EleWise проекты

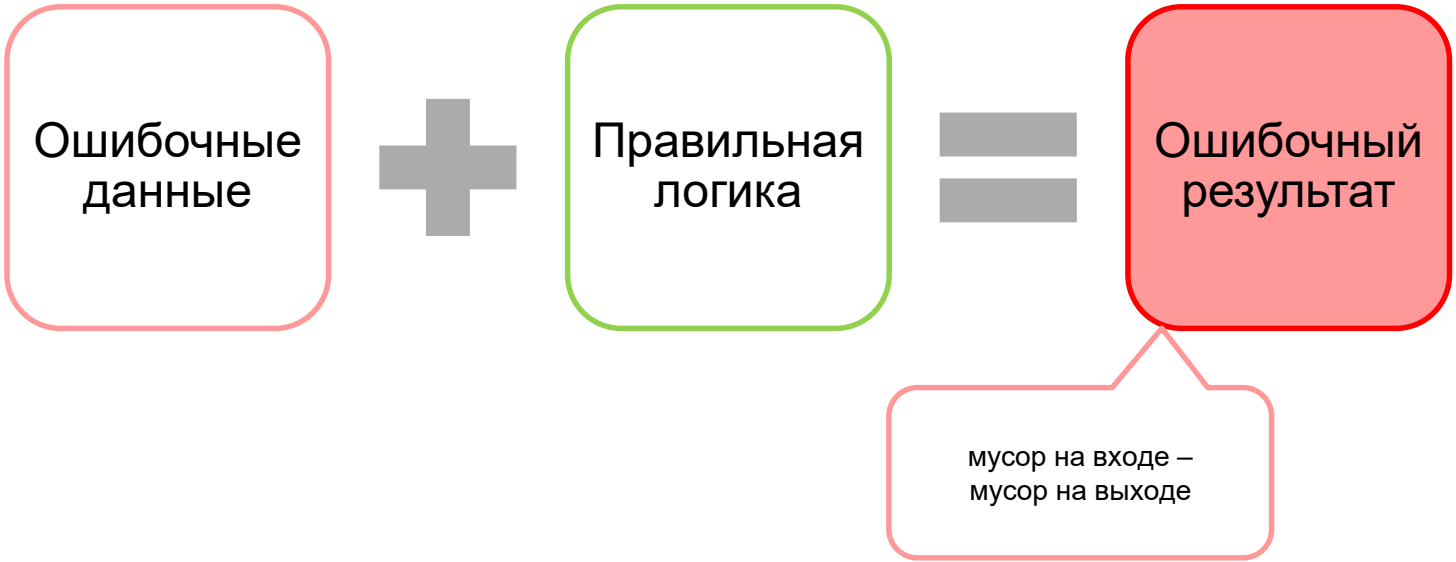
Лекция 2

Работа с данными

2.1. Изучение данных

2.1. Изучение данных

Качество данных: проблема



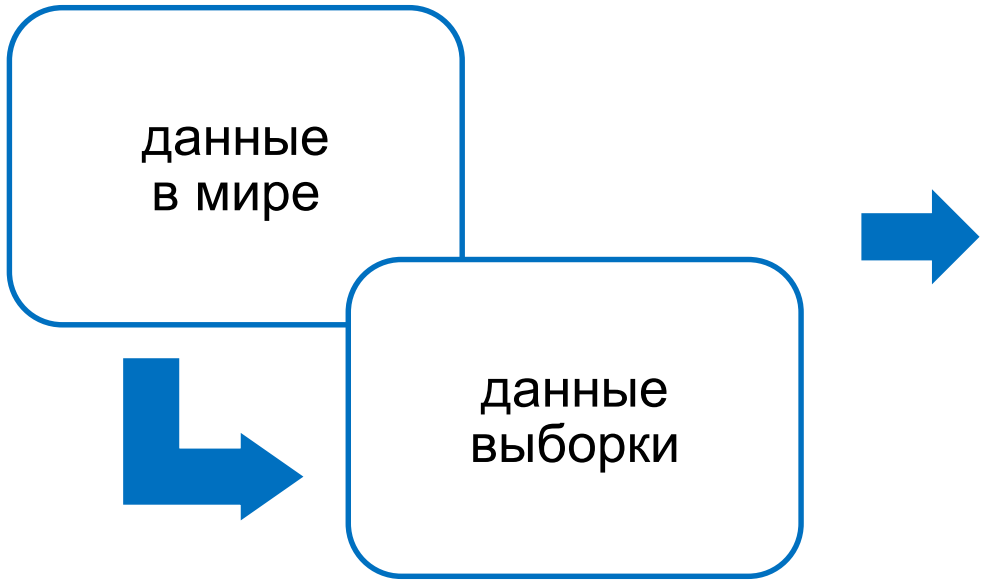
2.1. Изучение данных

- распределение
- среднее, медиана, мода
- доля пропущенных значений
- диапазон значений характеристики

+

визуализация

2.1. Изучение данных



репрезентативность выборки

2.1. Изучение данных
2.1.1. Пропущенные значения и выбросы

значения, которые выпадают из нормального диапазона значений той или иной характеристики:

- поля не используются;
- значения полей не зафиксировано;
- поля недоступны;
- поля не заполнены;
- неправильно введены значения;
- выбросы;
- резко выделяющиеся значения

Пример:
возраст – популяция 18-55 лет
единичные случаи: 115, 125 лет

2.1. Изучение данных
2.1.1. Пропущенные значения и выбросы

1. исключить все данные с пропущенными значениями [строки и столбцы]
2. исключить все характеристики, для которых доля пропущенных значений значительна (~50%) [столбцы]
3. исключить все записи, для которых доля пропущенных значений значительна (~50%) [строки]

PS: это не все, см. далее

2.1. Изучение данных
2.1.1. Пропущенные значения и выбросы

данные
надо
изучить!

2.1. Изучение данных
2.1.2. Предобработка данных

Одномерные преобразования

- **Масштабирование признаков**
  - Перевод в [0, 1]
  - Стандартизация
  - Робастная стандартизация

$$\tilde{x} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

$$\tilde{x} = \frac{x - x_i}{\sigma_x}$$

$$\tilde{x} = \frac{x - q_\alpha(x)}{q_{1-\alpha}(x) - q_\alpha(x)}$$

- **Обработка выбросов**
  - Клипование

$$\tilde{x} = \min(100500, \max(x, 0))$$

- **Нормализация**
  - Логарифмирование
  - Квантильное отображение

$$\tilde{x} = \log(1 + x)$$

$$\tilde{x} = rank(x)$$

2.1. Изучение данных
2.1.2. Предобработка данных

Заполнение пропусков

- **Откладывание**
  - Заполнение другим значением (9999)
  - Заполнение нулем, сохранение бинарной метки
- **Одномерное заполнение**
  - Среднее, мода, медиана
  - Предыдущее значение ряда
- **Моделирование**
  - По ближайшим соседям
  - Предсказать регрессией
  - Взять из условного совместного распределения

НО: возникнет всплеск данных

2.1. Изучение данных
2.1.2. Предобработка данных

EleWise проекты

Feature engineering

- Многомерные преобразования
  - Отношения, min/max
- Признаки признаков
  - Частоты
  - Классификации (город, большой город, очень большой город)
- Учет динамики
  - Лаги
  - Скользящие средние
- Баловство

2.1. Изучение данных
2.1.2. Предобработка данных

EleWise проекты

Сила характеристики

- предсказательная сила каждого атрибута (WOE)
- диапазон и тренд веса факторов (WOE) по сгруппированным атрибутам в рамках одной характеристики
- предсказательная сила характеристики (IV)
- операционные и бизнес-аспекты (регионы, продукты и т.п.)

Дополнительно:

- R-квадрат
- Хи-квадрат

2.1. Изучение данных
2.1.2. Предобработка данных

EleWise проекты

WOE и IV

Цель кредита	Хорошие	Доля хороших	Плохие	Доля плохих	Общее	Доля общих	Доля плохих в характеристике	WOE	IV
автомобиль (подержанный)	70	0,1397	13	0,0653	83	0,1186	0,1566	-0,7605	0,0566
бытовая техника, переподготовка	14	0,0279	3	0,0151	17	0,0243	0,1765	-0,6139	0,0079
ремонт	11	0,022	3	0,0151	14	0,0200	0,2143	-0,3763	0,0026
радио / телевидение	145	0,2894	48	0,2412	193	0,2757	0,2487	-0,1822	0,0088
мебель / оборудование, автомобиль (новый)	197	0,3932	89	0,4472	286	0,4086	0,3112	0,1287	0,0069
бизнес, другие	47	0,0938	27	0,1357	74	0,1057	0,3649	0,3693	0,0155
образование	17	0,0339	16	0,0804	33	0,0471	0,4848	0,8636	0,0402
отдых	0	0	0	0	0	0	0	0	0
Итого	501	1	199	1	700	1			**0,1384**

2.1. Изучение данных
2.1.2. Предобработка данных

WOE и IV

$$IV = \sum \left(S_{B_i} - S_{G_i} \right) \cdot WOE$$

$$WOE = \ln \frac{S_{B_i}}{S_{G_i}}$$

Для непрерывной переменной:

- делим на ~50 равных групп
- изучаем
- перегруппировываем

Значения IV:

меньше 0,02 — предсказательная сила отсутствует

0,02 – 0,1 — предсказательная сила мала

0,1 – 0,3 — предсказательная сила средняя

свыше 0,3 — предсказательная сила высокая

2.1. Изучение данных
2.1.2. Предобработка данных

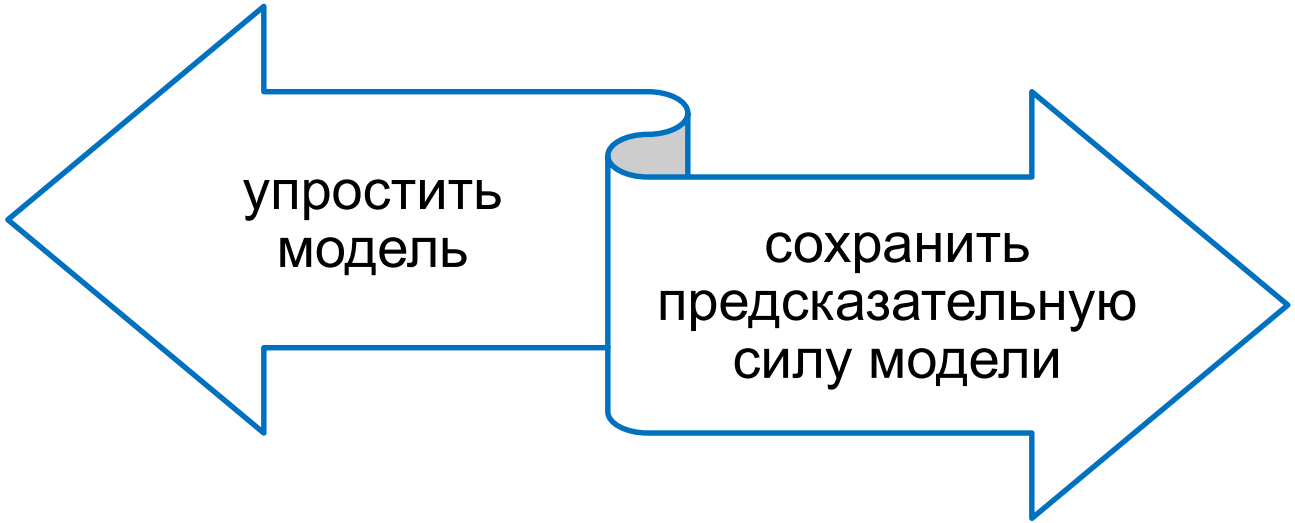
EleWise проекты

Сила характеристики

- корреляция
- мультиколлинеарность (исчезает при большой выборке)
- частичные взаимосвязи

- удаляем переменные, которые не коррелируют с целевой функцией
- удаляем/комбинируем переменные, которые сильно коррелируют между собой
- при разбиении на группы (кластеры) снова считаем корреляцию

2.1. Изучение данных
2.1.2. Предобработка данных



2.1. Изучение данных
2.1.2. Предобработка данных

Резюмируем:

- считать данные
- логическое понимание характеристик
- проанализировать на выбросы
- проанализировать полноту
- статистическое изучение данных:
  распределение, медиана, среднее, мода
- корреляционный анализ
- сила характеристики

2.1. Изучение данных
2.1.2. Предобработка данных

EleWise проекты

В выборке из ста муравьев и одного кита
средняя масса муравья может оказаться больше килограмма.

2.2. Проблема переобучения

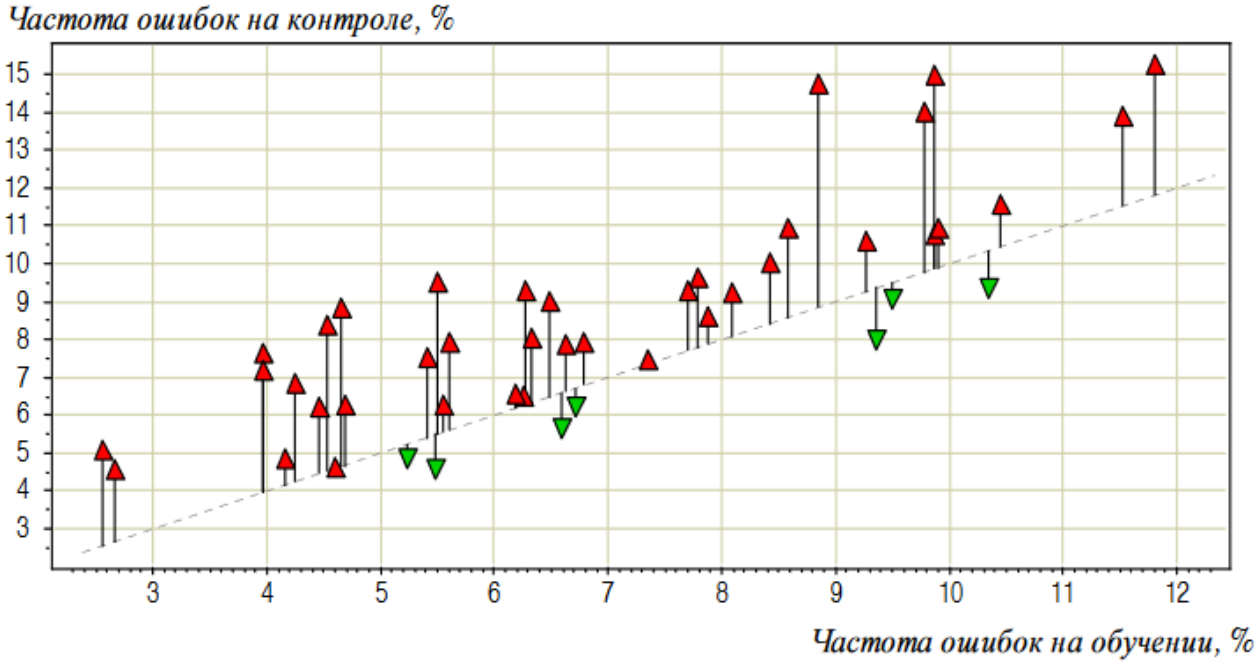
2.2. Проблема переобучения

EleWise проекты

- Представьте, что вы готовитесь к экзамену.
- Вопросы, которые задавались на предыдущем экзамене, и ответы на них выложены в сеть.
- Вы начинаете отвечать на старые вопросы и сравнивать свои ответы с опубликованными. Но вы слишком увлеклись и тратите все свое время на запоминание ответов на старые вопросы.
- Если на предстоящем экзамене будут задаваться только старые вопросы, то все у вас сложится прекрасно.
- Но если материал останется тем же, а вопросы будут другими, то окажется что ваша оценка будет гораздо ниже той, что вы заслужили бы при традиционной подготовке.
- В таком случае можно сказать, что вы переобучились на вопросах прошлых лет и приобретенные знания не обобщаются на вопросы будущего экзамена.

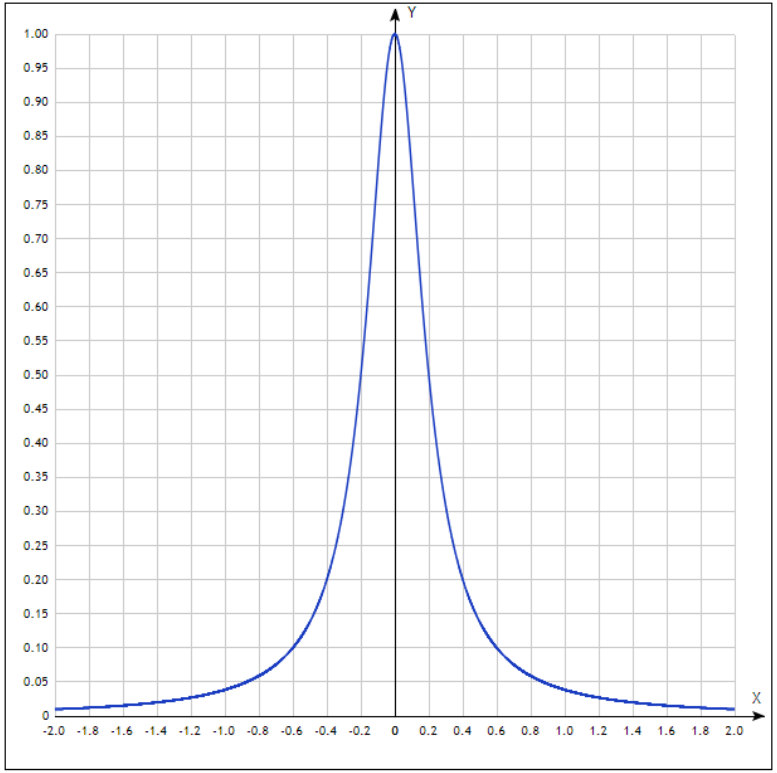
2.2. Проблема переобучения

Задача предсказания отдаленного результата хирургического лечения атеросклероза. Точки - различные алгоритмы.



2.2. Проблема переобучения

Зависимость $y(x) = \frac{1}{1 + 25x^2}$ на отрезке x∈ [-2, 2]



2.2. Проблема переобучения

Зависимость $y(x) = \frac{1}{1+25x^2}$ на отрезке x∈ [-2, 2]

Признаковое описание $x \mapsto (1, x^1, x^2, \ldots, x^n)$.

Модель полиномиальной регрессии

$$a(x, \theta) = \theta_0 + \theta_1 x + \cdots + \theta_n x^n$$ -

полином степени n

Обучение методом наименьших квадратов:

$$Q(a, X^\ell) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \cdots + \theta_n x_i^n - y_i)^2 \to \min_{\theta_0,\ldots,\theta_n}.$$

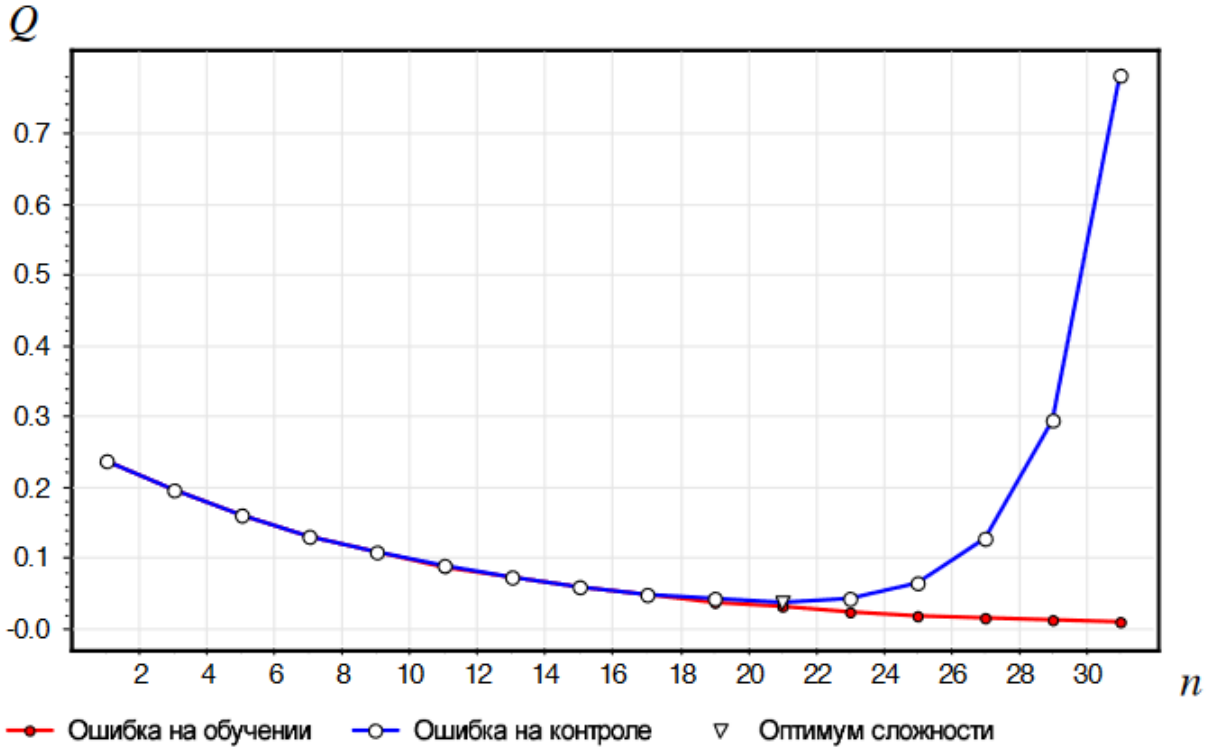
Обучающая выборка: $X^\ell = \left\{ x_i = 4\frac{i-1}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell \right\}$.

Новая выборка: $X^k = \left\{ x_i = 4\frac{i-0.5}{\ell-1} - 2 \;\middle|\; i = 1, \ldots, \ell-1 \right\}$.

Что происходит с $Q(a, X^\ell)$ и $Q(a, X^k)$ при увеличении n?

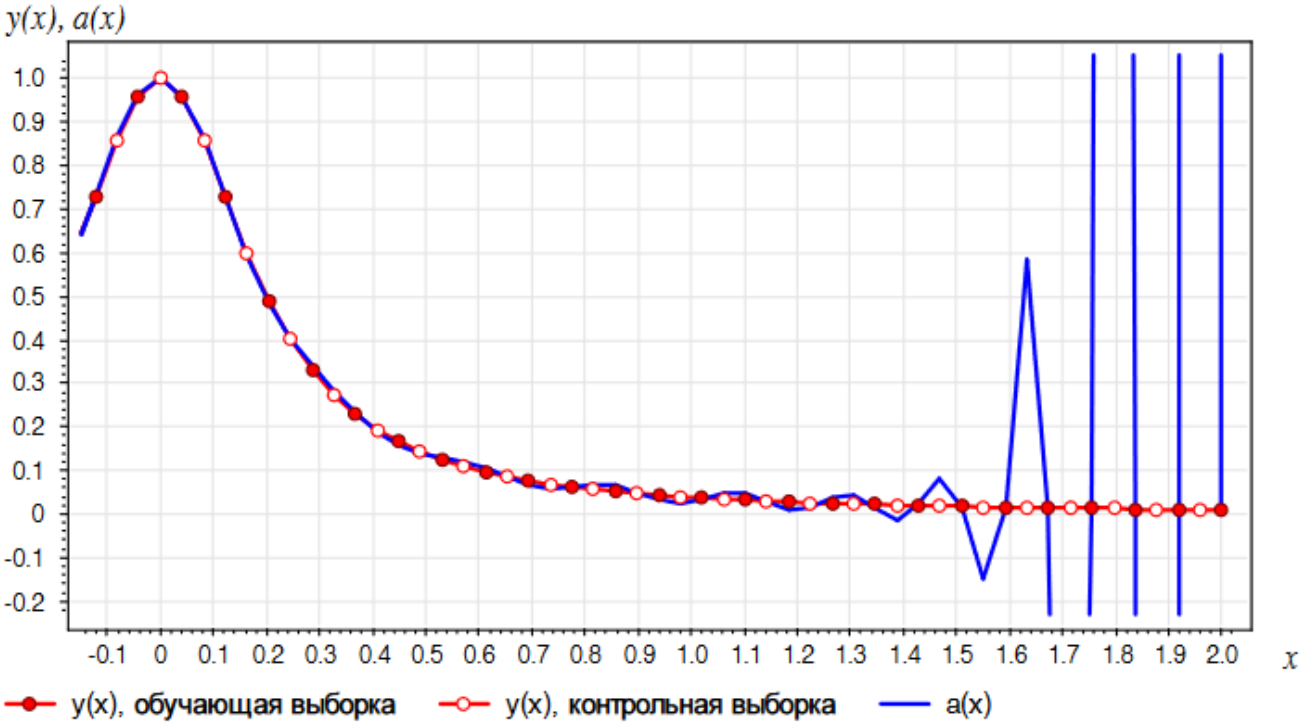
2.2. Проблема переобучения

Переобучение - это когда $Q(\mu(X^\ell), X^k) \gg Q(\mu(X^\ell), X^\ell)$:

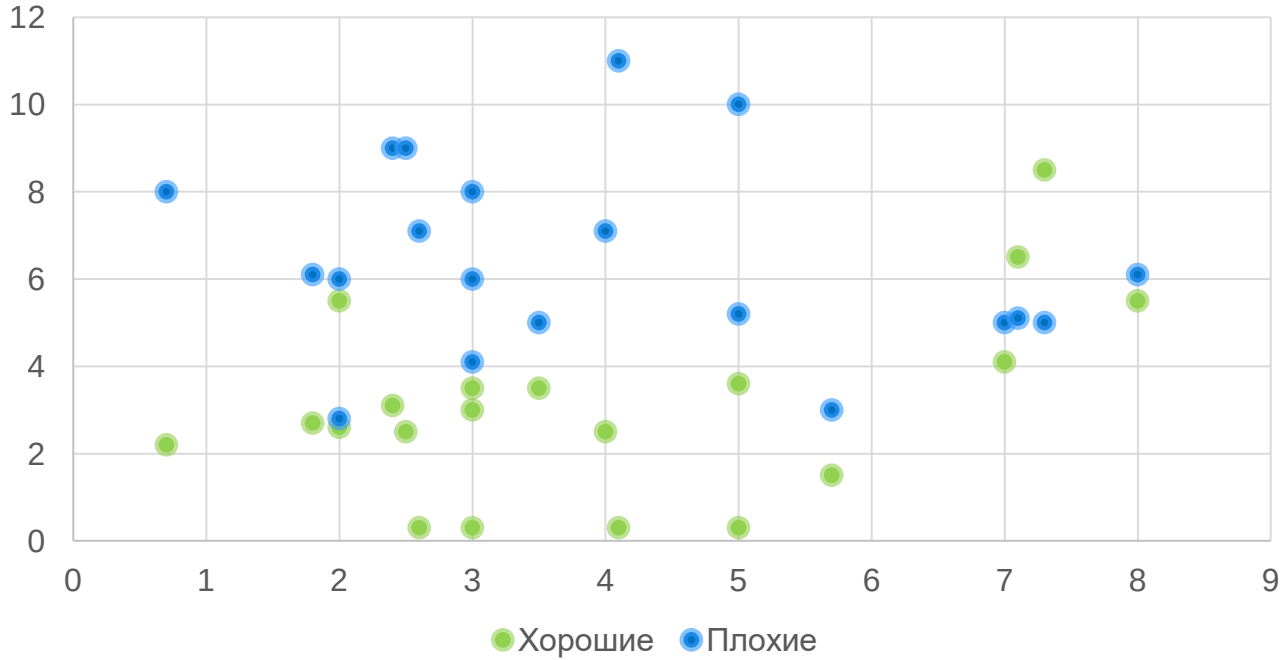


2.2. Проблема переобучения

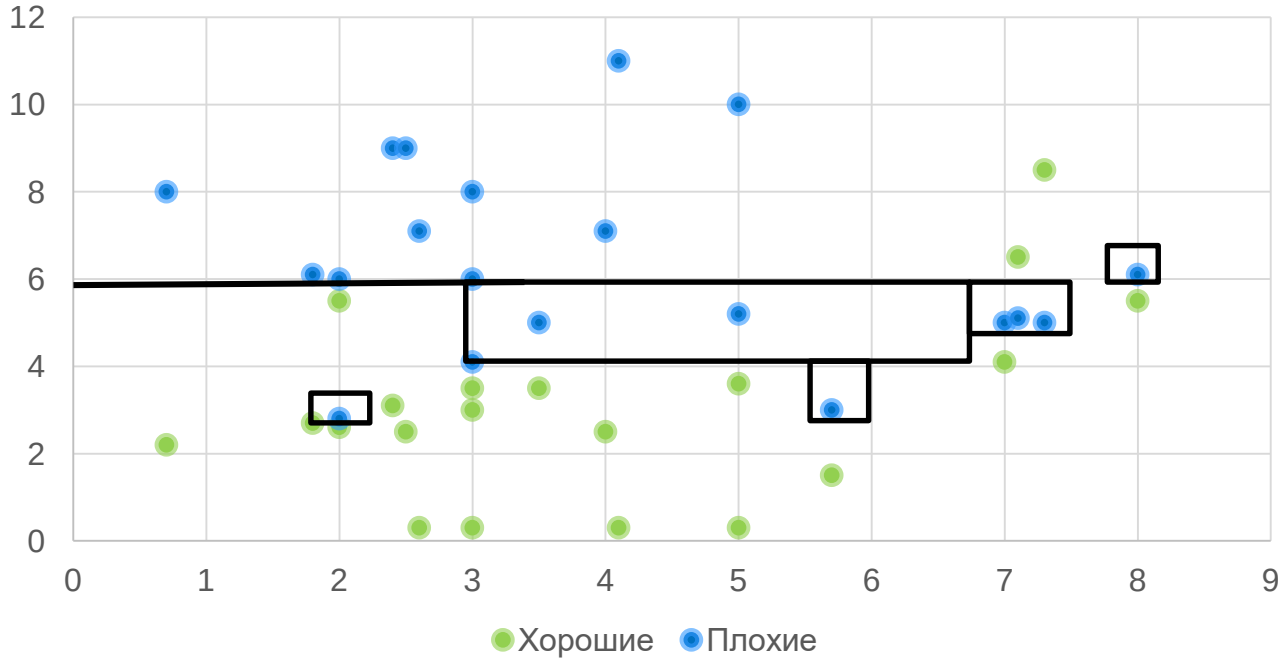
$$y(x) = \frac{1}{1 + 25x^2}; \quad a(x)$$ - полином степени n = 38



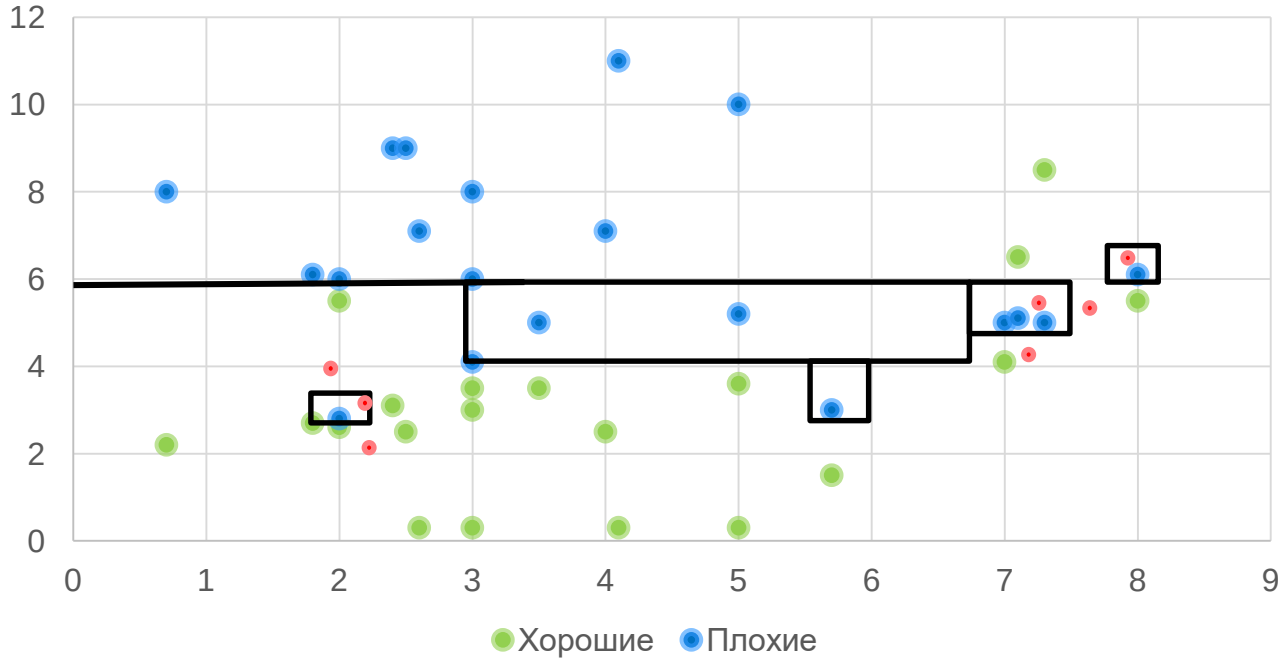
2.2. Проблема переобучения



2.2. Проблема переобучения

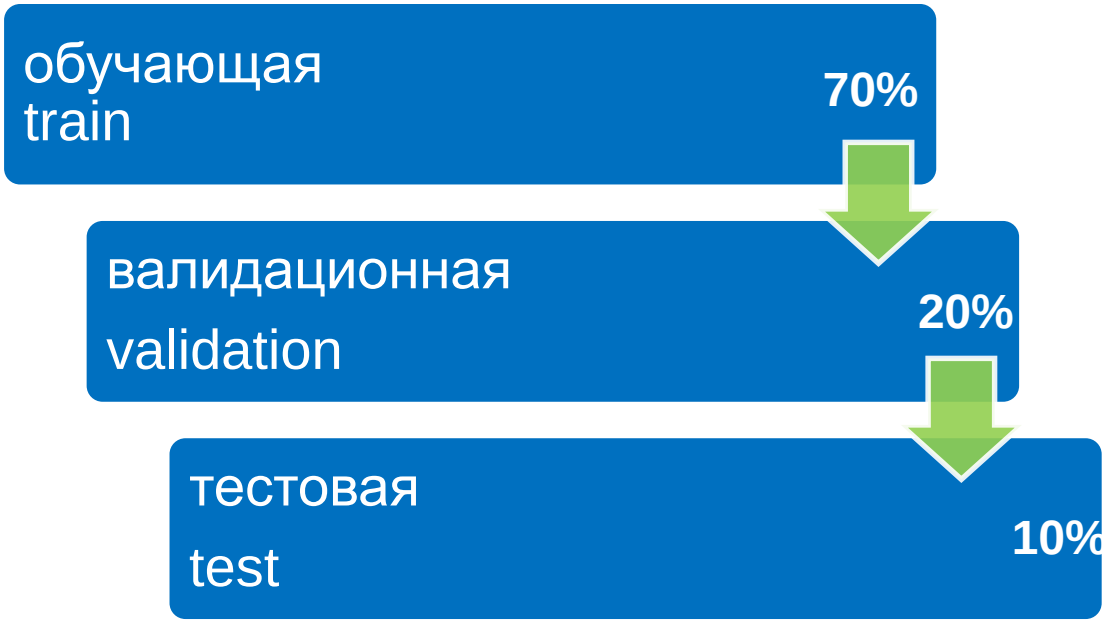


2.2. Проблема переобучения



2.3. Формирование выборки

2.3. Формирование выборки



2.3. Формирование выборки

- качество решения на обучающих данных не волнует –
  мы и так знаем их метки

- волнует, как поведет себя классификатор на будущих данных

2.3. Формирование выборки

Эмпирический риск на тестовых данных (hold-out):

$$\mathrm{HO}(\mu, X^\ell, X^k) = Q(\mu(X^\ell), X^k) \to \min$$

Скользящий контроль (leave-one-out) $L = \ell + 1$

$$\mathrm{LOO}(\mu, X^L) = \frac{1}{L}\sum_{i=1}^{L} \mathscr{L}(\mu(X^L \setminus \{x_i\}), x_i) \to \min$$

Кросс-проверка (cross-validation) по N разбиениям,

$X^L = X_n^\ell \sqcup X_n^k$, $L = \ell + k$:

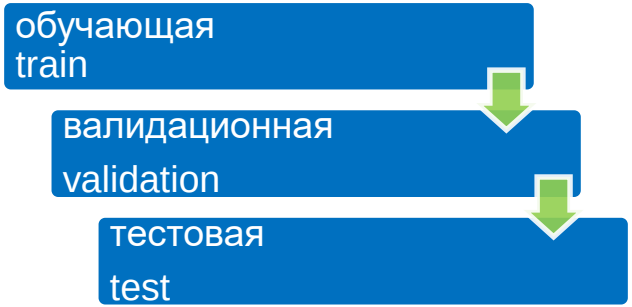
$$\mathrm{CV}(\mu, X^L) = \frac{1}{N}\sum_{n=1}^{N} Q(\mu(X_n^\ell), X_n^k) \to \min$$

2.3. Формирование выборки

Резюмируем:

- считать данные
- логическое понимание характеристик
- проанализировать на выбросы
- проанализировать полноту
- статистическое изучение данных: распределение, медиана, среднее, мода
- корреляционный анализ
- сила характеристики

должно быть однородным на всех выборках



EleWise проекты

2.4. Погладь Python-а

ML or not ML?

ML or not ML?
Преимущества

- Точность
- Автоматизация
- Скорость
- Возможность настройки
- Масштабируемость

ML or not ML?
Недостатки

- 80% времени уходит на то, чтобы привести данные к пригодному для моделирования виду
- Выбор подходящих признаков и оптимального алгоритма требует больших усилий
- «переобучение» - модель идеально работает на обучающих данных, но демонстрирует полную неспособность к предсказанию нового (обобщению)