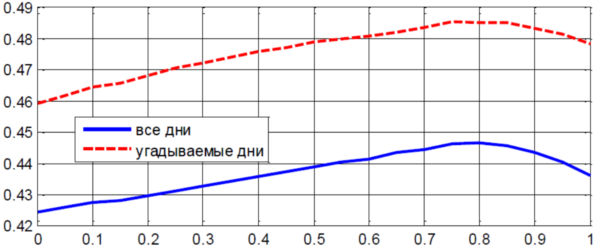
Качество прогноза в зависимости от степеней при $\beta = 0.8$



Предсказание суммы покупки

Как настраивать параметры, точнее где

- на всей выборке
- только на угадываемых днях



Качество прогноза суммы покупок от параметра β при $\rho = 0.7$, $\tilde{\rho} = 1.6$. Видно, что точки максимума практически одинаковые

Предсказание суммы покупки

Улучшение алгоритма

- метод предсказания даты визита
- метод предсказания суммы покупки (точка максимума оценки плотности)

Как объединить два независимых прогноза?

Простой путь: сначала определяем день и затем для этого дня строим прогноз суммы Но эти события могут быть зависимыми, как тогда?

Предсказание суммы покупки

Рассмотри клиента, который ходит часто в понедельник и совершает покупки на разные суммы, а во вторник ходит реже,но сумма одна и та же

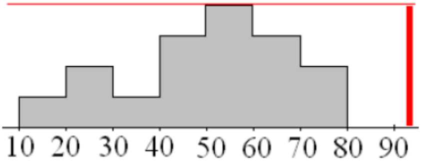
Понедельник: 10\$, 50\$, 220\$, 100\$, 310\$, 5\$, 250\$, 75\$, 500\$

Вторник: 40\$, 42\$, 40\$

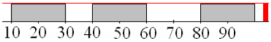
(вероятность угадать день) $\times$ (вероятность угадать сумму):
$0.9 \cdot 0.1 = 0.09$
$0.1 \cdot 1 = 0.1$ выгоднее ставить на вторник

Новая стратегия: вычислить вероятности угадывания дня и суммы, т.е. при выборе наиболее вероятного дня будем учитывать насколько хорошо в этот день предсказывается сумма покупки

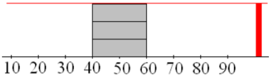
Как вычислить стабильность поведения клиента?



максимум плотности q_j можно принять за аналог вероятности угадать сумму (мера стабильности)



низкая стабильность



высокая стабильность

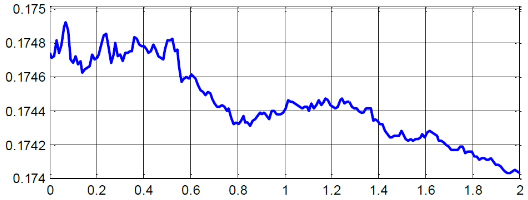
учёт стабильности $\Rightarrow$ улучшение результата

Неполный учёт стабильности (при оценке вероятности)

$$\widehat{p}_j(q_j + h) \to \max_j$$

это и регуляризация и ансамблирование:

$$\widehat{p}_j q_j \to \max\,, \quad h\widehat{p}_j \to \max$$



Качество предсказания в зависимости от параметра h, который меняет вес прогноза для дня и для суммы покупки

Предсказание суммы покупки

Итог

- Каждый метод — система предположений
- Можно решать задачи простыми методами
- Весовые схемы улучшают качество
- Есть методы, в которые хорошо интегрируются весовые схемы
- Умная состыковка методов

Предсказание суммы покупки

Домашнее задание

1. Дана статистика посещений ресурса за 1099 дней для 300 000 пользователей (Train)
2. Требуется спрогнозировать для каждого пользователя день недели его следующего визита, или его отсутствие
3. Функционал качества: процент правильных ответов, например,

$$\mathrm{performance}([1,2,2,7],[3,2,2,7]) = 0.75$$

Формат ответа (solutionex): в файле по строкам id пользователей и номера дней их первых визитов по версии вашего алгоритма:

id,nextvisit

1, 7

Ссылка на данные