

БАЗОВЫЕ АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ В SCIKIT-LEARN

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ

Вопросы занятия

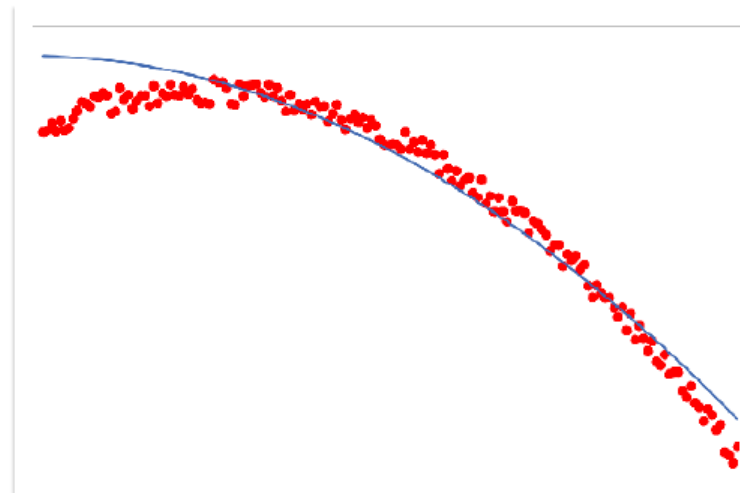
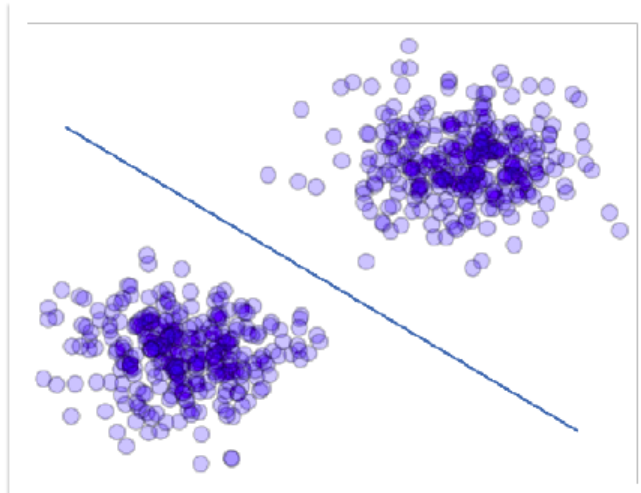
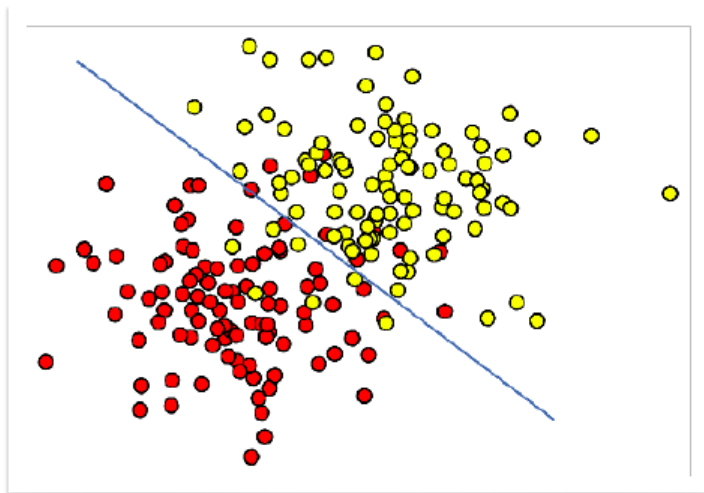
1. Вспомним типы задач, решаемые ML
2. Обзорно познакомимся с различными методами, реализованными в *sklearn*
3. На практике используем несколько из них
4. Разберёмся, как можно улучшить качество решения при помощи *sklearn*
5. Отработаем это на практике

В конце занятия научимся:

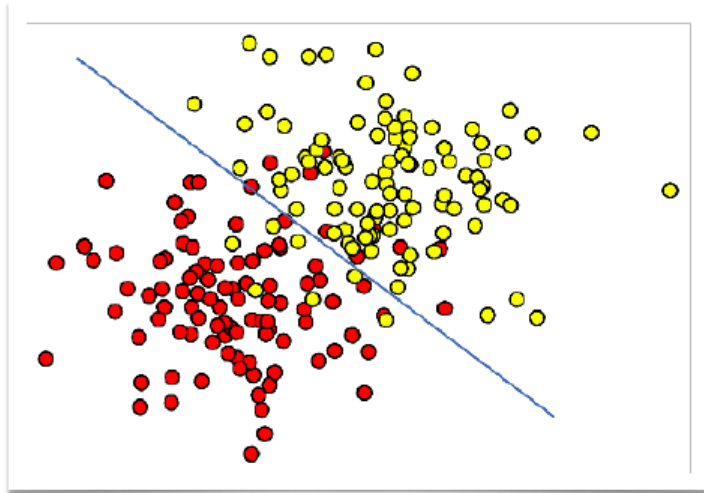
- решать основные задачи машинного обучения при помощи реализованных в `scikit-learn` методах;
- оценивать качество решения;
- предобрабатывать данные и подбирать параметры моделей для улучшения качества решения.

БИБЛИОТЕКА SCIKIT-LEARN

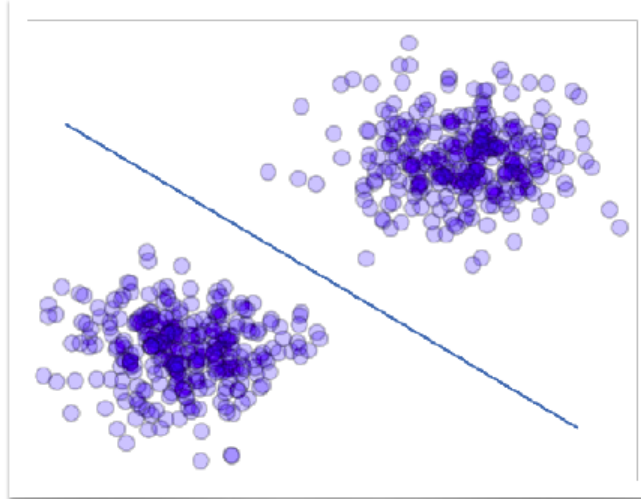
ТИПЫ ЗАДАЧ



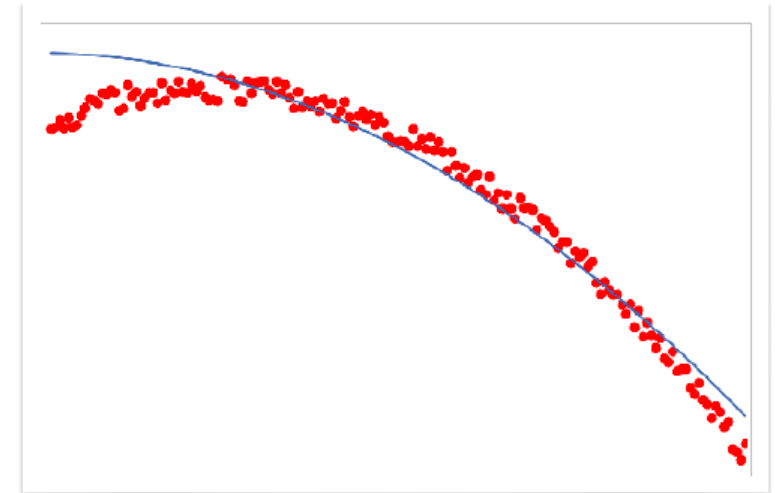
ТИПЫ ЗАДАЧ



классификация
есть разметка: X, y

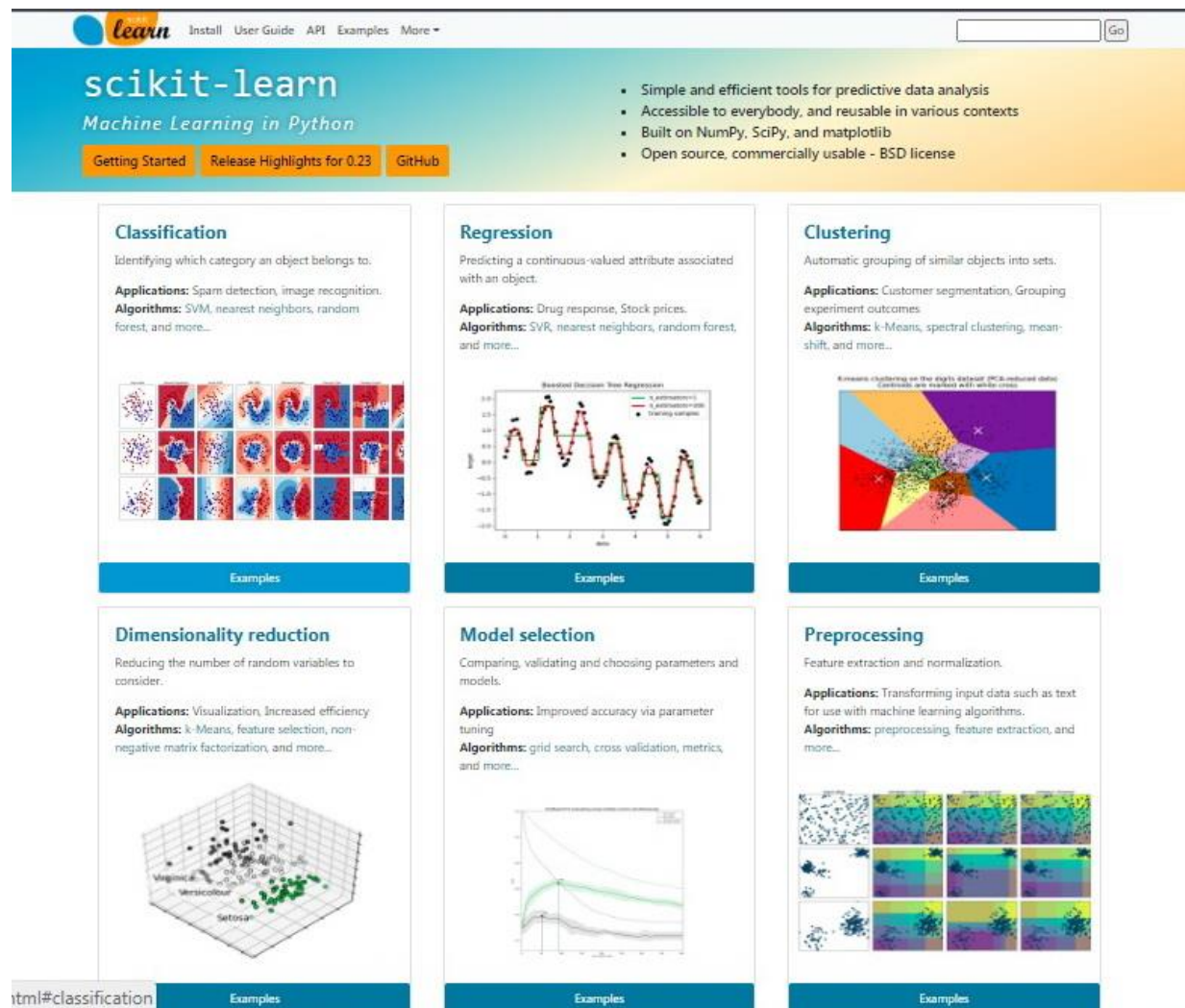


кластеризация
нет разметки: X



регрессия
есть разметка: X, y

БИБЛИОТЕКА SCIKIT-LEARN



scikit-learn
Machine Learning in Python

Getting Started | Release Highlights for 0.23 | GitHub

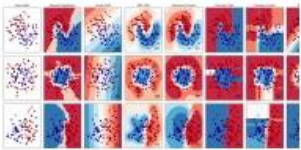
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: SVM, nearest neighbors, random forest, and more...



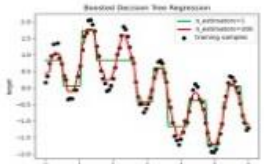
Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, nearest neighbors, random forest, and more...




Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes.

Algorithms: k-Means, spectral clustering, mean-shift, and more...



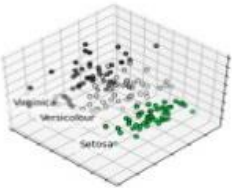
Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency.

Algorithms: k-Means, feature selection, non-negative matrix factorization, and more...



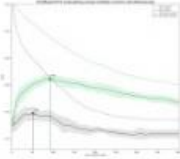
Examples

Model selection

Comparing, validating and choosing parameters and models.

Applications: Improved accuracy via parameter tuning.

Algorithms: grid search, cross validation, metrics, and more...



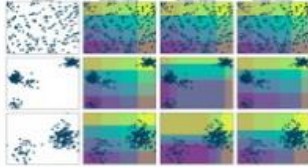
Examples

Preprocessing

Feature extraction and normalization.

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

Algorithms: preprocessing, feature extraction, and more...



Examples

Набор логически разделённых модулей

Единообразный API взаимодействия
fit + transform + predict

Отличная документация с примерами и описанием работы алгоритмов

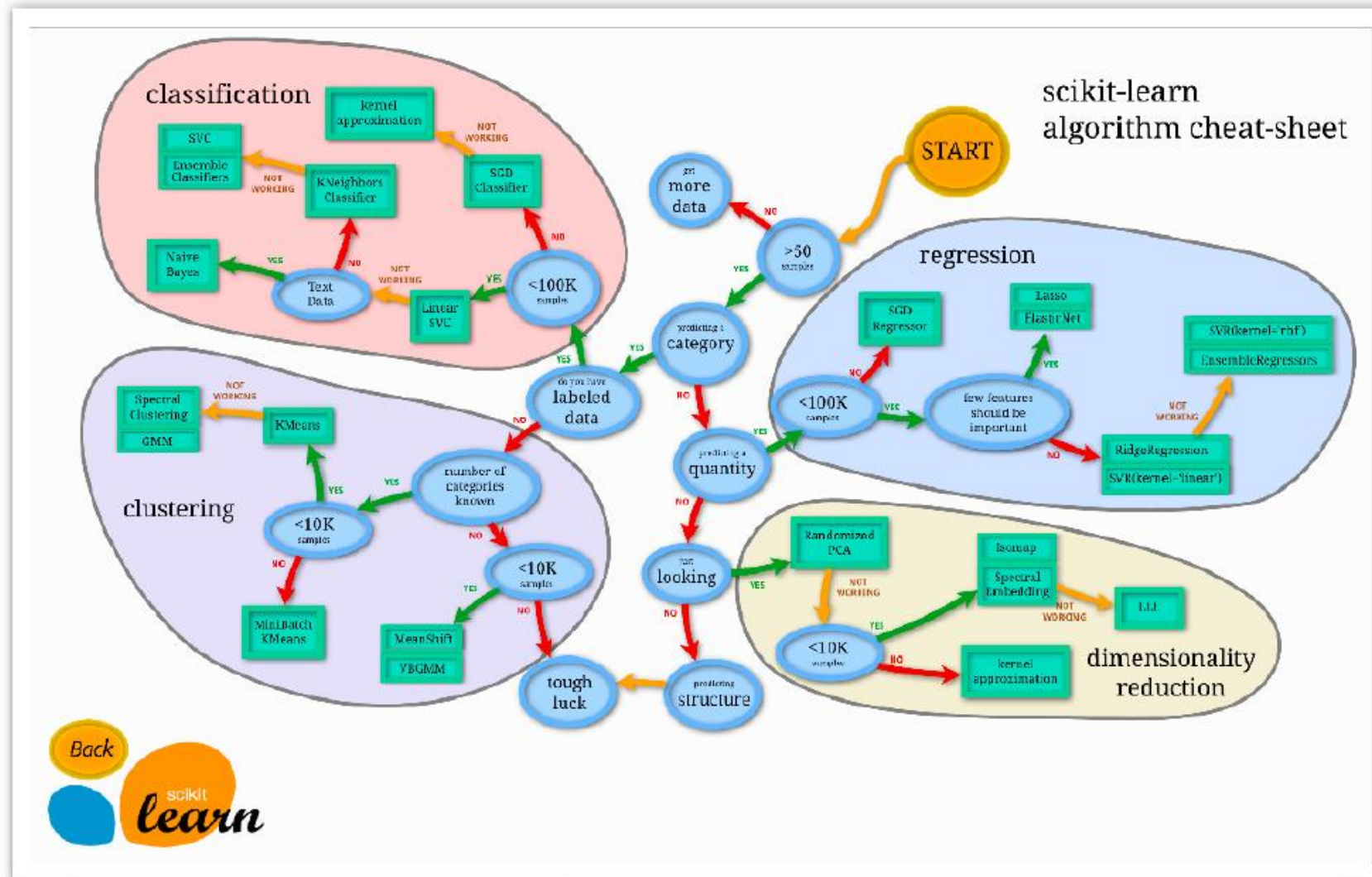
БИБЛИОТЕКА SCIKIT-LEARN

ПОЛЕЗНО ПОМНИТЬ

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

МОДУЛИ SCIKIT-LEARN

SKLEARN ALGO CHEATSHEET



МОДУЛИ SCIKIT-LEARN

ПРАКТИКА 1

1. Загрузить данные по недвижимости Бостона
2. Разделить их на 2 части: обучающую и тестовую выборки

МОДУЛИ SCIKIT-LEARN

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

linear_model - линейные модели

- LinearRegression
- LogisticRegression

МОДУЛИ SCIKIT-LEARN

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

tree - дерево решений

- DecisionTreeClassifier
- DecisionTreeRegressor

ensemble - ансамбли решений: лес, бустинг

- RandomForestClassifier
- GradientBoostingClassifier

МОДУЛИ SCIKIT-LEARN

МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

cluster - различные методы кластеризации

- KMeans, MiniBatchKMeans
- DBSCAN
- AffinityPropagation

МОДУЛИ SCIKIT-LEARN

ИСПОЛЬЗОВАНИЕ МЕТОДОВ ML

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

1. `model = LinearRegression()`

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

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

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

оценка `a` должна приближаться к `y`

МОДУЛИ SCIKIT-LEARN

ПРАКТИКА 1

3. Сделать предсказание по тестовой выборке
4. Сравнить реальные значения с предсказанием

МОДУЛИ SCIKIT-LEARN

ПРЕДОБРАБОТКА ДАННЫХ

preprocessing - нормировка

- StandardScaler

feature_extraction - векторизация

- HashingVectorizer
- TfidfVectorizer

МОДУЛИ SCIKIT-LEARN

ПОДБОР ПАРАМЕТРОВ МОДЕЛИ

model_selection - оценка качества + подбор гиперпараметров моделей

- GridSearchCV
- cross_val_score

МОДУЛИ SCIKIT-LEARN

СНИЖЕНИЕ РАЗМЕРНОСТИ

decomposition - разложение матриц и снижение размерности

- PCA
- TruncatedSVD

МОДУЛИ SCIKIT-LEARN

ОЦЕНКА КАЧЕСТВА

metrics - различные метрики качества решений

- `classification_report`
- `mean_squared_error`

МОДУЛИ SCIKIT-LEARN

ПРАКТИКА 2

1. Взять данные «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

ВЫВОД

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

БАЗОВЫЕ АЛГОРИТМЫ МАШИННОГО ОБУЧЕНИЯ В SCIKIT-LEARN

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ