«Прикладные задачи анализа данных»

CASE: Прогнозирование визитов покупателей супермаркетов и сумм их покупок

Александр Дьяконов

(ВМК МГУ имени М.В. Ломоносова)

7-8 ноября 2019 года

План лекции

Постановка задачи

Предположения метода

Оценки вероятности / весовые схемы

Оценки плотности / весовые схемы

«Состыковка» алгоритмов

в этой лекции будут плохие картинки

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:

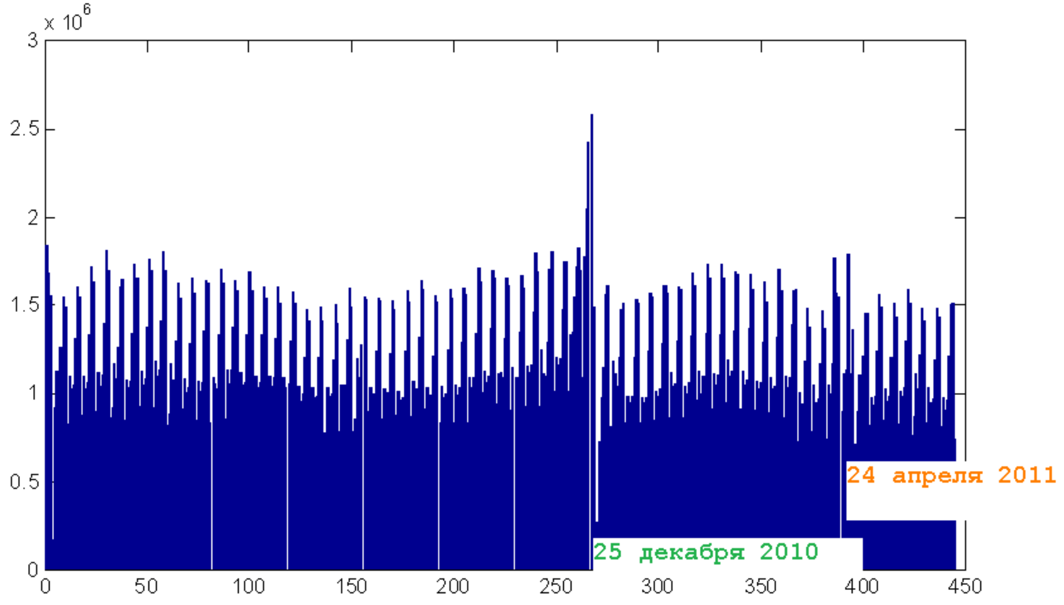
рт 1	Март 22	Март 23	Март 24	Март 25	Март 26	Март 27	Март 28	Март 29	Март 30	Март 31	Апрель 1	Апрель 2	Апр 3
5$		45$	5$				35$		60$		?	?	?

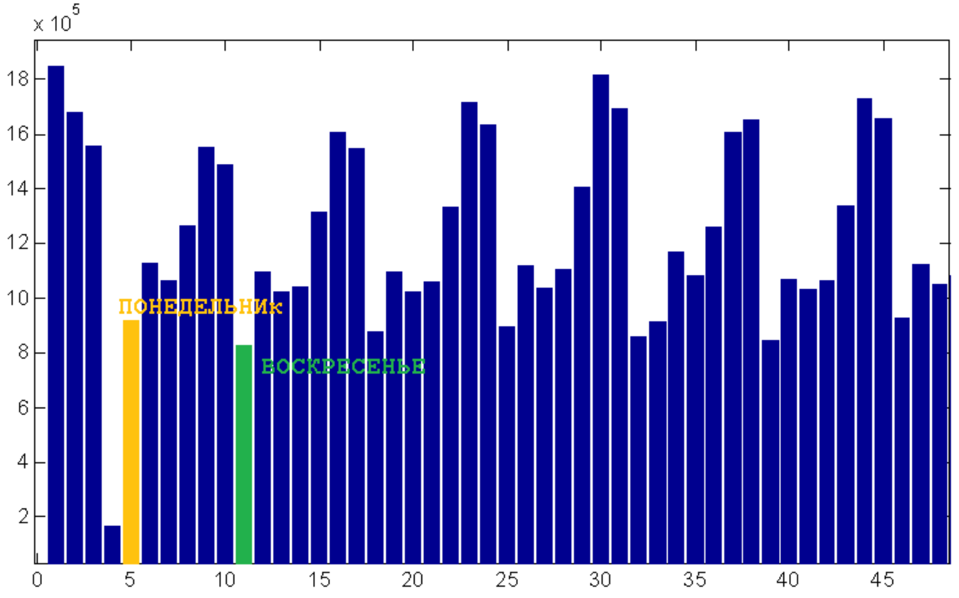
Опишем лучший алгоритм из 287

#	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

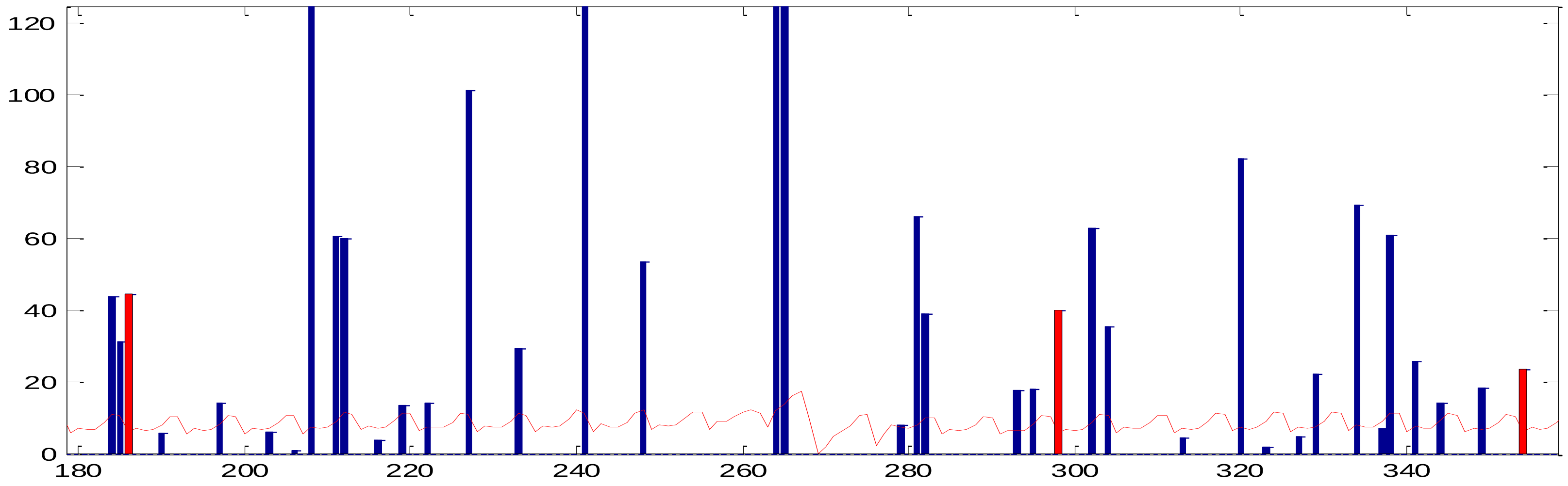
Агрегированная статистика всегда лучше

Суммы покупок всех клиентов





Покупки одного клиента

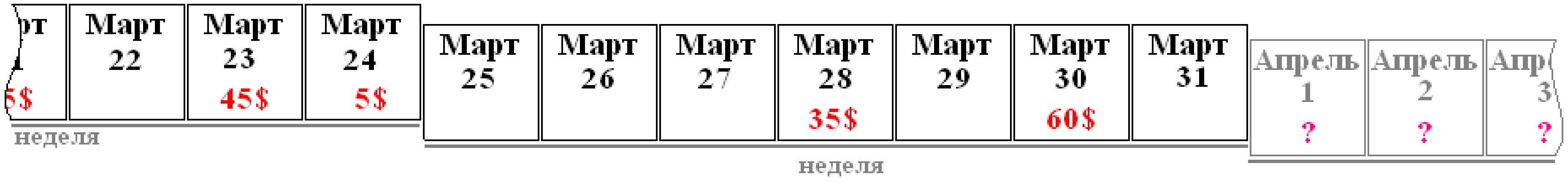


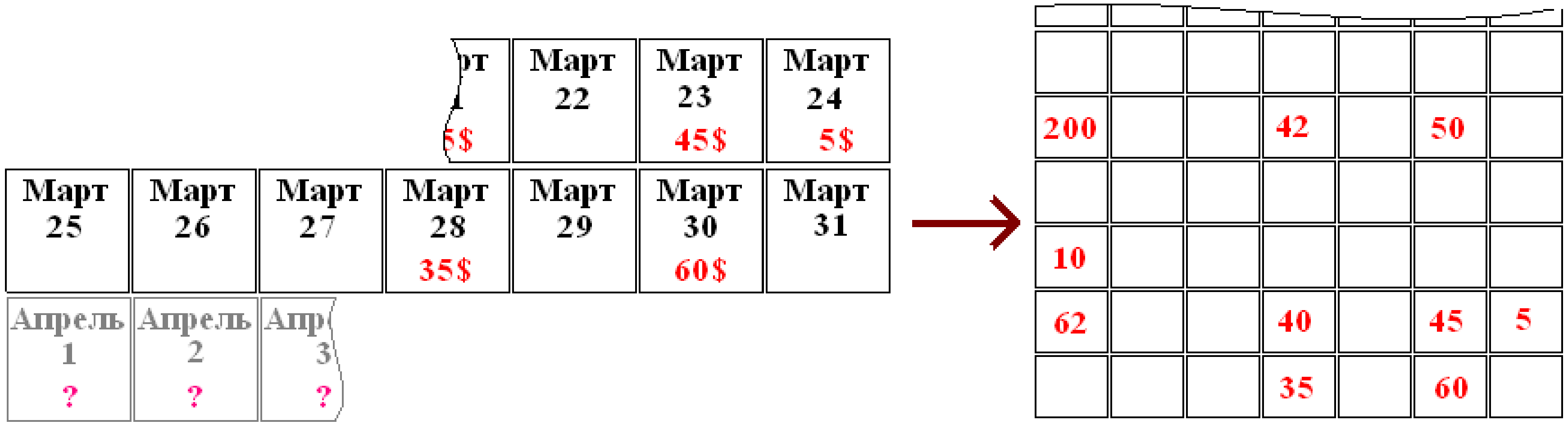
Предположения

Все клиенты независимы

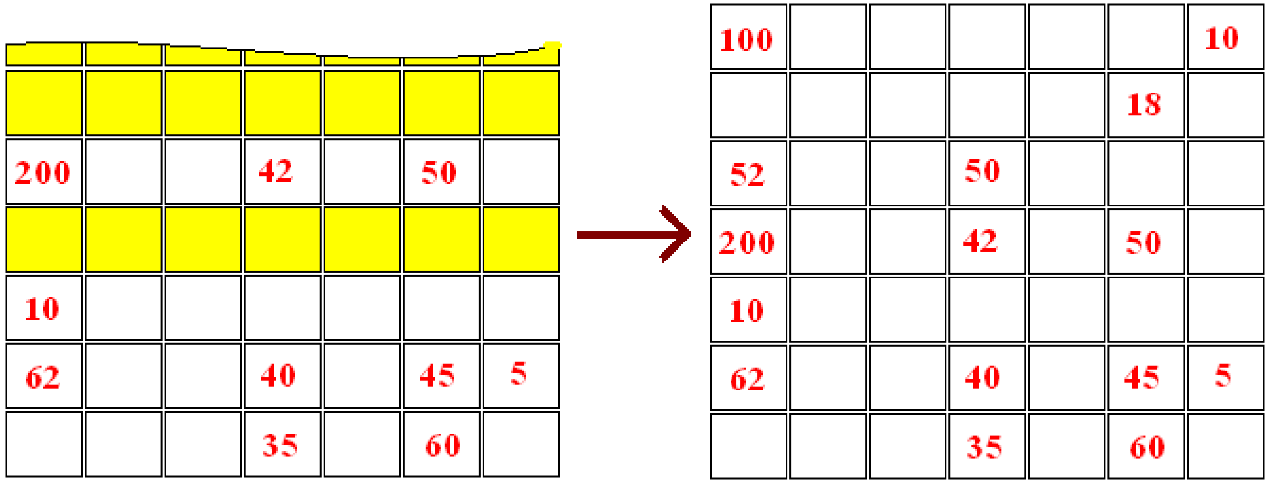
Будем анализировать каждого клиента отдельно

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

Вероятности визитов

оценки вероятностей… **первых визитов**

100						10
					18	
52			50			
200			42		50	
10						
62			40		45	5
			35		60	
5/N	0	0	4/N	0	4/N	2/N

вероятности визитов

5/N ((N-5)/N)·0=0

((N-5)/N)·1·0 = 0

((N-5)/N)·1·1·(4/N) …

вероятности первых визитов

$$p_1 \qquad \tilde{p}_1 = p_1$$

$$p_2 \qquad \tilde{p}_2 = (1 - p_1)p_2$$

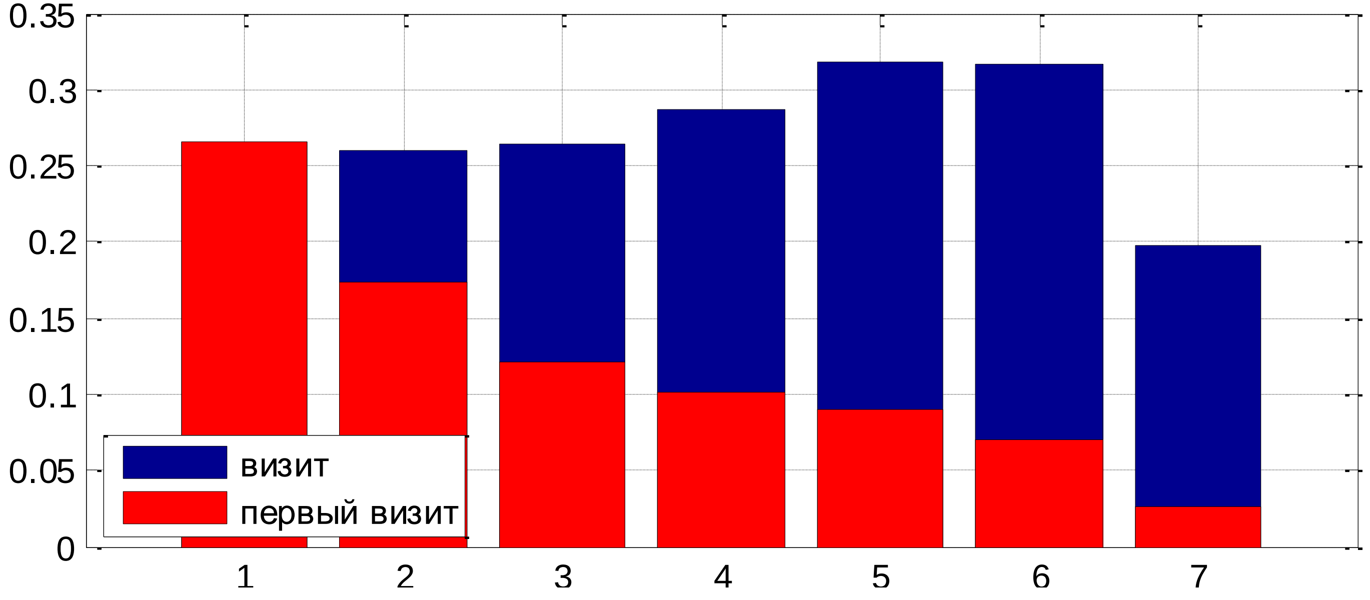
$$\ldots \qquad \ldots$$

$$p_7 \qquad \tilde{p}_7 = \prod_{i=1}^{6}(1 - p_i)p_7$$

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

100						10	1
					18		2
52			50				3
200			42		50		4
10							5
62			40		45	5	6
			35		60		7

$$\frac{6+5+4+3+1}{7+6+\ldots+1} \qquad \frac{7+6+4+3}{7+6+\ldots+1}$$

вероятности визитов

Весовые схемы!

Взвешенная схема оценки вероятности

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

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

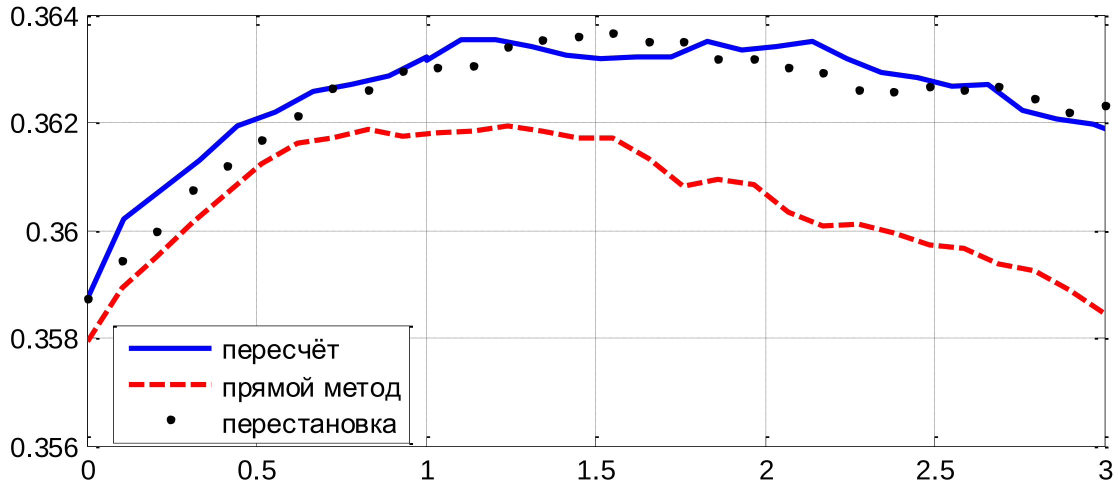
Способы

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

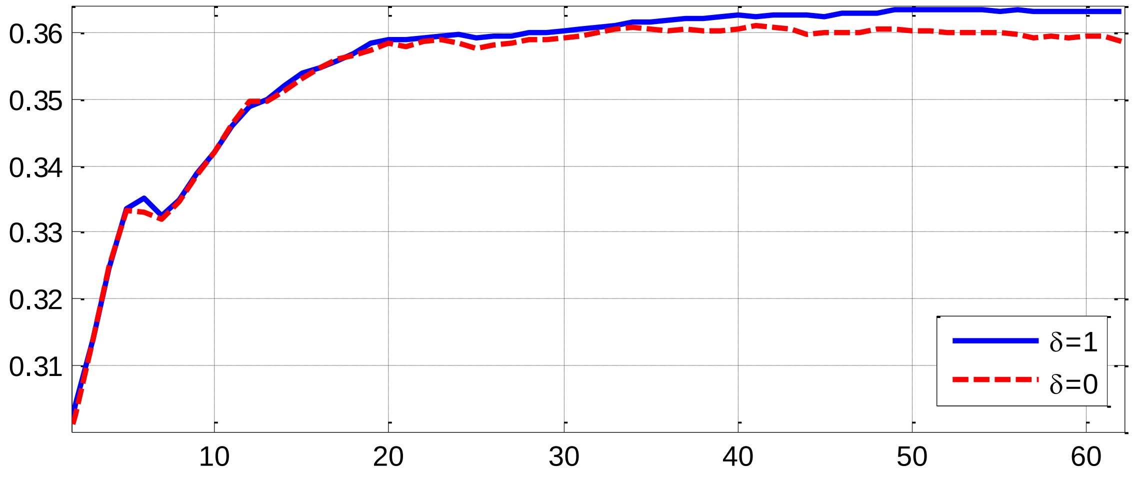
$$w_i = \frac{w_i^{\mathrm{N}}}{\sum_{i=1}^{d} w_i^{\mathrm{N}}}, i \in \{1,2,\ldots,d\}.$$ [просто нормировка]

Параметр $\delta \in [0,+\infty)$.

Веса – от равномерных к «агрессивным»

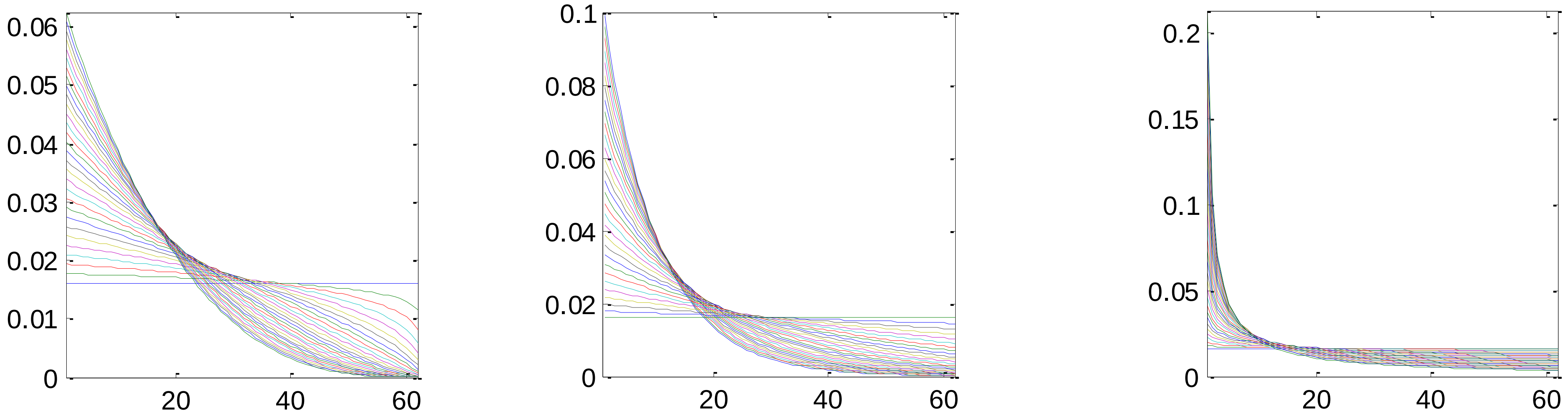


Зависимость качества прогноза от степени δ



Зависимость качества прогноза от числа учитываемых недель

Три разные весовые схемы



вес недели в зависимости от её номера

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

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

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

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

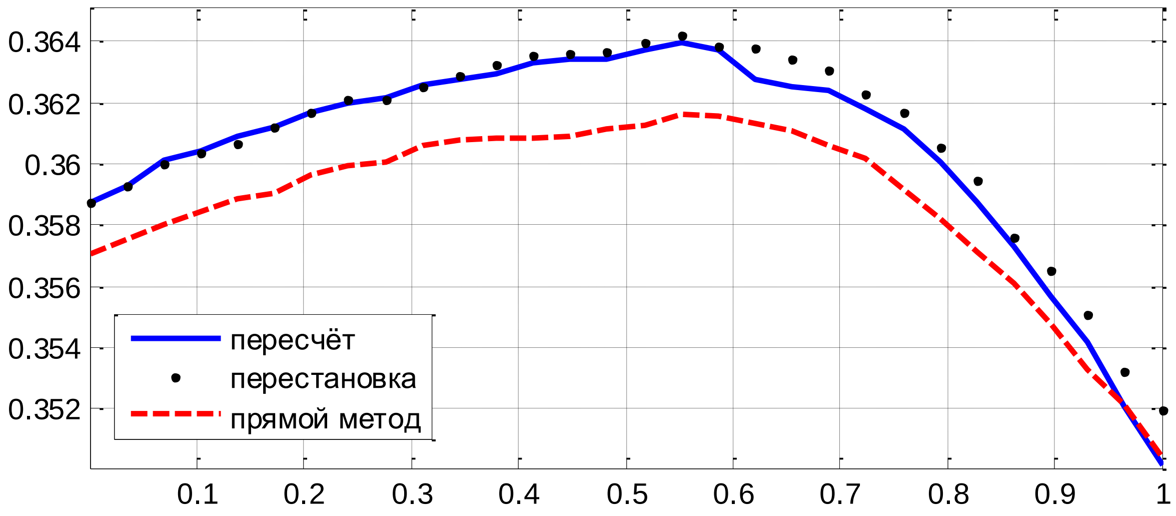
$$w_i^{\mathrm{N}} = \frac{1}{i^{\gamma}},$$

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

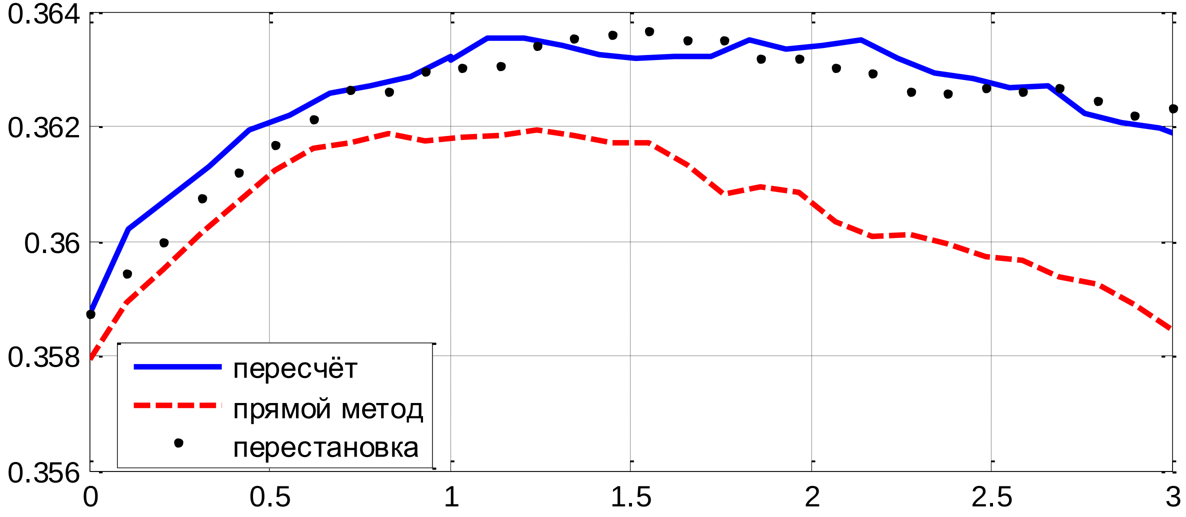
ДЗ исследовать эти схемы

ДЗ предложить другие – исследовать

Принципиально всё одинаково…



Третья весовая схема



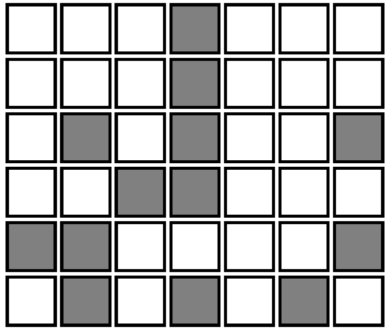
Первая весовая схема

Два способа оценки вероятности первого визита

Прямой метод

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

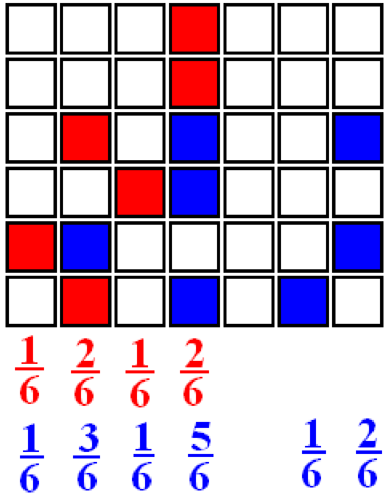
Более естественный, но хуже!



матрица первых визитов

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

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



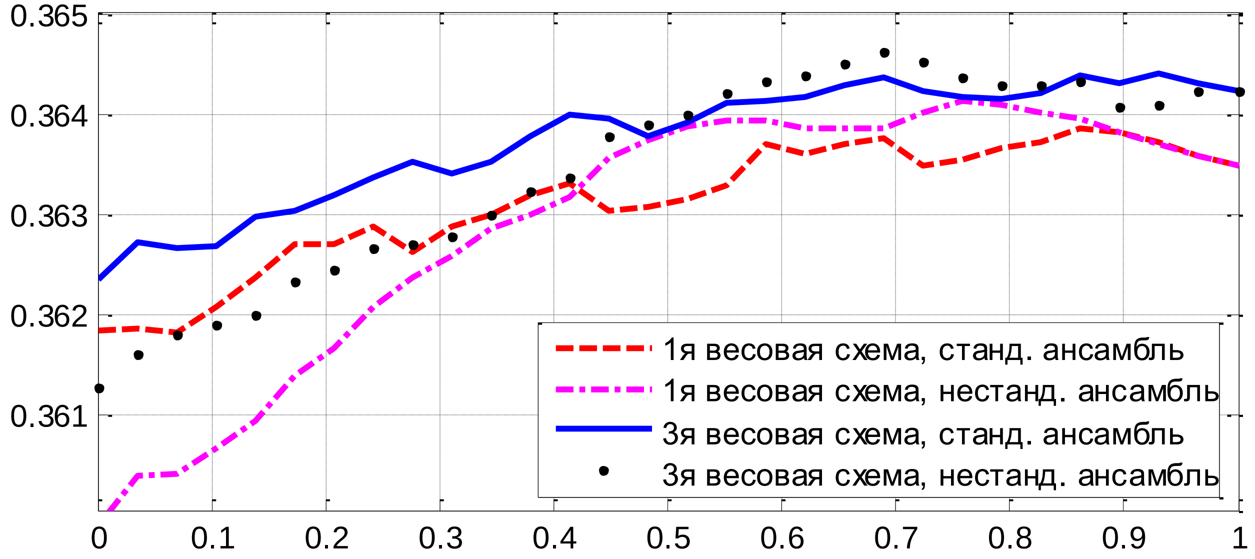
Ансамблирование

«Стандартный ансамбль» – взять выпуклую комбинацию:

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

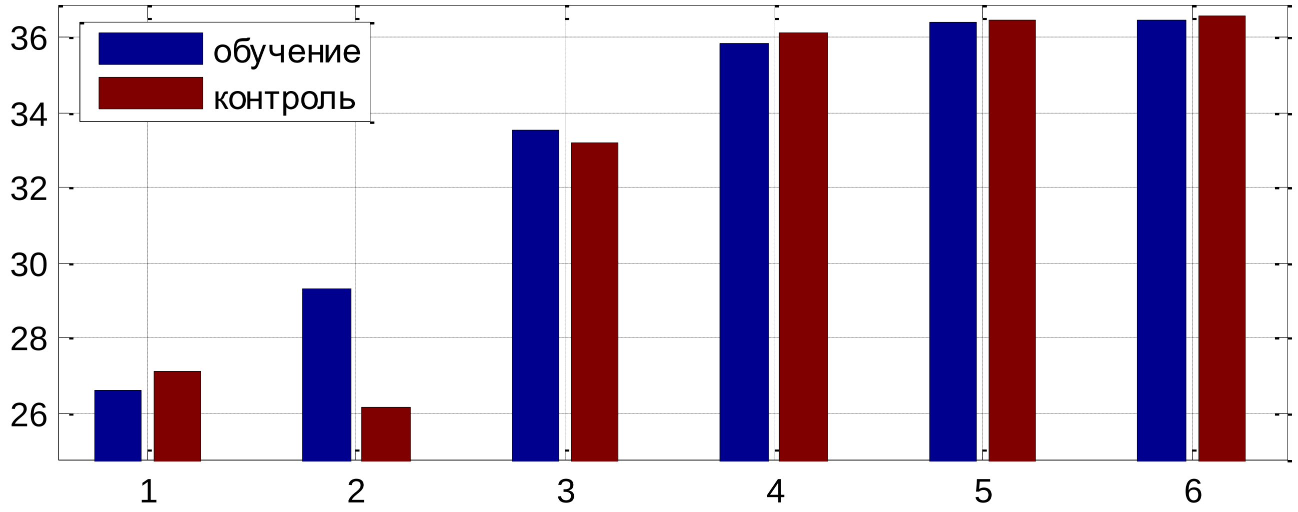
«Нестандартный ансамбль»

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



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

Про переобучение

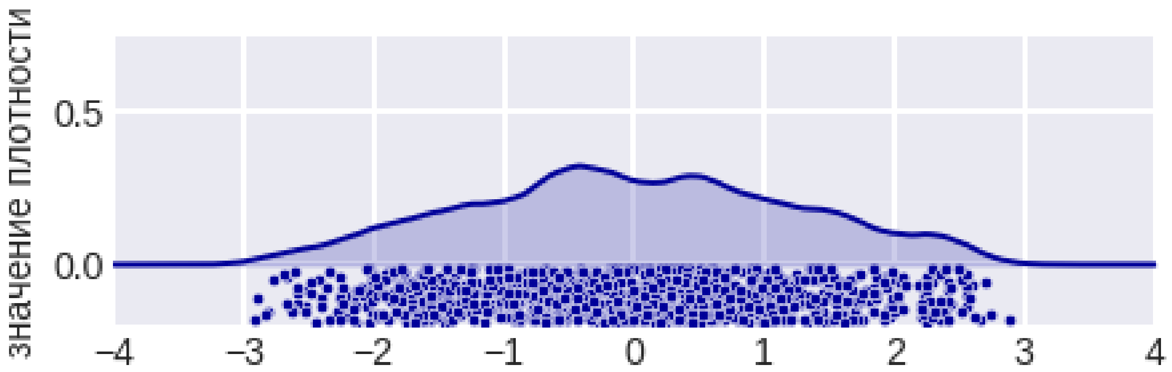


Качество на обучении и отложенном контроле для шести алгоритмов

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!

Восстановление плотности



Какие методы знаете?

Восстановление плотности

1. Параметрические

Плотность известна с точностью до параметров

2. Непараметрические

Вид плотности не известен

3. Восстановление смесей

Плотность = сумме плотностей

Непараметрические методы восстановления плотности

Парзеновский подход

Выборка $x^1,\ldots,x^m$ **в пространстве** $\mathbf{R}^d$

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

где $K(x)$ **– функция окна**

$$K(z)\geq 0$$

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

ДЗ Исследование применимости непараметрических методов (подбор параметров)

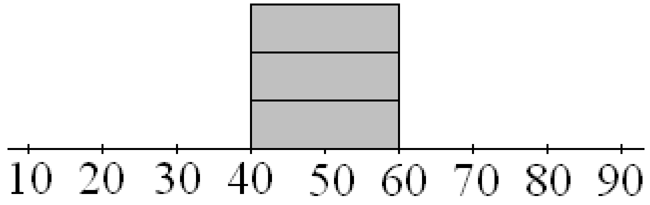
Непараметрические методы восстановления плотности

Предсказание суммы покупки

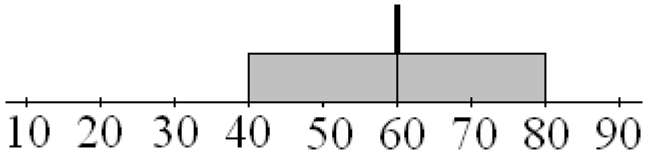
= непараметрическое восстановление плотности по Парзену

«Суммы ступенек» при покупках

50, 50, 50



50, 70

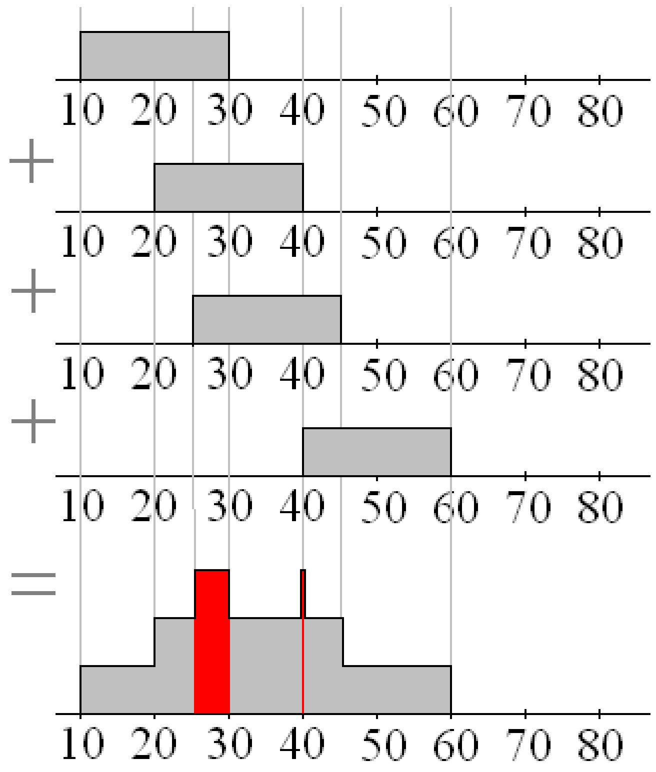


Наилучшая стратегия предсказания суммы при условии, что пользователь ведёт себя как раньше

т.е. это оценка среднего

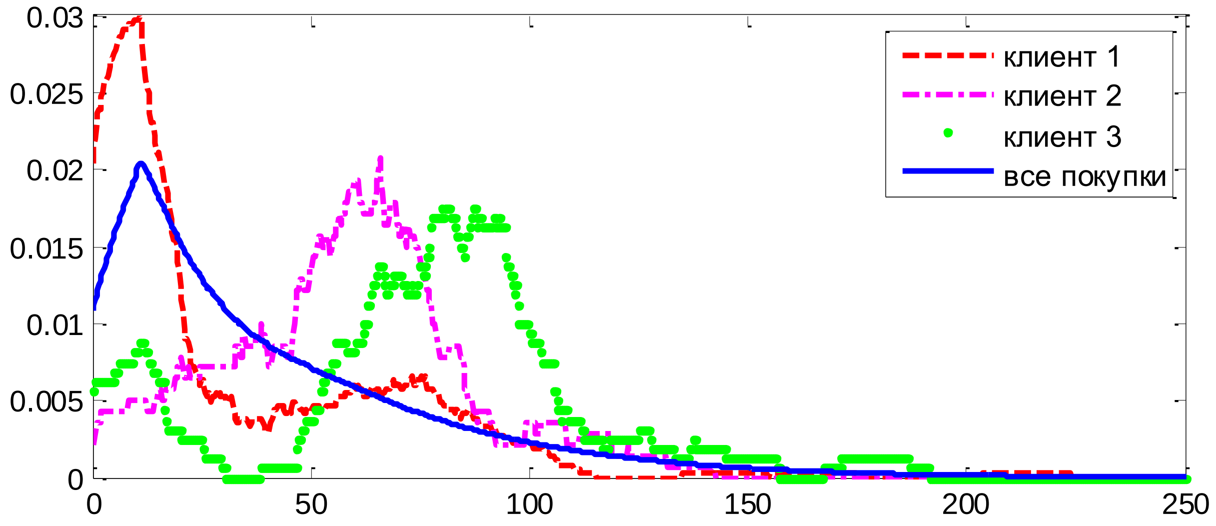
Прогноз с помощью моды

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

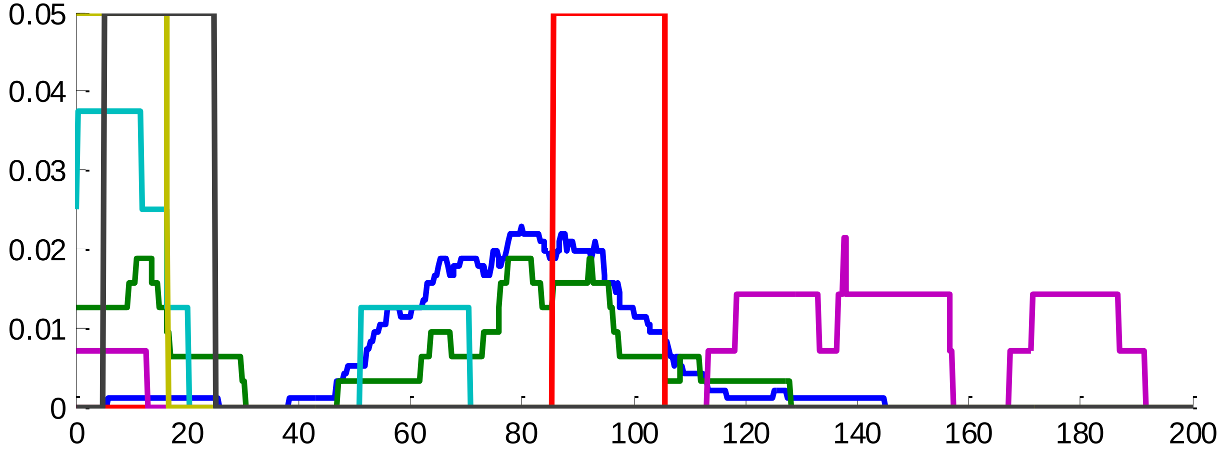


максимум достигается на отрезке [25, 30] и в точке 40.

Как выглядят плотности



Плотности распределения покупок



Плотности покупок одного пользователя в разные дни недели

И здесь сделаем весовую схему!

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

$$2\int_{0}^{+\infty}K(x)\,dx=1.$$

$$K(|s-x|)=\begin{cases}1/2\varepsilon, & |s-x|\le\varepsilon,\\ 0, & |s-x|>\varepsilon.\end{cases}$$

Весовая схема:

$$f(x)=\sum_{i=1}^{m}w_iK(|s_i-x|)$$

Весовая схема

учёт времени, дня недели

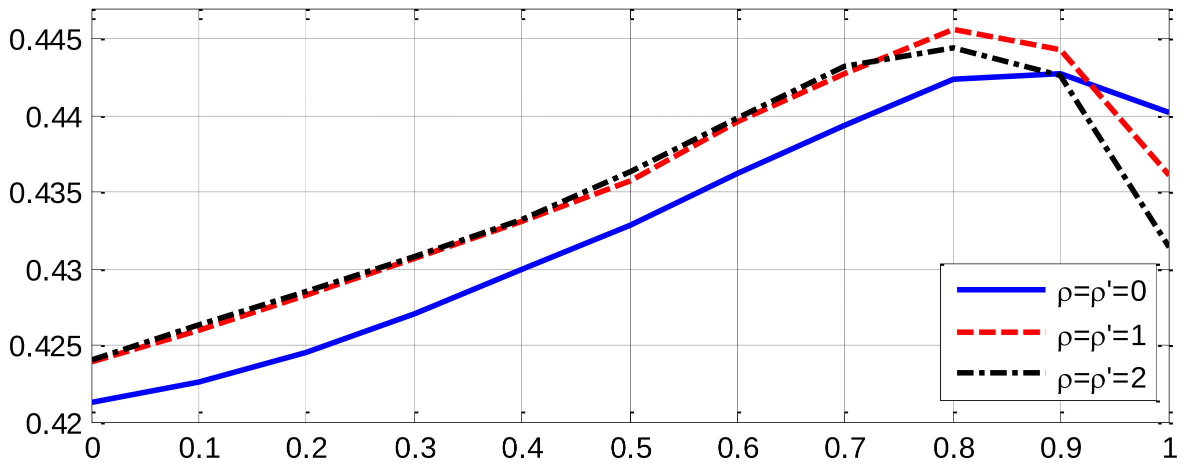
Пусть $s_1, \ldots, s_m$ **– все упорядоченные покупки пользователя,**
$s'_1, \ldots, s'_{m'}$ **– покупки, сделанные в этот день недели.**

Плотность будем восстанавливать для расширенного набора $s'_1, \ldots, s'_{m'}, s_1, \ldots, s_m$**.**

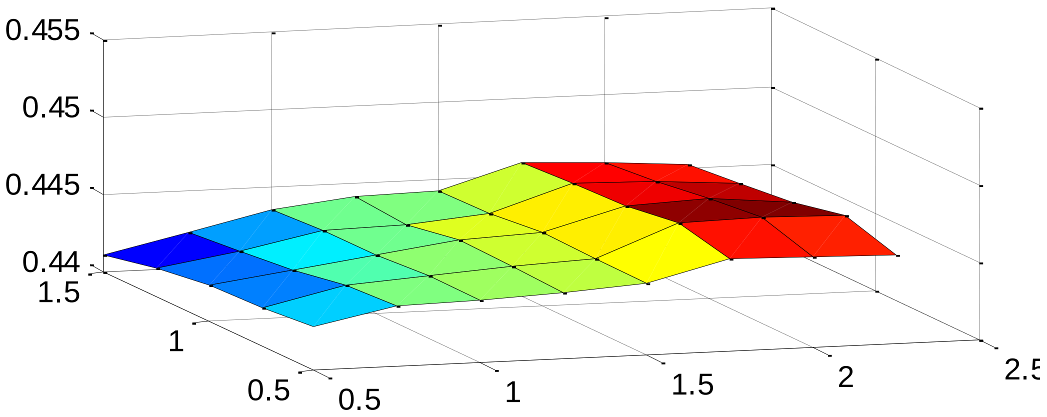
Веса:

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

Весовая схема



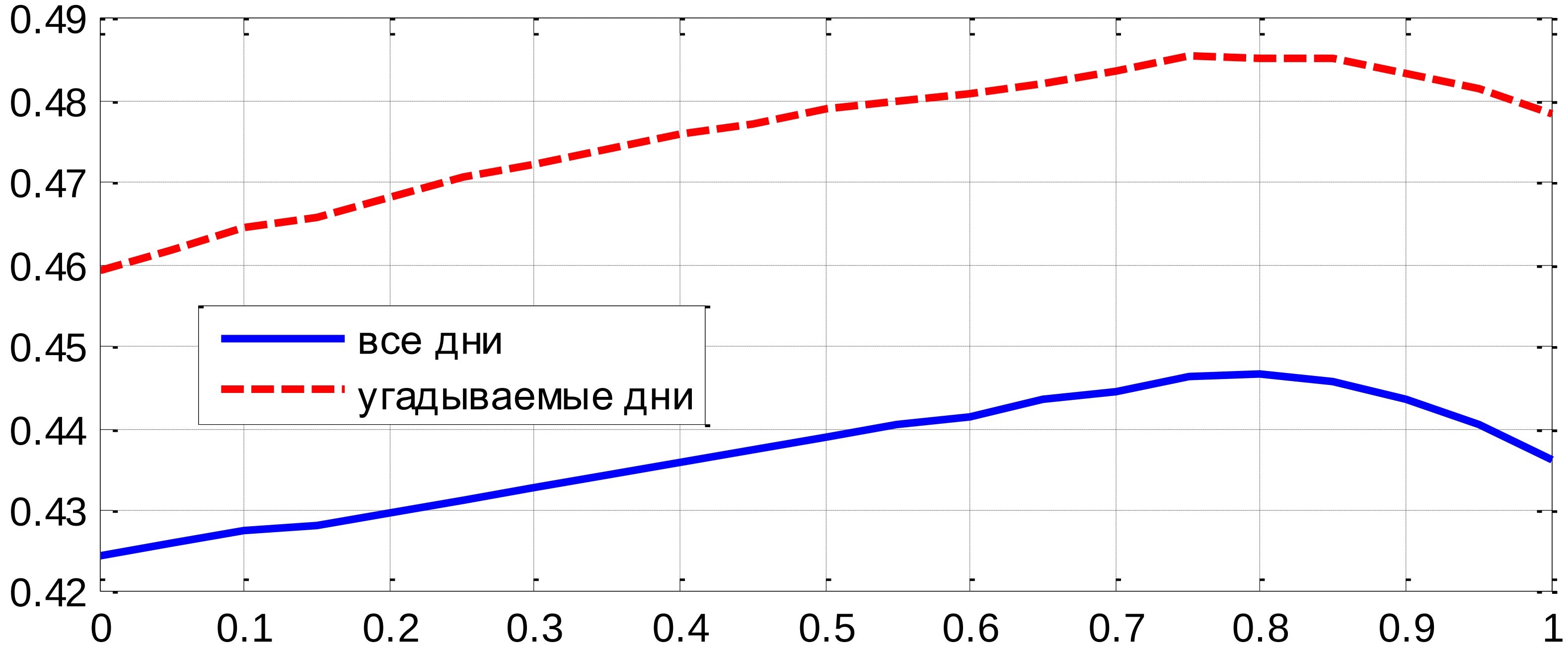
Качество прогноза суммы покупок от параметра β



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

Как настраивать, точнее где…

- **на всей выборке**
- **на угадываемых днях**
  (на остальных – бесполезно для функционала)



Качество прогноза суммы покупок от параметра β при $\rho = 0.7$, $\rho' = 1.6$.

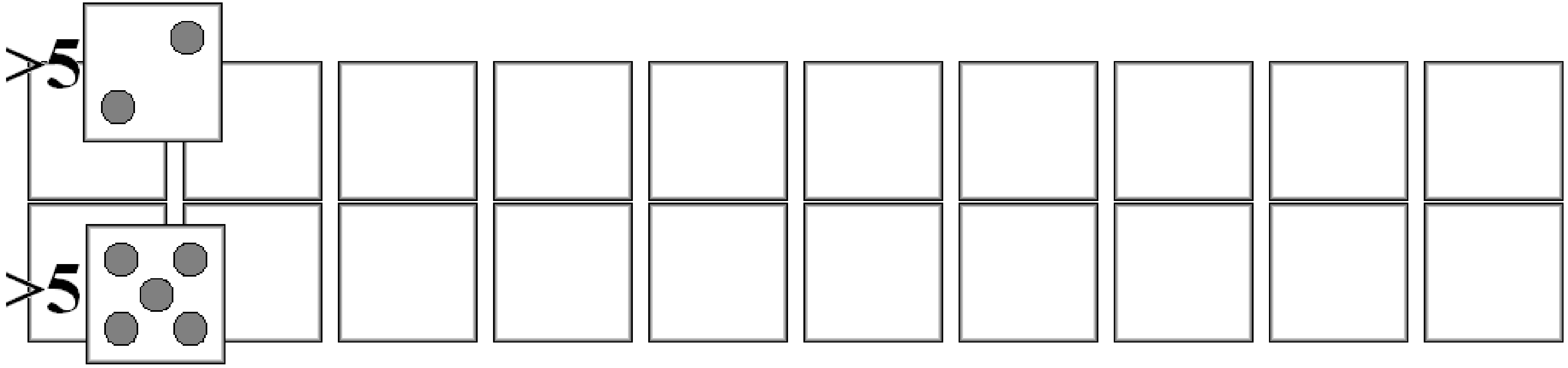
Улучшение алгоритма

Есть:

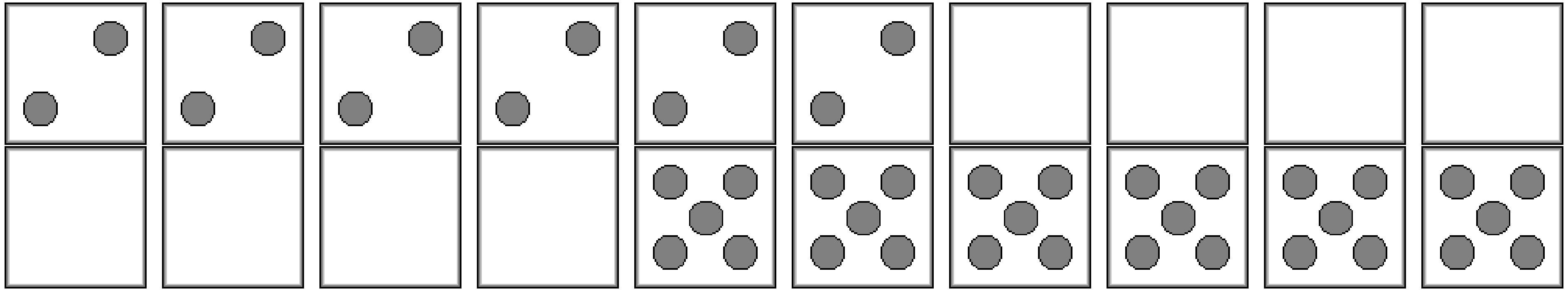
- **метод предсказания даты визита** (вероятностный пересчёт)
- **метод предсказания суммы покупки** (непараметрическое восстановление)

Можно ли так осуществить прогноз?

Все прогнозировали так…



Почему метод работает не очень хорошо…



«И» в условии не означает «И» в решении

Найти день И сумму.

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

Вторник: 40$, 42$, 40$

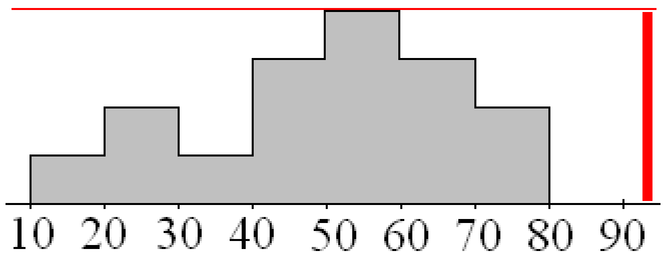
(вероятность угадать день) $*$ (вероятность угадать сумму)

$0.9 * 0.1 = 0.09$

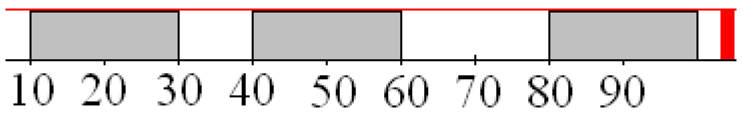
$0.1 * 1 = 0.1$ выгоднее ставить на вторник

Надо: вычислить вероятность угадывания дня и суммы

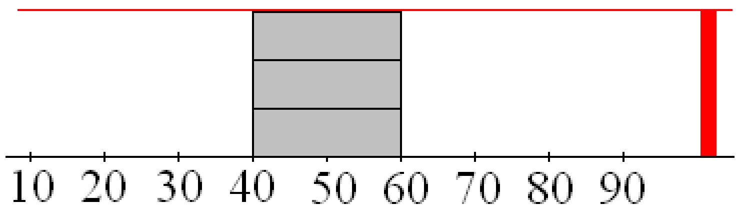
Как вычислить стабильность поведения клиента?



высота графика плотности



низкая стабильность



высокая стабильность

учёт стабильности = улучшение результата

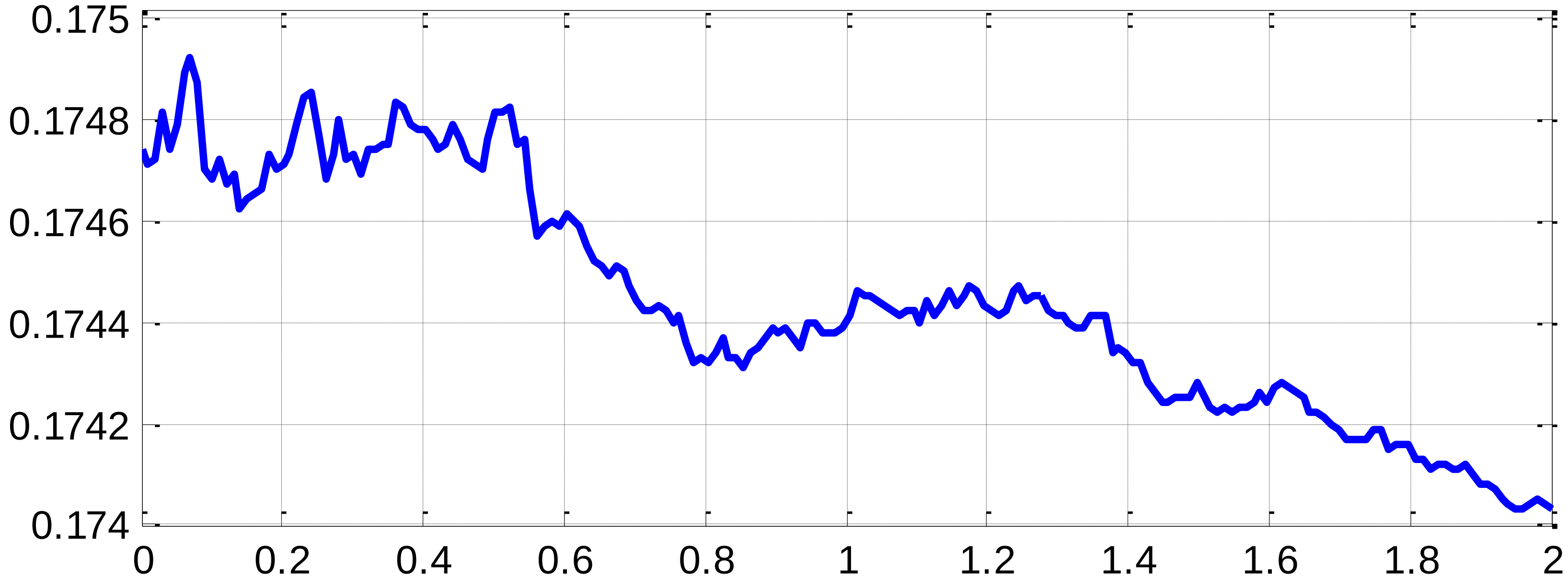
Неполный учёт стабильности

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

это и регуляризация

и ансамблирование $(\tilde{p}_j q_j + h\tilde{p}_j)$

max max



Качество предсказания поведения в зависимости от параметра h.

Итог

Каждый метод – система предположений

Можно решать задачи простыми методами
ещё об этом поговорим

Весовые схемы улучшают качество

Есть методы, в которые хорошо интегрируются весовые схемы
оценки среднего, вероятности
парзеновские методы

Умная состыковка методов

ДЗ повторить (часть) исследований: подтвердить / опровергнуть графики

Литература

- **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**