



# **Признаки и модели**

Евгений Борисов

# признаки и модели

изучаем предметную область

извлекаем признаки из объекта

подбираем преобразования признаков

отбираем хорошие признаки

# признаки и модели

изучаем предметную область

извлекаем признаки из объекта

подбираем преобразования признаков

отбираем хорошие признаки

собираем учебный набор

удаляем выбросы

обучаем модель

тестируем модель

запускаем модель в работу

# признаки и модели

## **собираем и обрабатываем признаки**

feature extraction / feature engineering – формализация данных

feature transformation – трансформация данных

feature selection – отбор наиболее удачных признаков

# признаки и модели

## **feature extraction / feature engineering**

отображение данных, специфических для предметной области,  
в точки пространства признаков

признаки

бинарные (да/нет)

категориальные (ограниченный список значений)

количественные ( $\mathbb{R}$ )

собираем признаки формируем учебный датасет

# признаки и модели

## **feature extraction / feature engineering**

объект → вектор признаков

### примеры

для текстов - TF-IDF, Word2Vec

для изображений - SIFT,  
Haar-like features,  
Histogram of Oriented Gradients (HOG),  
Bag of visual Words (BoW),  
deep convolutional neural networks

# признаки и модели

## **feature transformation**

трансформация данных для улучшения результатов работы модели  
(повышения точности)

# признаки и модели

## **feature transformation**

стандартизация, StandardScaling, Z-score normalization

приведение к нулевому мат.ожиданию ( $\mu$ ) и единичной дисперсии ( $\sigma$ )

$$x := \frac{x - \mu}{\sigma}$$

улучшает ситуацию с выбросами

применяют совместно с метрическими методами



# признаки и модели

## **feature transformation**

MinMaxScaling масштабирование в отрезок [0,1]

похож на StandartScaling,

MinMaxScaling полезен для визуализации,  
легко перенести признаки на отрезок [0, 255]

$$x := \frac{x - x_{min}}{x_{max} - x_{min}}$$

# признаки и модели

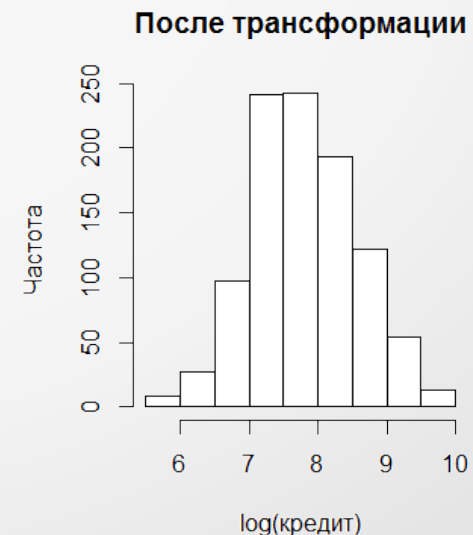
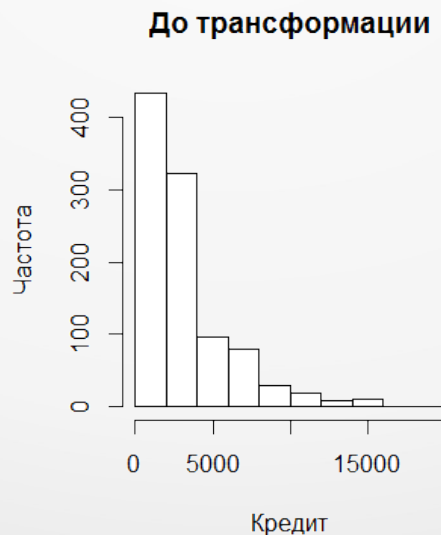
## feature transformation

log-трансформация

многие модели хорошо работают с нормально распределёнными данными (параметрические методы)

если данные описываются распределением похожим на логнормальное то их можно привести к распределению близкому к нормальному

$$x := \log(x)$$



# признаки и модели

## feature transformation

метод пространственных знаков (spatial sign)

проецирует значения на поверхность многомерной сферы, данные становятся равноудаленными от центра этой сферы

$$x_{ij} := \frac{x_{ij}}{\sqrt{\sum_{k=1}^P x_{ik}^2}}$$

i- примеры,  
j - признаки,  
P - количество признаков

**применяется после стандартизации признаков**

# признаки и модели

## **зависимость признаков**

мультиколлинеарность - наличие линейной зависимости у признаков

зависимость признаков

не позволяет однозначно оценить параметры модели

# признаки и модели

## **обработка пропусков**

удалить объект из выборки

заполнить средним (медианой) вещественных переменных

заполнить наиболее частым значением для категориальных

заменить пропуск на редкое (мало вероятное) значение

заменить на соседнее значение для упорядоченных данных

# признаки и модели

## **отбор признаков и метрики качества**

(считаем на тестовом наборе)

- погрешность (accuracy)
- матрица ошибок ( confusion matrix )
- точность (precision)
- полнота (recall)
- F-мера
- ROC/AUC

# признаки и модели

## **метрики качества**

погрешность (accuracy)

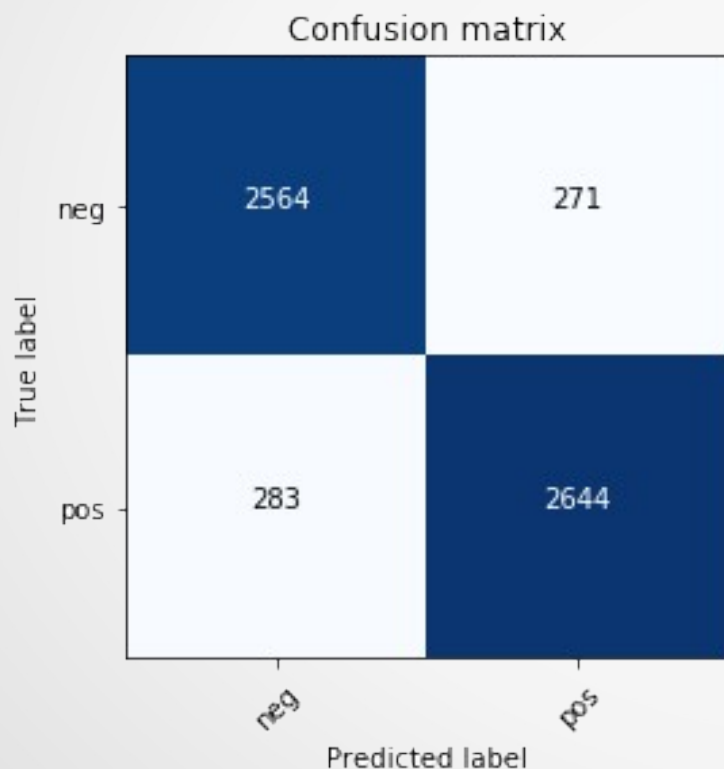
правильные ответы / всего примеров

оценка для сбалансированного набора, т.е.  
количество примеров в классах +- одинаковое

# признаки и модели

## метрики качества

матрица ошибок ( confusion matrix )



два класса — четыре группы

- TP истинно положительные
- TN истинно отрицательные
- FP ложно положительные
- FN ложно отрицательные



# признаки и модели

## метрики качества

точность (precision)

$$TP / ( TP + FP )$$

(метрики для отдельного класса)

доля объектов действительно принадлежащих данному классу относительно всех объектов, которые классификатор отнес к этому классу

полнота (recall)

$$TP / ( TP + FN )$$

доля объектов, найденных классификатором, относительно всех объектов этого класса

F-мера

$$( precision * recall ) / ( precision + recall )$$

усреднение точности и полноты

# признаки и модели

## метрики качества

**Пример** *classification\_report*

	precision	recall	f1-score	support
0	0.90	0.90	0.90	2835
1	0.91	0.90	0.91	2927
avg / total	0.90	0.90	0.90	5762

# признаки и модели

## **методы отбора признаков**

цель: для минимизации ошибки модели на контроле

полный перебор подмножеств признаков

добавление признаков по одному с минимизацией ошибки (жадный)

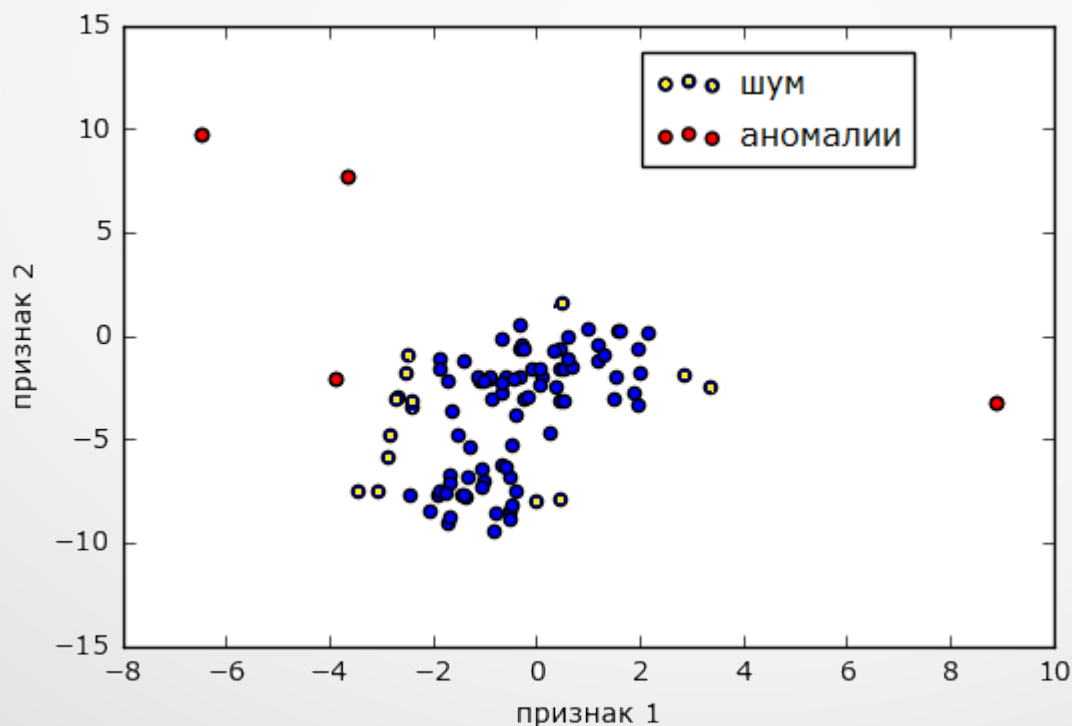
поочерёдное добавление/удаление

# признаки и модели

## поиск выбросов / Outlier Detection

выброс или аномалия это то, что не вписывается в общие правила

задачи детектирования аномалий  
не имеют единой формальной постановки



# признаки и модели

## поиск выбросов / Outlier Detection

Статистические тесты (признаки обрабатываем отдельно)  
простой метод - отсечение по квантили 0.95

Модельные тесты - строим модель данных,  
точки, которые сильно отклоняются от модели - аномалии

Итерационные методы - на каждой итерации удаляем группу  
«подозрительных» объектов (последовательное удаление выпуклы  
оболочек).

Метрические методы - у выброса мало соседей

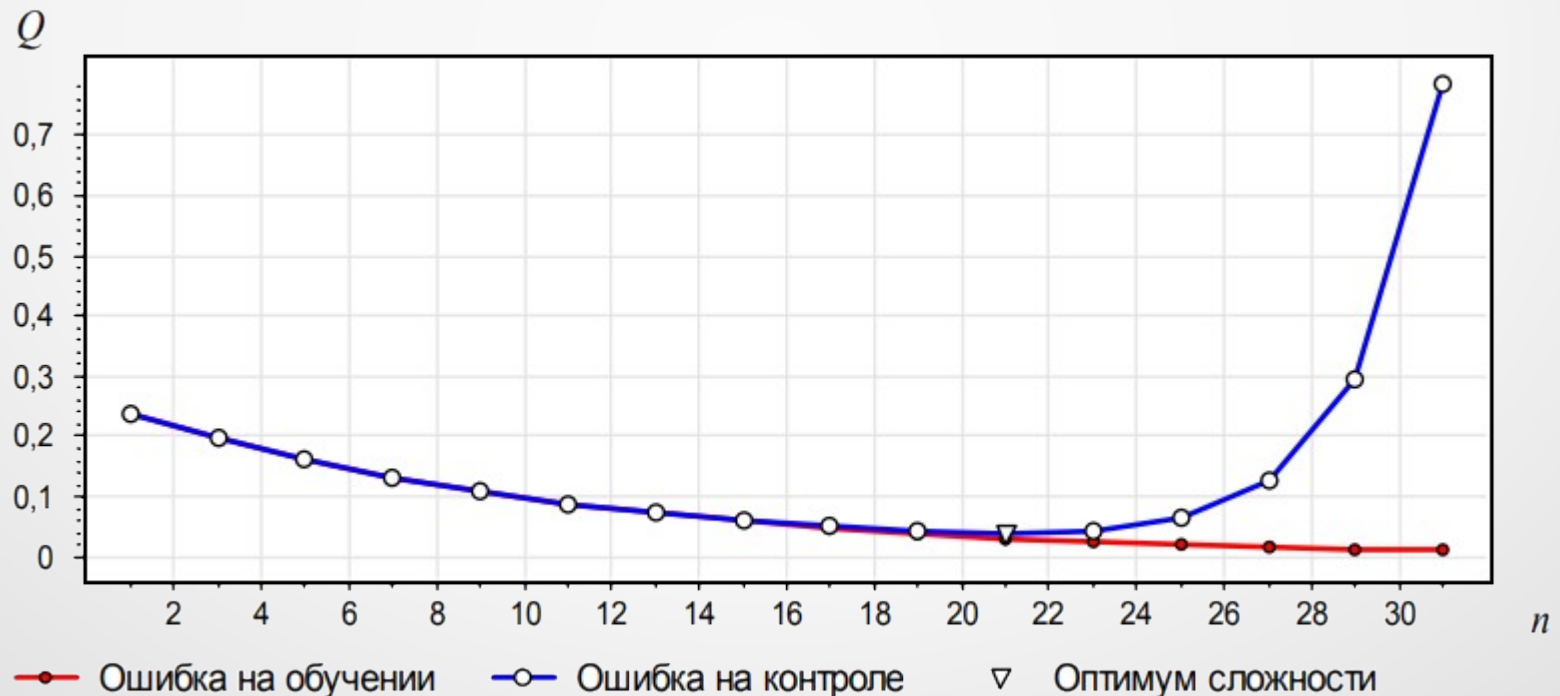
Методы машинного обучения - IsolationForest,  
при построении деревьев выбросы будут попадать в листья на  
ранних этапах (на небольшой глубине дерева)

# признаки и модели

## оценка и выбор моделей

формируем 3 набора: учебный / контрольный / тестовый

обучаем на учебном  
проверяем на контрольном  
итоговый тест на тестовом



# признаки и модели

## оценка и выбор моделей

кроссвалидация (CV)

скользящий контроль - Leave One Out (LOO CV)

$$\text{LOO}(\mu, X^L) = \frac{1}{L} \sum_{i=1}^L Q_{\mu}(X^L \setminus \{x_i\}, \{x_i\})$$

вынимаем пример из учебной выборки  
обучаем модель без него  
проверяем ошибку на этом примере

LOO CV это долго

повторяем для всех объектов выборки  
результат суммируем

# признаки и модели

## оценка и выбор моделей

кроссвалидация (CV)

q-fold CV

аналогично LOO, но будем вместо одного объекта использовать подмножество из q объектов

$$CV_q(\mu, X^L) = \frac{1}{q} \sum_{n=1}^q Q_{\mu}(X^L \setminus X_n^{\ell_n}, X_n^{\ell_n})$$

оценка зависит от разбиения  
на подмножества примеров



# признаки и модели

## оценка и выбор моделей

кроссвалидация (CV)

t x q-fold CV

t раз выполняем q-fold CV,  
учебный набор t раз случайно разбиваем на q блоков

$$CV_{t \times q}(\mu, X^L) = \frac{1}{t} \sum_{s=1}^t \frac{1}{q} \sum_{n=1}^q Q_{\mu}(X^L \setminus X_{sn}^{\ell_n}, X_{sn}^{\ell_n}).$$

# признаки и модели

## Литература

git clone [https://github.com/mechanoid5/ml\\_lectorium.git](https://github.com/mechanoid5/ml_lectorium.git)

К.В. Воронцов Обобщающая способность. Методы отбора признаков. - курс "Машинное обучение" ШАД Яндекс 2014

Александр Дьяконов Поиск аномалий <https://dyakonov.org>

<http://www.machinelearning.ru>

# признаки и модели



**Вопросы ?**

# признаки и модели



Конкурс BigData от Beeline

<https://special.habrahabr.ru/beeline/>



Александр Куменко Как я победил в конкурсе BigData от Beeline

7 ноября 2015

<https://habr.com/post/270367/>