

ЗАНЯТИЕ 1.4

БАЗОВЫЕ АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ В SKLEARN



ДЕНИС
КИРЬЯНОВ

Сбербанк



[@kirdin](https://www.telegram.me/kirdin)

ЦЕЛИ ЗАНЯТИЯ

В КОНЦЕ ЗАНЯТИЯ ВЫ НАУЧИТЕСЬ:

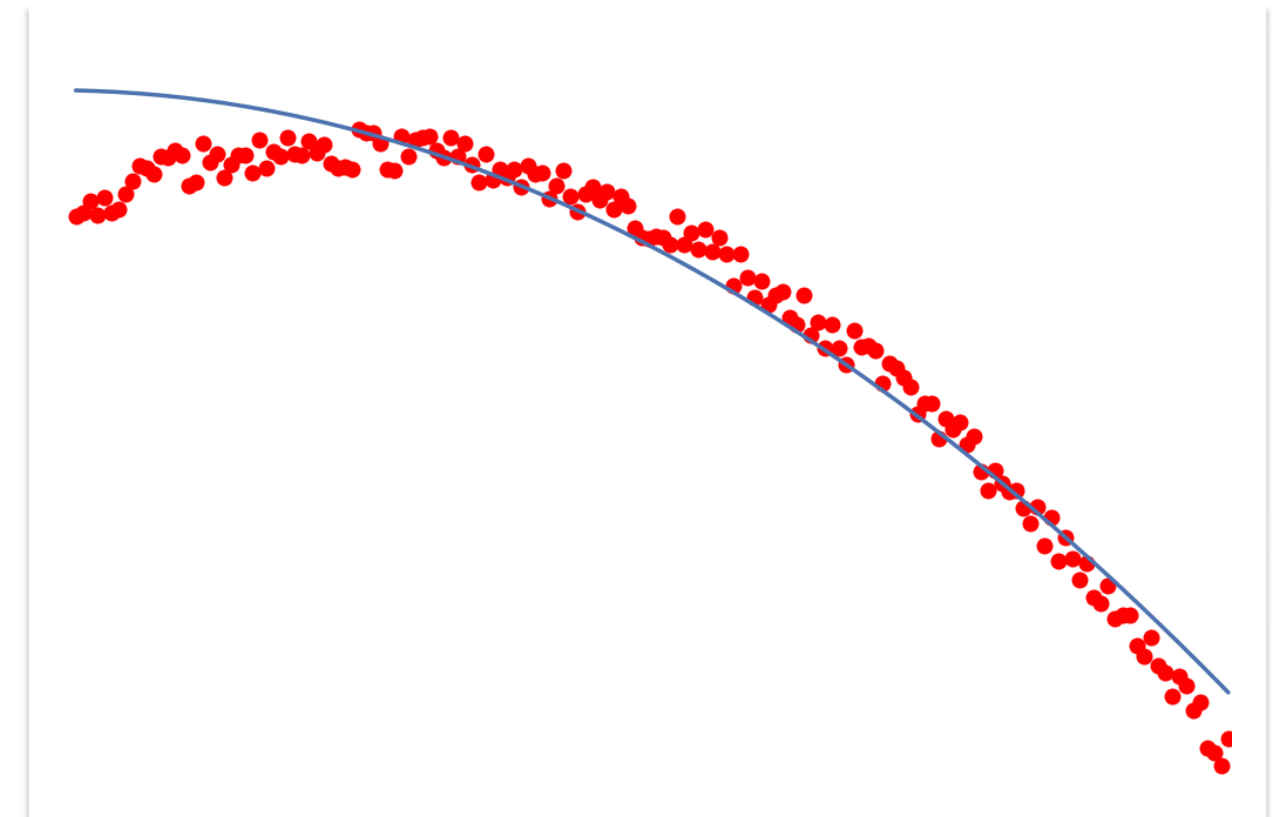
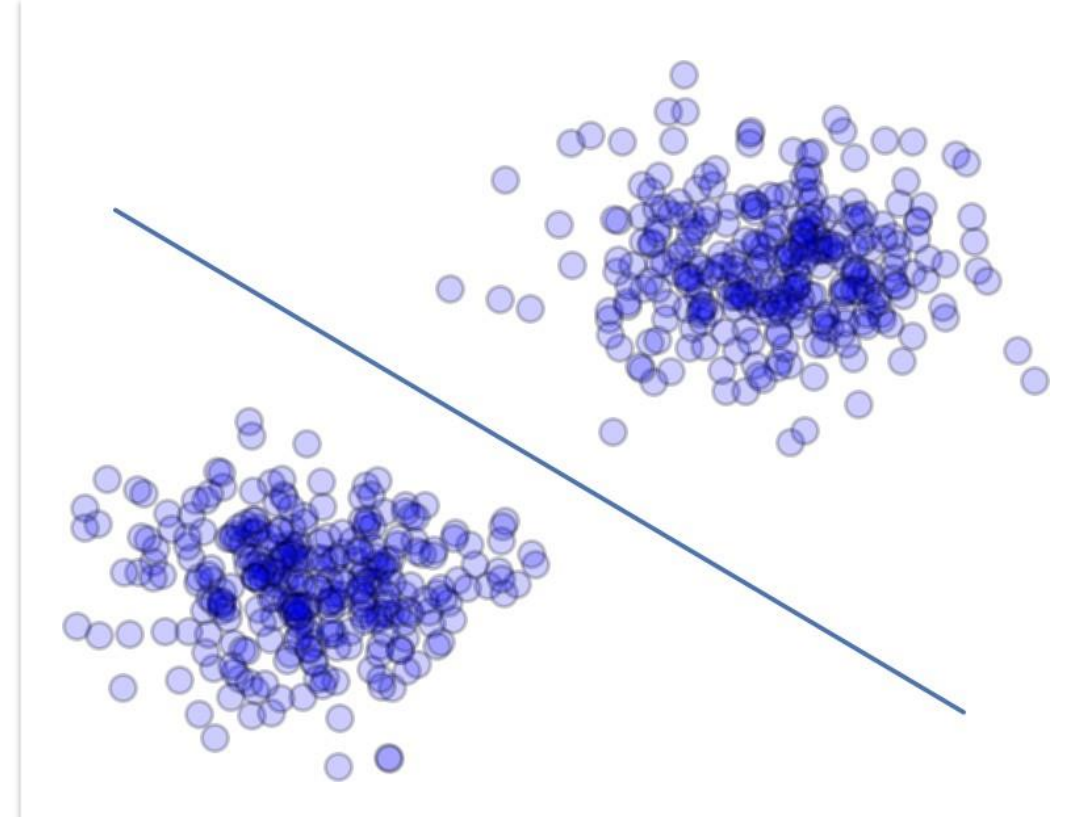
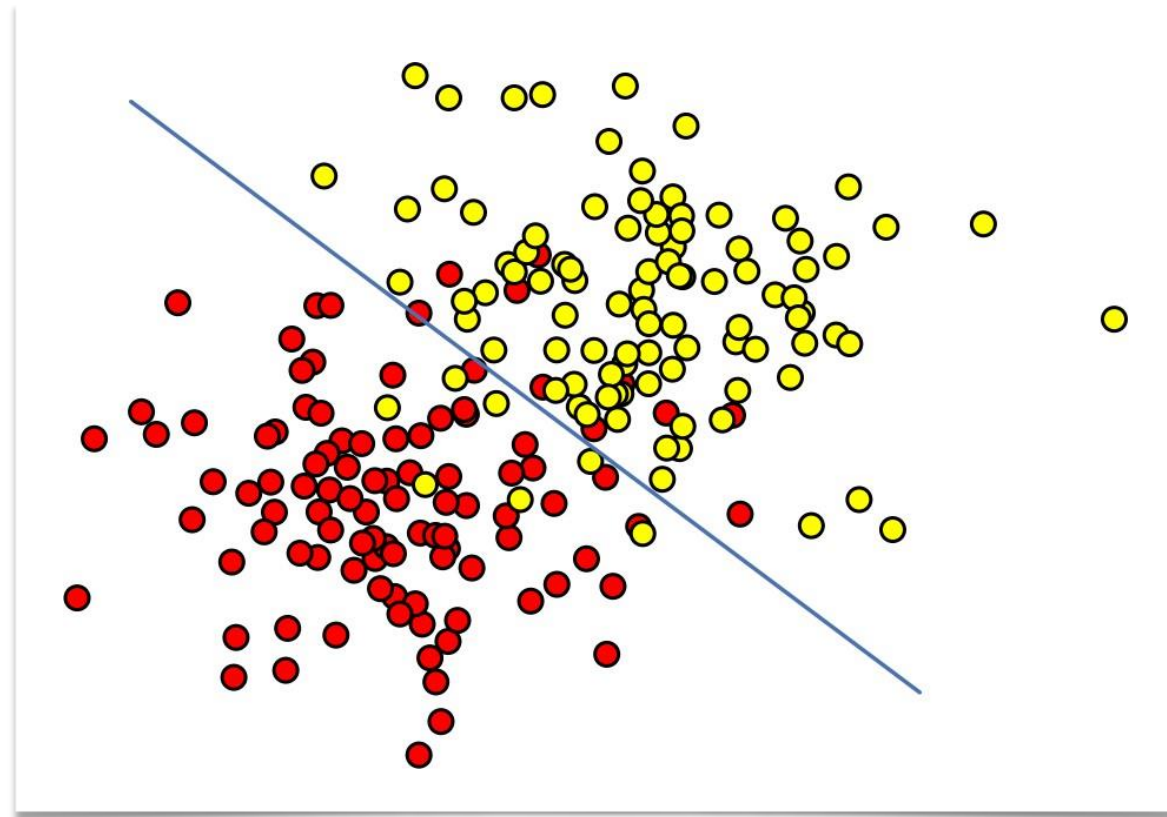
- **решать основные задачи машинного обучения** при помощи реализованных в sklearn методах
- **оценивать качество** решения
- **предобрабатывать данные и подбирать параметры** моделей для улучшения качества решения

О ЧЁМ ПОГОВОРИМ И ЧТО
СДЕЛАЕМ

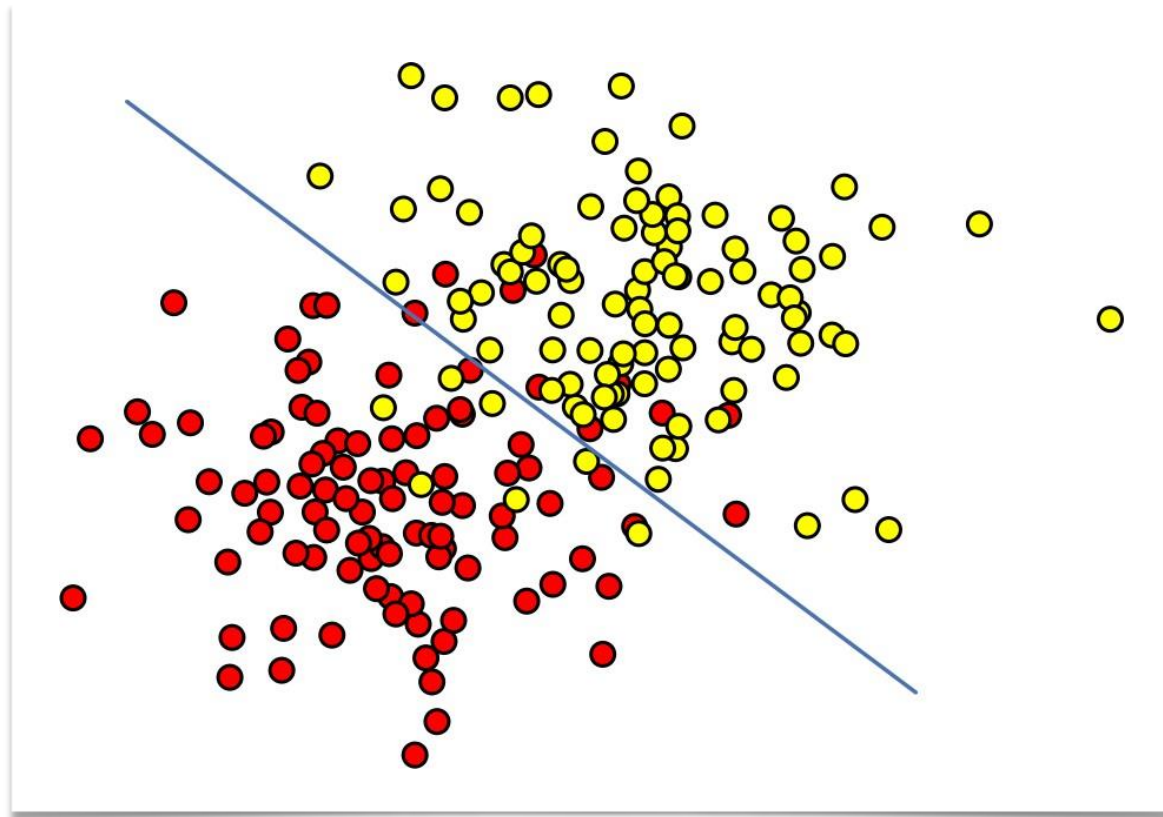
-
1. Вспомним **типы задач**, решаемые ML
 2. Обзорно познакомимся с **различными методами**, реализованными в sklearn
 3. **На практике** используем несколько из них
 4. Разберёмся, как можно **улучшить качество** решения при помощи sklearn
 5. Отработаем это **на практике**

1. БИБЛИОТЕКА SCIKIT-LEARN

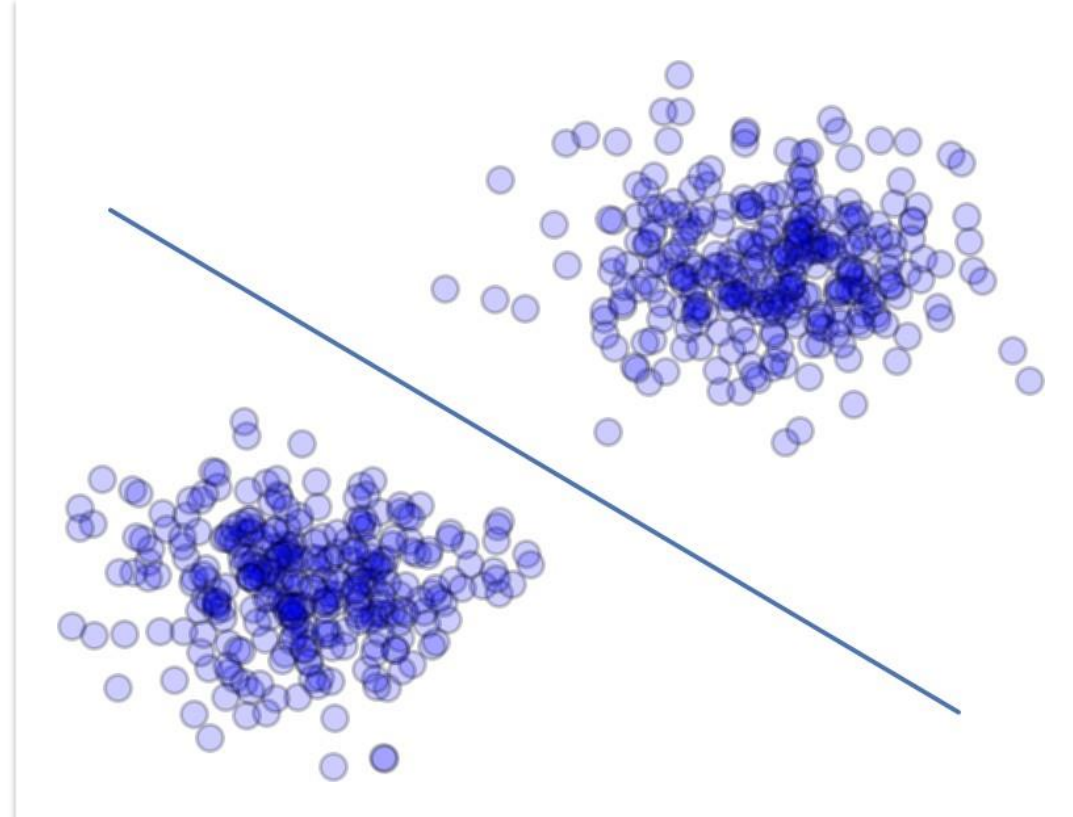
ТИПЫ ЗАДАЧ



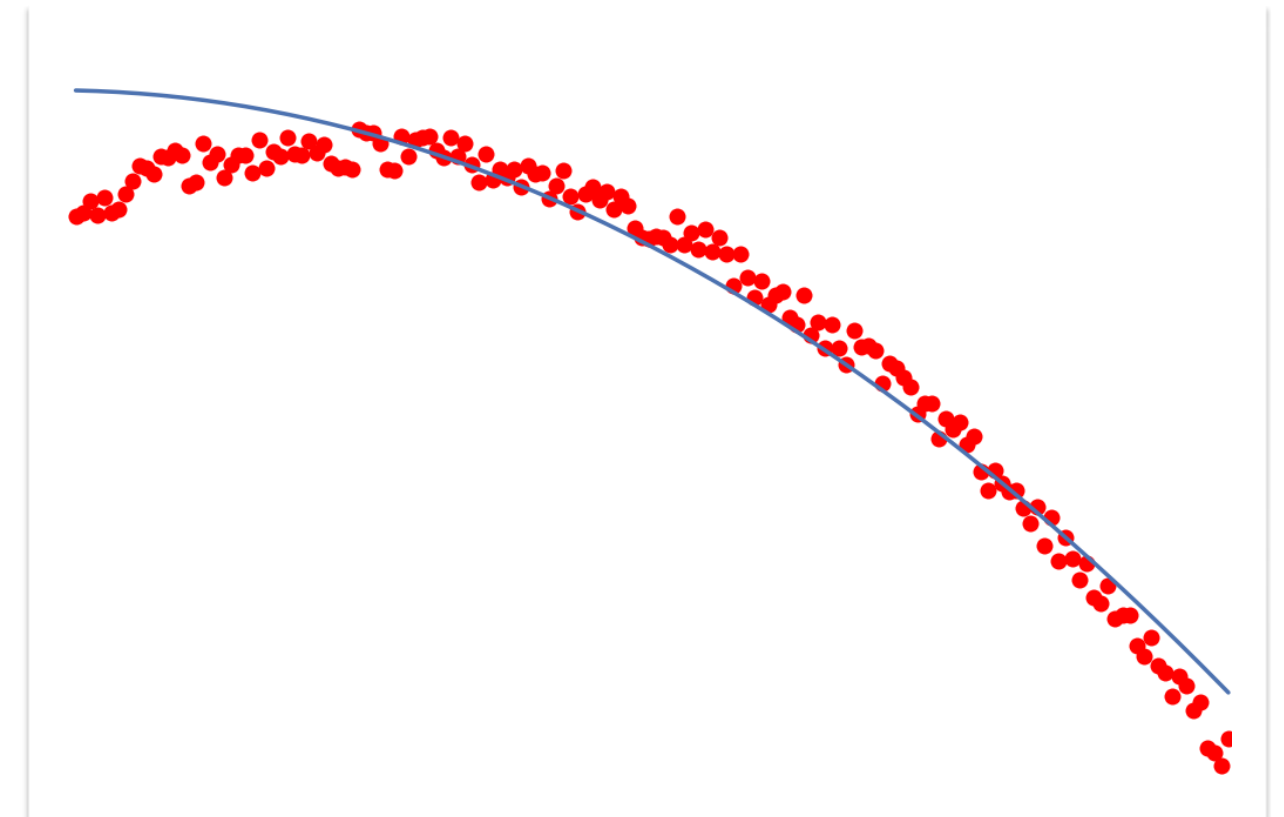
ТИПЫ ЗАДАЧ



классификация
есть разметка: X, y

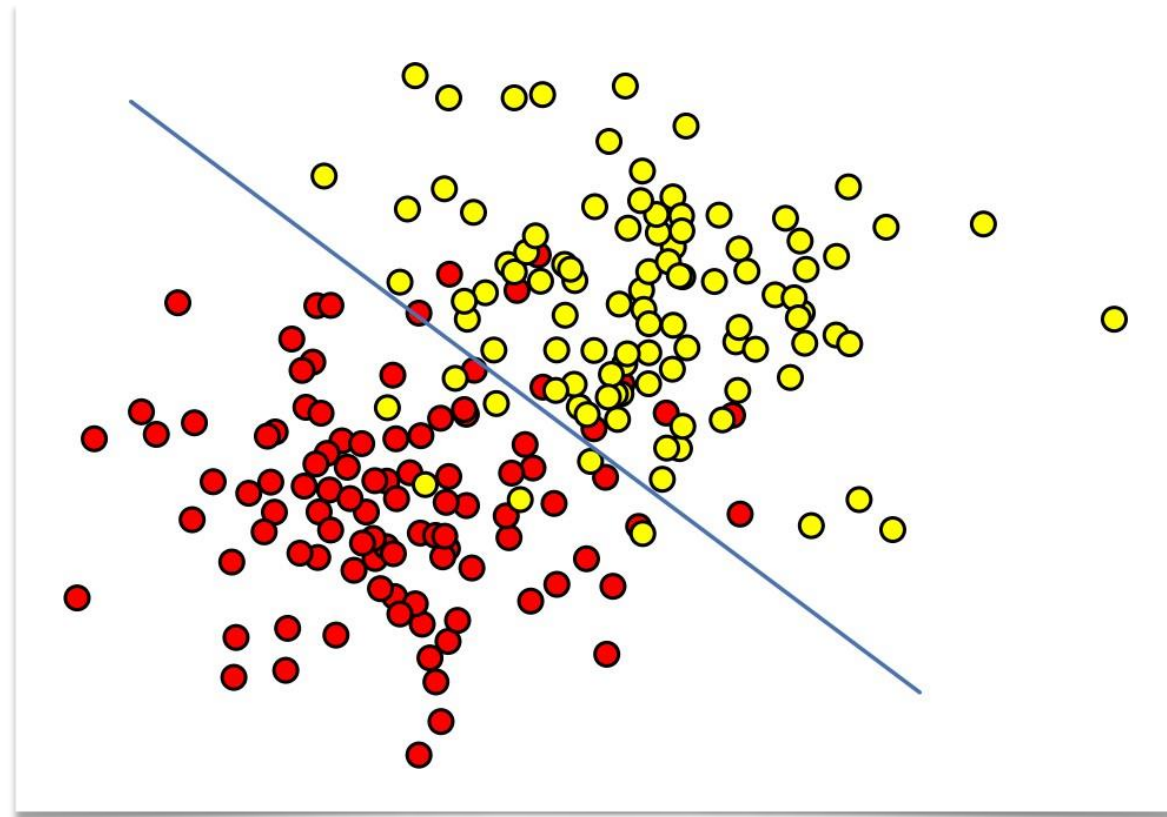


кластеризация
нет разметки: X



регрессия
есть разметка: X, y

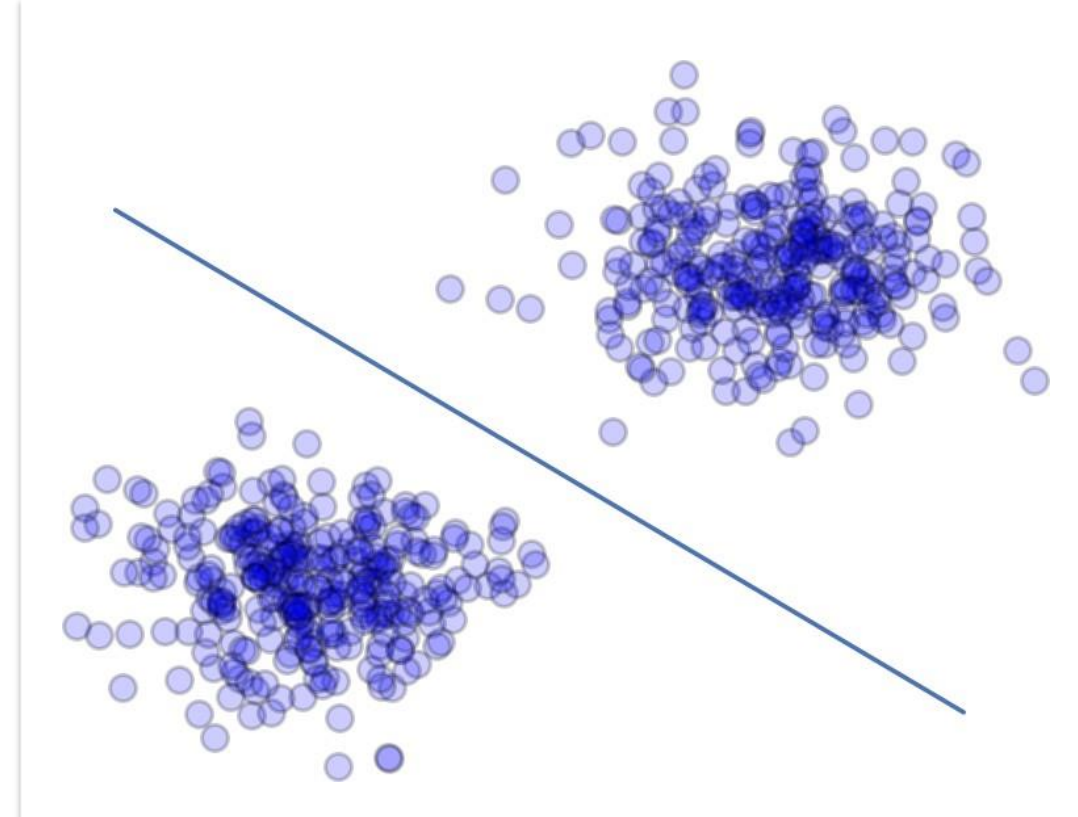
ТИПЫ ЗАДАЧ



классификация

есть разметка: X, y

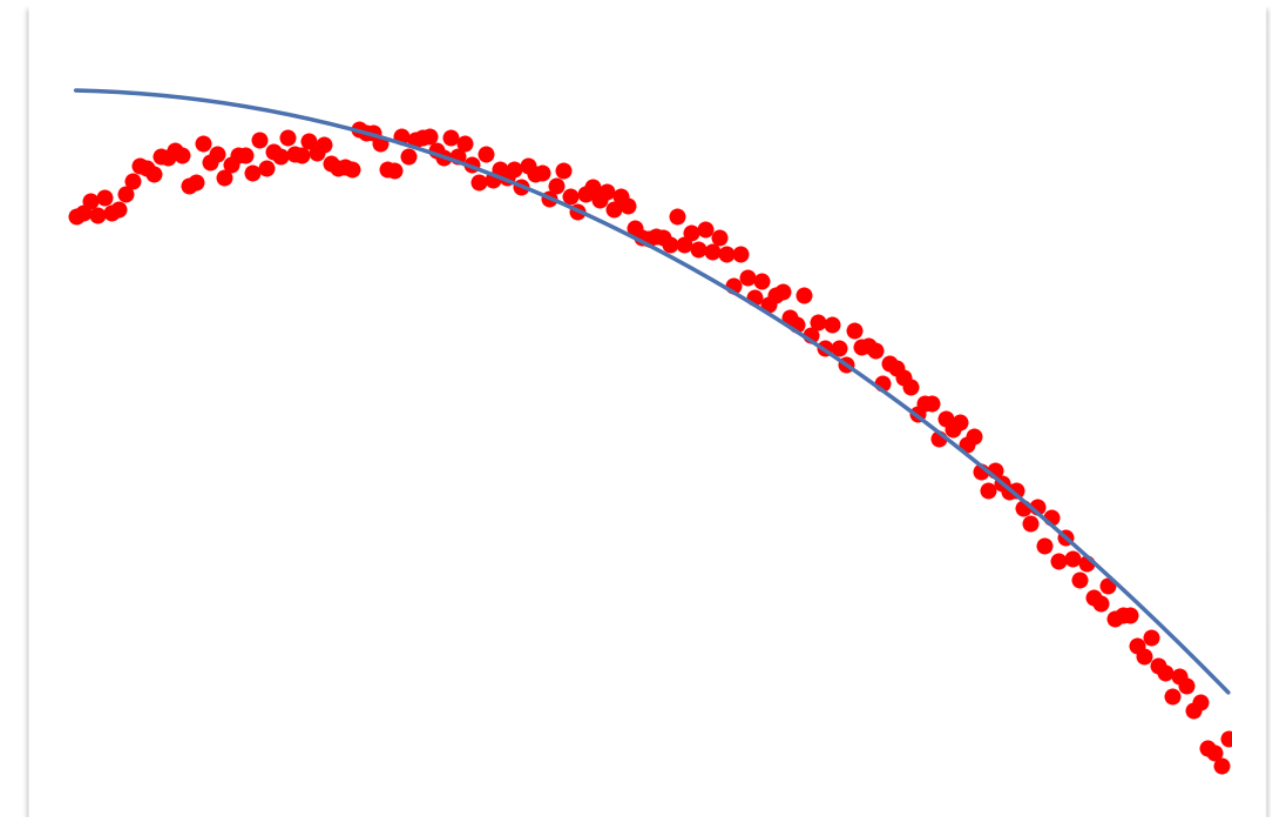
y - перечислимо



кластеризация

нет разметки: X

y - перечислимо



регрессия

есть разметка: X, y

y в непрерывном диапазоне

SKLEARN - 👍

The screenshot shows the scikit-learn website homepage. At the top, there's a navigation bar with links: Home, Installation, Documentation, and Examples. A search bar is also present. The main banner features the scikit-learn logo and the tagline "Machine Learning in Python". Below the banner, there's a grid of 12 small images showing various machine learning visualizations. To the right of the grid, there's a list of features: Simple and efficient tools for data mining and data analysis, Accessible to everybody, and reusable in various contexts, Built on NumPy, SciPy, and matplotlib, and Open source, commercially usable - BSD license.

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.

Algorithms: SVM, nearest neighbors, random forest, ... — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, ridge regression, Lasso, ... — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, ... — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning

Modules: grid search, cross validation, metrics. — Examples

Preprocessing

Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.

Modules: preprocessing, feature extraction. — Examples

Набор логически
разделённых модулей

Единообразный API
взаимодействия

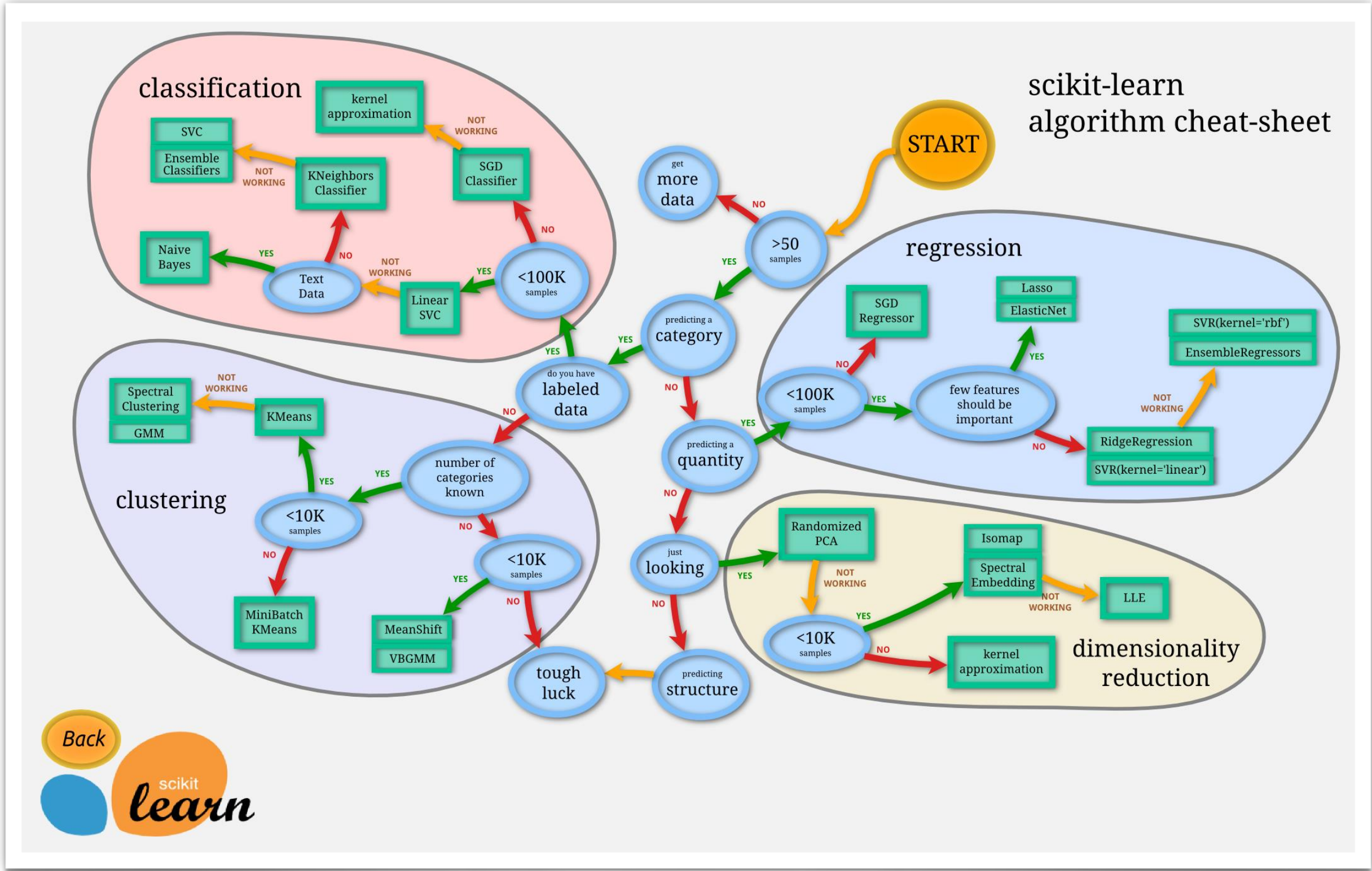
fit + transform + predict

Отличная документация с
примерами и описанием
работы алгоритмов

ЧТО ЕЩЁ НАДО ЗНАТЬ?

- ▶ Обученные модели **можно сохранять**
- ▶ Для обучения данные должны содержаться **целиком в оперативной памяти**
- ▶ Внутри python + cython,
через rpycharm, например, можно посмотреть, что внутри :)
- ▶ Для работы необходимы **numpy / pandas**

SKLEARN ALGO CHEATSHEET



2. МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ SCIKIT-LEARN

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

linear_model - линейные модели

- ▶ [LinearRegression](#)
- ▶ [LogisticRegression](#)

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

tree - дерево решений

- ▶ [DecisionTreeClassifier](#)
- ▶ [DecisionTreeRegressor](#)

ensemble - ансамбли решений: лес, бустинг

- ▶ [RandomForestClassifier](#)
- ▶ [GradientBoostingClassifier](#)

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

cluster - различные методы кластеризации

▶ [KMeans](#), [MiniBatchKMeans](#)

▶ [DBSCAN](#)

▶ [AffinityPropagation](#)

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ML

```
from sklearn.linear_model import LinearRegression  
X, y = ...
```

1. `model = LinearRegression(fit_intercept=True)`

2. `model.fit(X, y)`

3. `a = model.predict(X)`

(если это классификация, то есть также и `predict_proba`)

оценка `a` должна приближать `y`

3. ПРАКТИЧЕСКОЕ ЗАДАНИЕ - 1

ПРАКТИЧЕСКОЕ ЗАДАНИЕ - 1

1. Загрузить данные по недвижимости Бостона
2. Разделить их на 2 части: обучающую и тестовую выборки
3. Сделать предсказание по тестовой выборке
4. Сравнить реальные значения с предсказанием

4. ДРУГИЕ ПОЛЕЗНЫЕ ФУНКЦИИ SCIKIT-LEARN

ОЦЕНКА КАЧЕСТВА

metrics - различные метрики качества решений

▶ [classification report](#)

▶ [mean squared error](#)

ПОДБОР ПАРАМЕТРОВ МОДЕЛИ

model_selection - оценка качества + подбор гиперпараметров моделей

▶ [GridSearchCV](#)

▶ [cross_val_score](#)

ПРЕДОБРАБОТКА ДАННЫХ

preprocessing - нормировка

▶ [StandardScaler](#)

feature_extraction.text - векторизация

▶ [CountVectorizer](#)

▶ [TfidfVectorizer](#)

▶ [HashingVectorizer](#)

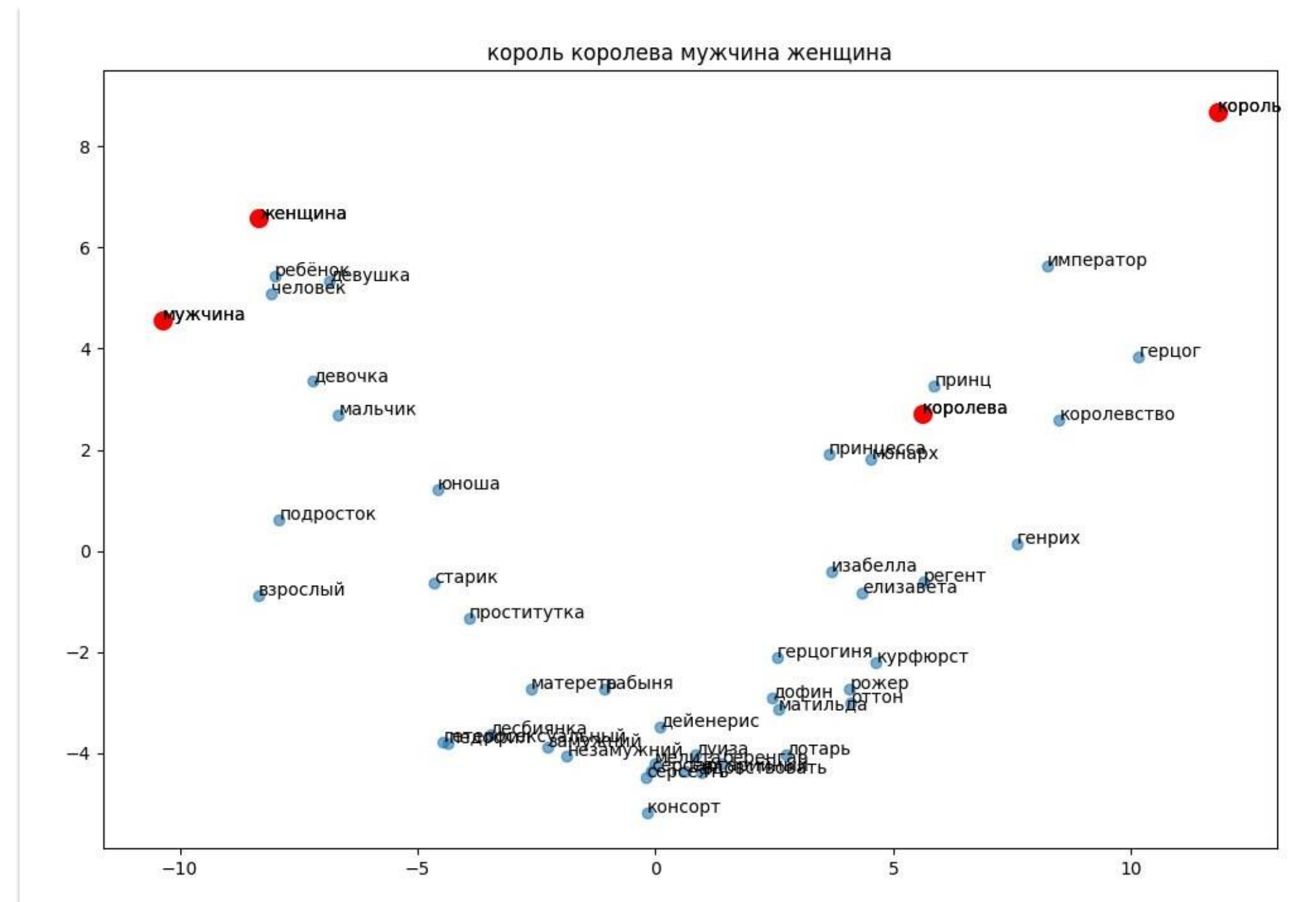
СНИЖЕНИЕ РАЗМЕРНОСТИ

decomposition - разложение матриц и снижение размерности

▶ [PCA](#)

▶ [TruncatedSVD](#)

СНИЖЕНИЕ РАЗМЕРНОСТИ



Пример:

РСА по векторам word2vec
для визуализации:
из 300-мерного пространства
в 2-мерное

5. ПРАКТИЧЕСКОЕ ЗАДАНИЕ - 2

ПРАКТИЧЕСКОЕ ЗАДАНИЕ - 2

1. Взять данные со соревнования [Титаник](#)
2. Перевести всё в числовой вид
3. Заполнить пропуски и отсортировать данные
4. При помощи кросс-валидации найти оптимальный параметр для логистической регрессии
5. Лучшей выбранной моделью оценить качество на отложенной выборке
6. Сделать предсказание по тестовой выборке

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

1. Scikit-learn - open-source библиотека для решения задач машинного обучения, содержащая различные методы решения со схожим набором методов для работы
2. Также в ней содержится набор методов для предобработки выборки, подбора гиперпараметров модели и оценки качества построенного решения
3. Библиотека имеет хорошую документацию и удобна в использовании

ПОЛЕЗНЫЕ МАТЕРИАЛЫ

ПОЛЕЗНЫЕ МАТЕРИАЛЫ

1. [Документация sklearn](#)
2. [Sklearn cheatsheet](#)



НЕТОЛОГИЯ
групп

Спасибо за внимание!

НИКИТА КУЗНЕЦОВ



oychorange@gmail.com



[@NikitaKuznetsov](https://www.instagram.com/NikitaKuznetsov)