

ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ

О ПРЕДМЕТЕ И
БЛИЖАЙШИХ СОСЕДЯХ

Москва, 2024

Материалы занятия <https://github.com/kshilin/machine-learning>

Что такое машинное обучение?



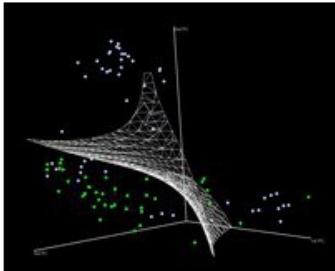
Как люди видят
мою работу

$$\begin{aligned} L_p &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^n \alpha_i \\ \alpha_i &\geq 0, \forall i \\ \mathbf{w} &= \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i, \sum_{i=1}^n \alpha_i = 0 \\ \nabla g(\theta_t) &= \frac{1}{n} \sum_{i=1}^n \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t). \\ \theta_{t+1} &= \theta_t - \eta_t \nabla \ell(x_{i(t)}, y_{i(t)}; \theta_t) - \eta_t \cdot \nabla r(\theta_t). \\ \mathbb{E}_{i(t)}[\ell(x_{i(t)}, y_{i(t)}; \theta_t)] &= \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t). \end{aligned}$$

Как программисты
видят мою работу



Как мои друзья
видят мою работу



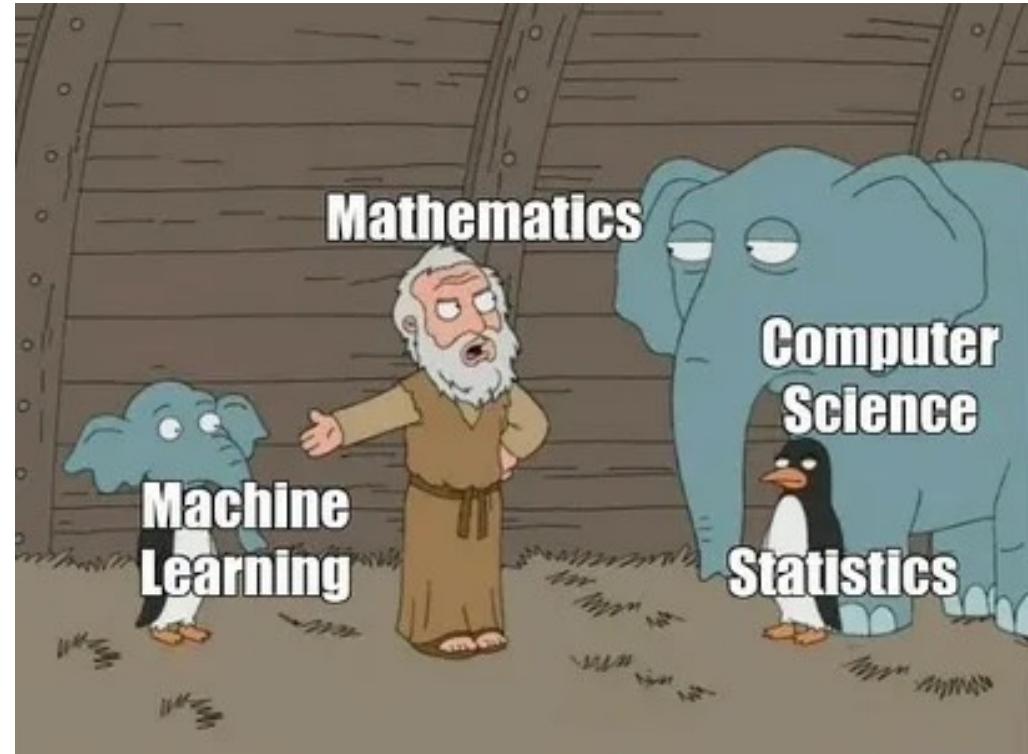
Как я вижу
свою работу



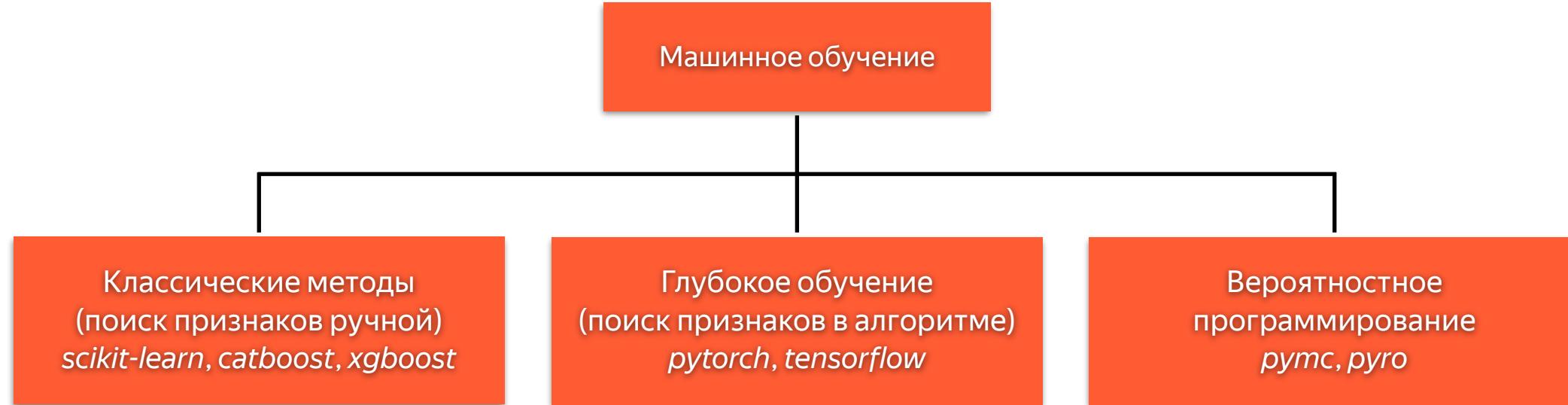
Как мои родители
видят мою работу

```
>>> from sklearn import svm
```

Что я на самом
деле делаю



Что такое машинное обучение?



Какие пререквизиты?

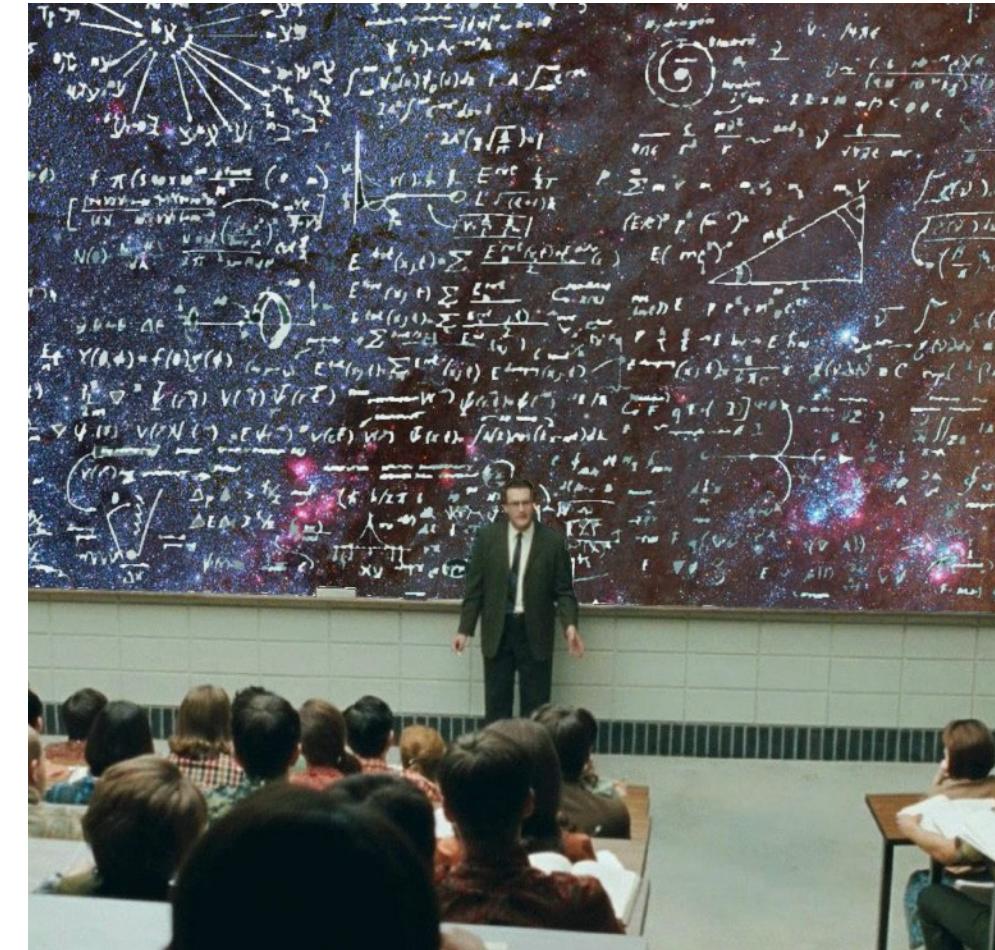
НЕОБХОДИМО

- Математический анализ
минимум 2 семестра
- Алгебра
минимум 1 семестра
- Теория вероятностей и
математическая статистика
минимум 1 семестр

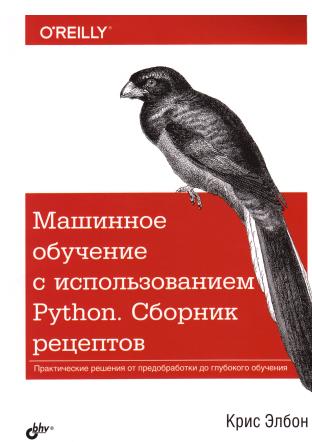
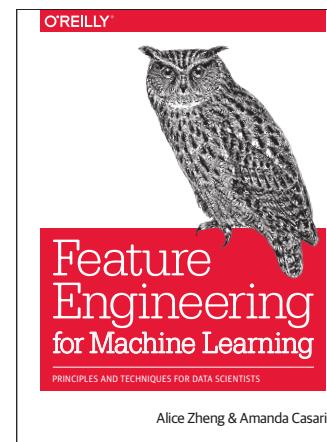
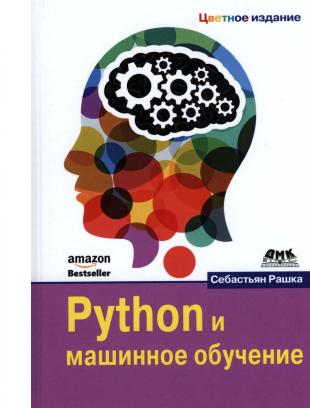
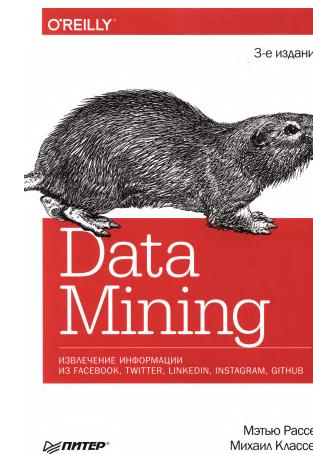
ЖЕЛАТЕЛЬНО

- Методы оптимизации
1 семестр
- Анализ данных (на Python)
1 семестр
- Программирование (на Python или R)
2 семестра

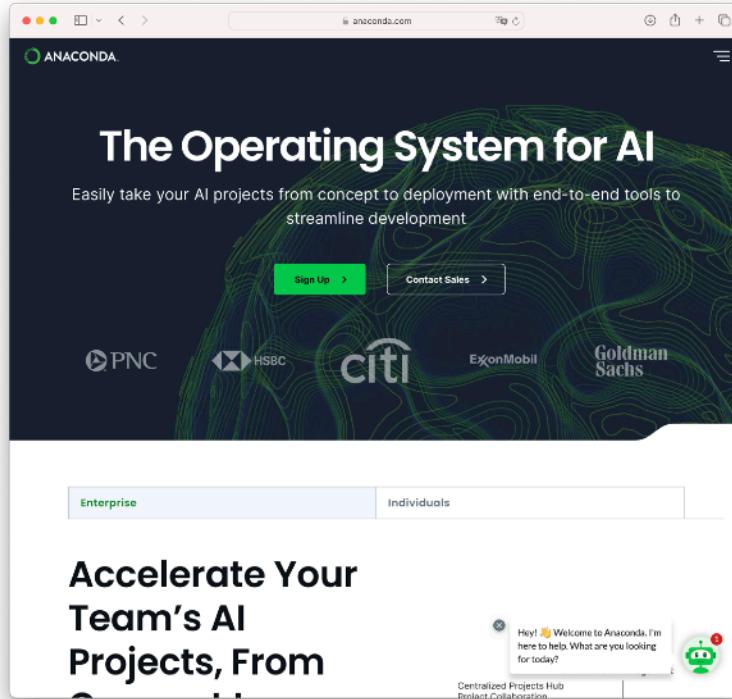
При минимуме входных знаний учится можно, но очень сложно



Что читать в процессе обучения?



В какой программной среде мы будем работать?



The Operating System for AI

Easily take your AI projects from concept to deployment with end-to-end tools to streamline development

[Sign Up >](#) [Contact Sales >](#)

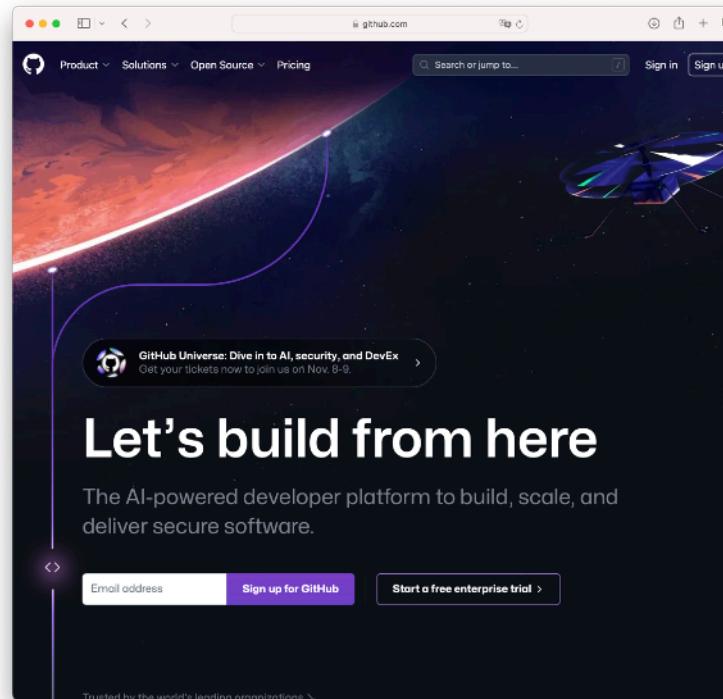
PNC HSBC Citi ExxonMobil Goldman Sachs

Enterprise Individuals

Accelerate Your Team's AI Projects, From

Hey! Welcome to Anaconda. I'm here to help. What are you looking for today?

Centralized Projects Hub, Project Collaboration



Product Solutions Open Source Pricing

Search or jump to... Sign in Sign up

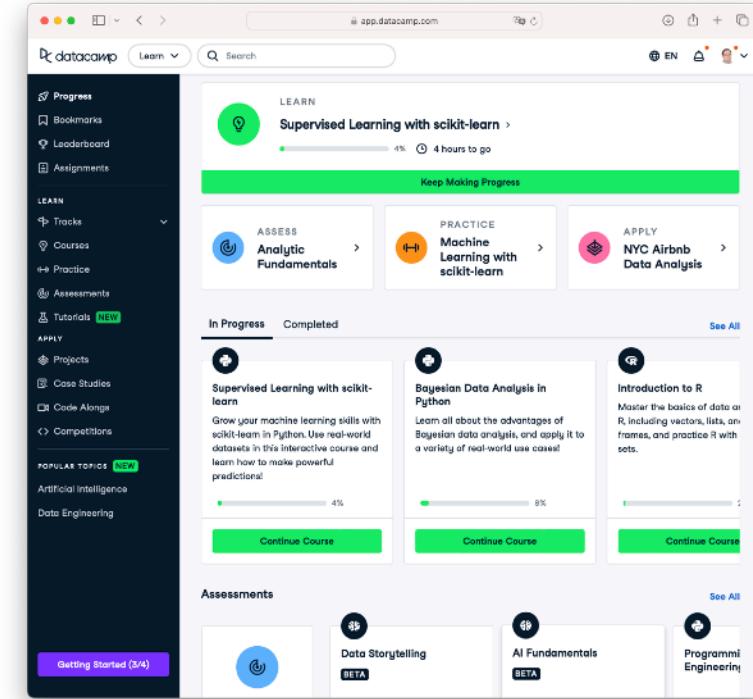
GitHub Universe: Dive in to AI, security, and DevEx Get your tickets now to join us on Nov. 8-9.

Let's build from here

The AI-powered developer platform to build, scale, and deliver secure software.

Email address [Sign up for GitHub](#) [Start a free enterprise trial >](#)

Trusted by the world's leading organizations



Progress Bookmarks Leaderboard Assignments

LEARN Tracks Courses Practice Assessments Tutorials NEW

APPLY Projects Case Studies Code Alongs Competitors

POPULAR TOPICS NOW Artificial Intelligence Data Engineering

Supervised Learning with scikit-learn 4% 4 hours to go

Keep Making Progress

ASSESS Analytic Fundamentals

PRACTICE Machine Learning with scikit-learn

APPLY NYC Airbnb Data Analysis

In Progress Completed

Supervised Learning with scikit-learn 4%

Bayesian Data Analysis In Python 8%

Introduction to R 1%

Getting Started (3/4)

Continue Course Continue Course Continue Course

Assessments

Data Storytelling BETA AI Fundamentals BETA Programming Engineering

Какие алгоритмы изучим?

Обучение с учителем

Классификация и регрессия

- Ближайшие соседи
- Линейные модели
- Наивные байесовские модели
- Машины опорных векторов
- Деревья решений
- Случайные леса
- Градиентный бустинг на деревьях решений

Обучение без учителя

Кластеризация

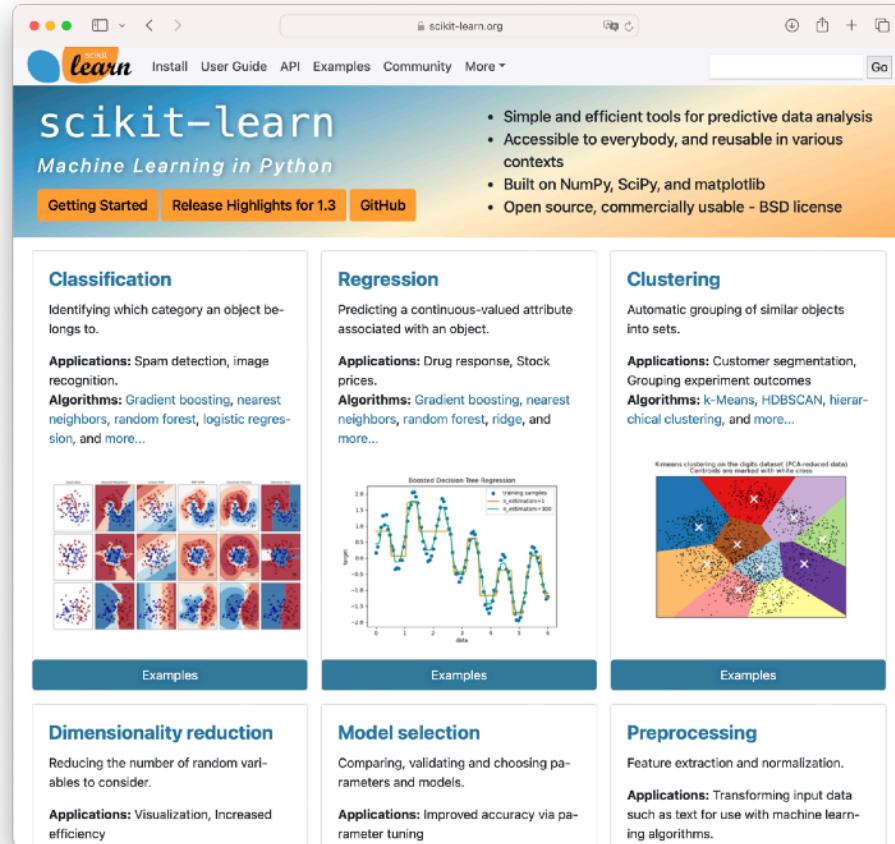
- Преобразование данных:
 - стандартизация
 - нормализация
 - обработка категориальных данных
 - анализ главных компонент
 - t-SNE
- Кластеризация:
 - k-средних
 - агломеративная
 - DBSCAN
- Поиск аномалий в данных

Дополнительно

Контейнеры

- Конструирование пайплайнов
- Отбор наиболее значимых признаков
- Трактовка значимости с использованием векторов Шепли
- Собственные метрики качества
- Регрессия и кластеризация временных рядов

С каким фреймворком мы научимся работать?



The screenshot shows the official website for scikit-learn (scikit-learn.org). The main navigation bar includes links for 'Install', 'User Guide', 'API', 'Examples', 'Community', and 'More'. Below the navigation, there's a 'Getting Started' section, 'Release Highlights for 1.3', and a 'GitHub' link.

Classification: Identifying which category an object belongs to. Applications: Spam detection, image recognition. Algorithms: Gradient boosting, nearest neighbors, random forest, logistic regression, and more... Examples: Boosted Decision Tree Regression plot and a scatter plot of digits.

Regression: Predicting a continuous-valued attribute associated with an object. Applications: Drug response, Stock prices. Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more... Examples: Boosted Decision Tree Regression plot and a scatter plot of digits.

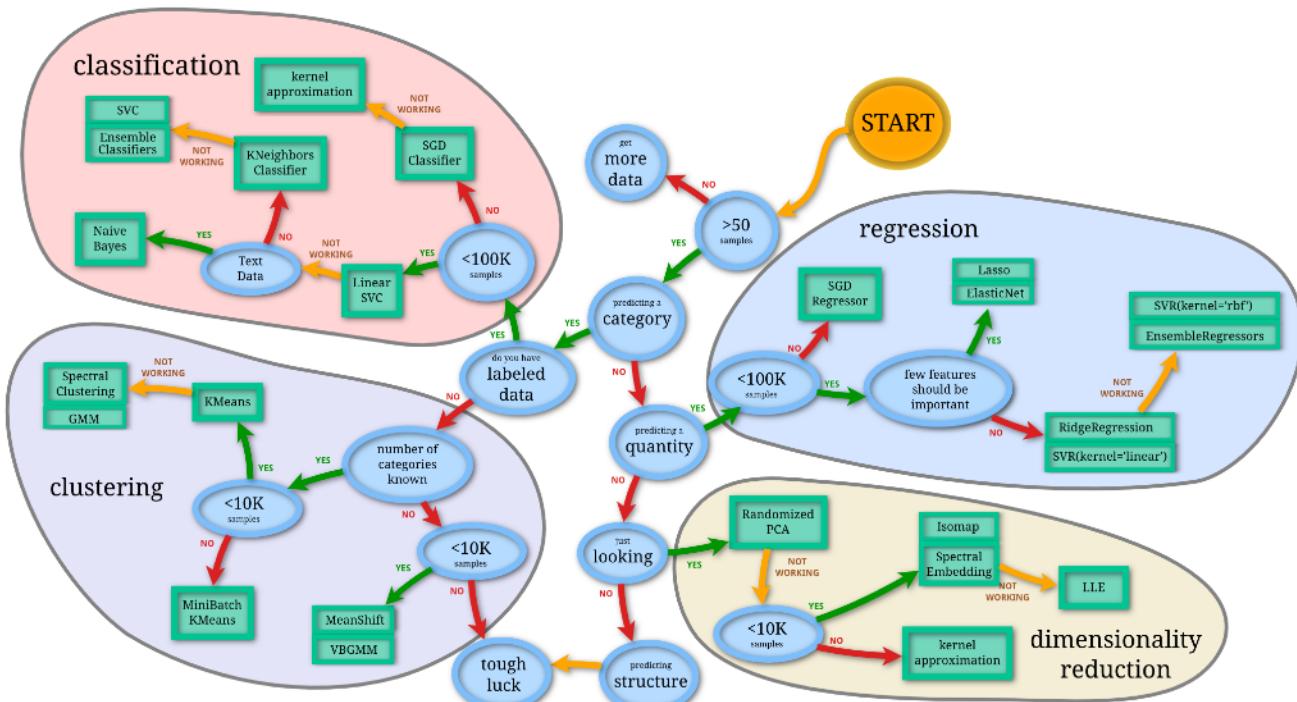
Clustering: Automatic grouping of similar objects into sets. Applications: Customer segmentation, Grouping experiment outcomes. Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more... Examples: A scatter plot of digits.

Dimensionality reduction: Reducing the number of random variables to consider. Applications: Visualization, Increased efficiency.

Model selection: Comparing, validating and choosing parameters and models. Applications: Improved accuracy via parameter tuning.

Preprocessing: Feature extraction and normalization. Applications: Transforming input data such as text for use with machine learning algorithms.

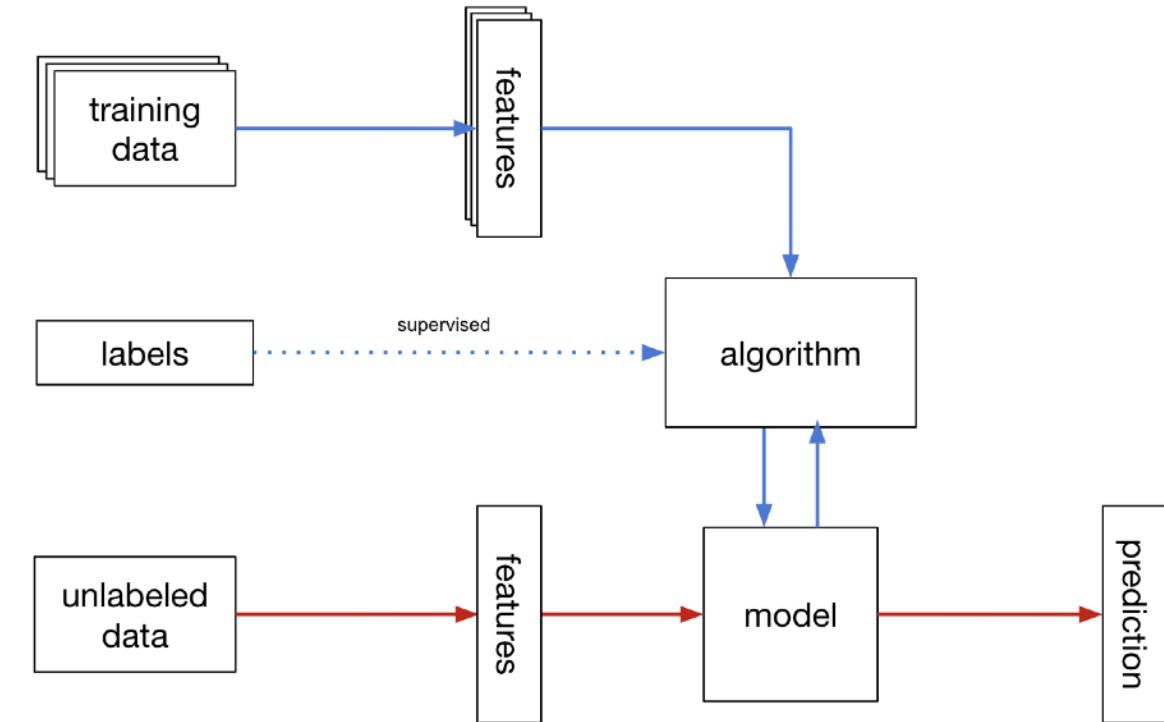
A sidebar on the right lists several tools: SVC, Ensemble Classifiers, kernel approximation, SGD Classifier, Naive Bayes, KNeighbors Classifier, NOT WORKING, <100K samples, Linear SVC, Text Data, and NOT WORKING.



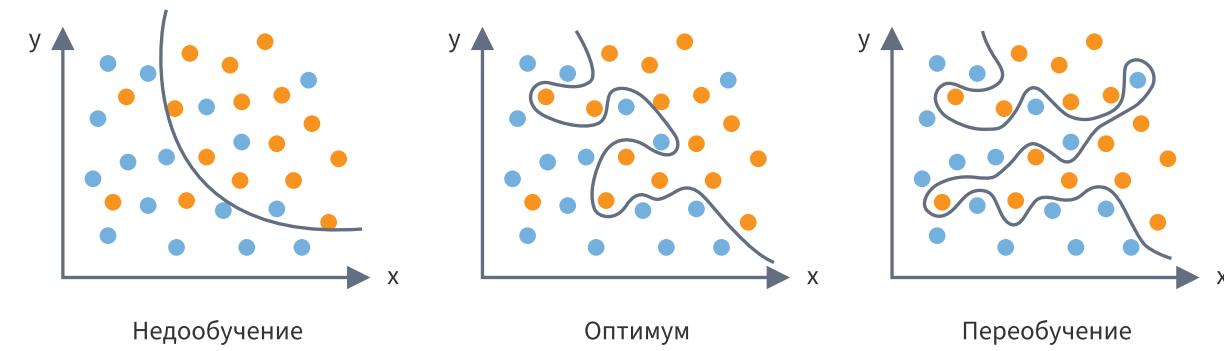
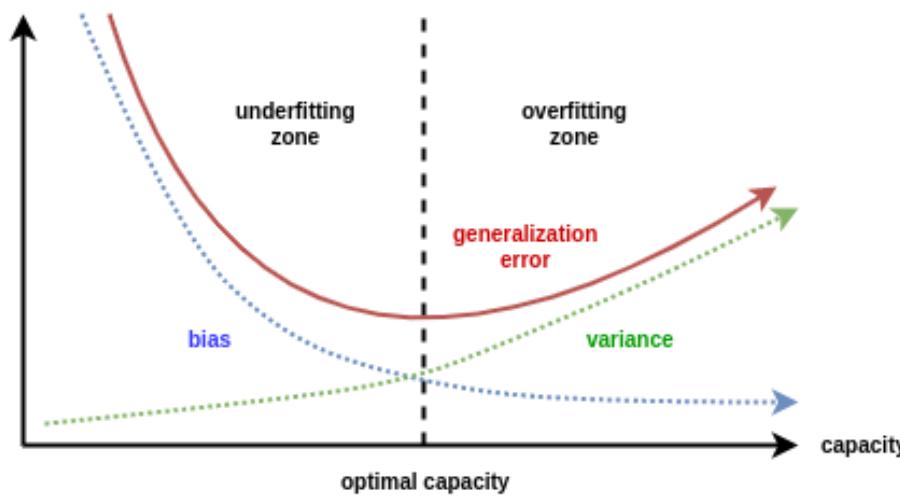
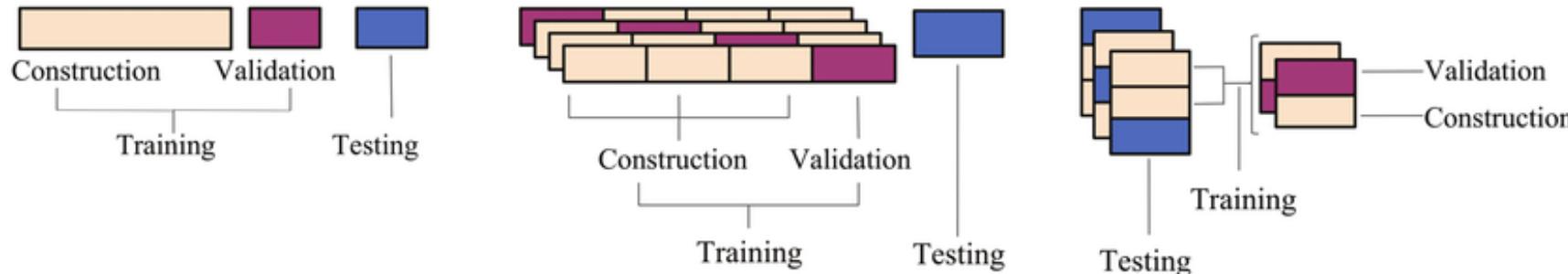
Как строится разработка модели ML

Обучение с учителем:

- Этап 1 Подготовка данных (препроцессинг)
- Этап 2 Подбор алгоритма машинного обучения
- Этап 3 Оценка качества модели и ее улучшение



Как проверить качество алгоритма





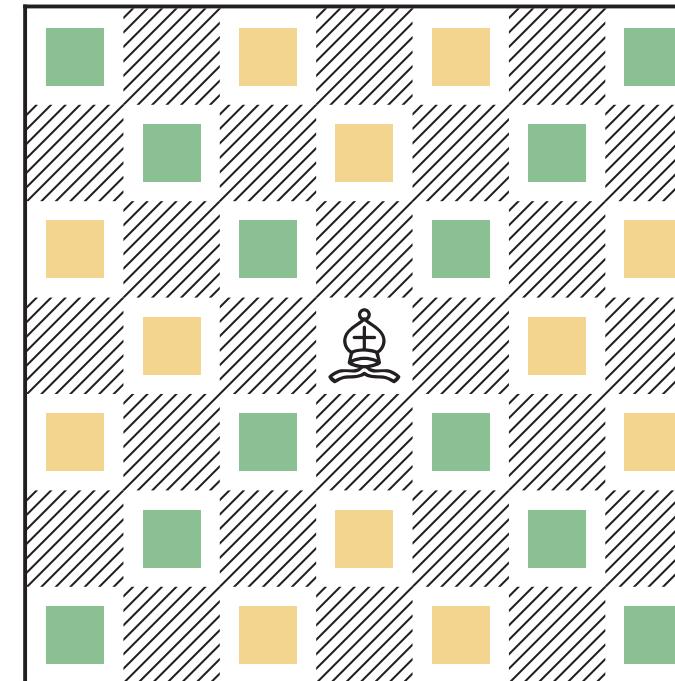
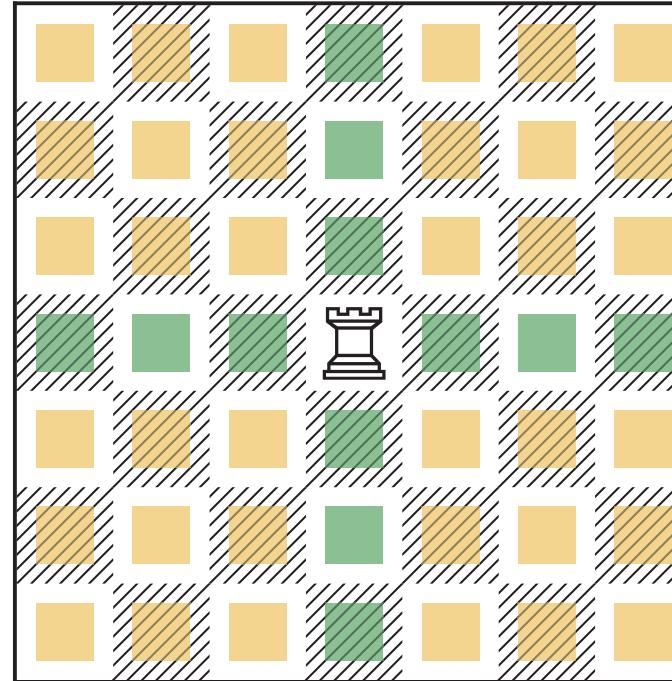
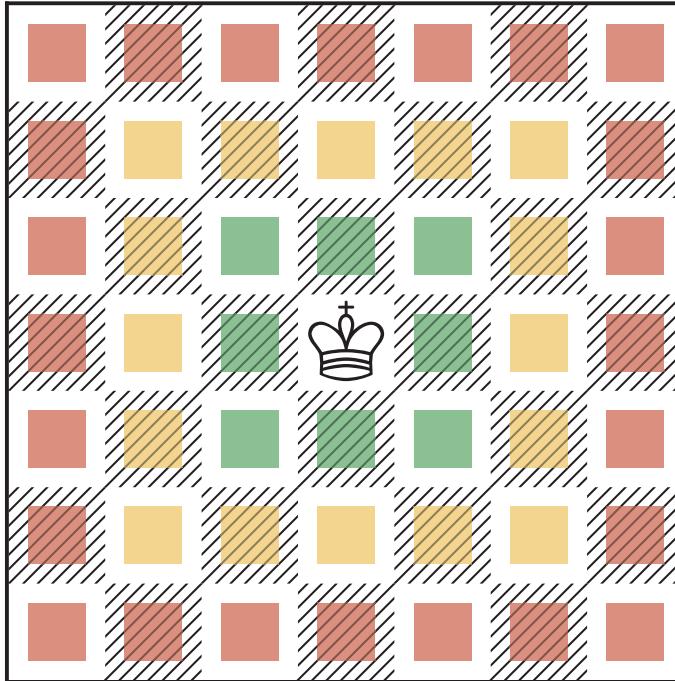
Институт экономики,
математики
и информационных
технологий



ПРЕЗИДЕНТСКАЯ
АКАДЕМИЯ

Пятиминутка математики ...

Шахматы



Метрическое пространство

Метрическое пространство есть пара (X, d) , где X – множество, а d – числовая функция, которая определена на декартовом произведении $X \times X$, принимает значения в множестве неотрицательных вещественных чисел, и такова, что

1. $d(x, y) = 0 \Leftrightarrow x = y$ (аксиома тождества).
2. $d(x, y) = d(y, x)$ (аксиома симметрии).
3. $d(x, z) \leq d(x, y) + d(y, z)$ (аксиома треугольника или неравенство треугольника).

При этом

- множество X называется подлежащим множеством метрического пространства.
- элементы множества X называются точками метрического пространства.
- функция d называется метрикой.

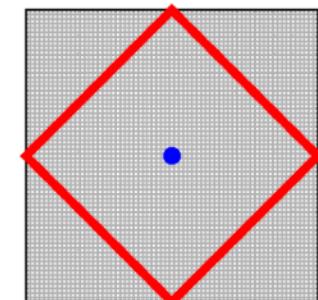
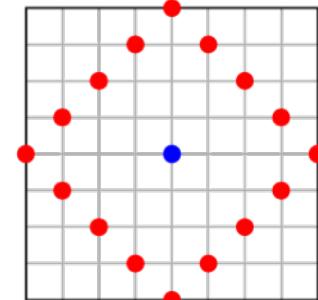
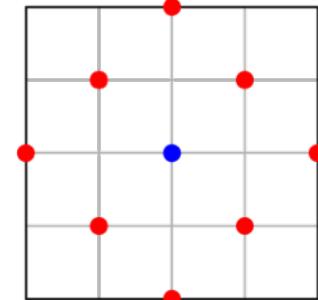
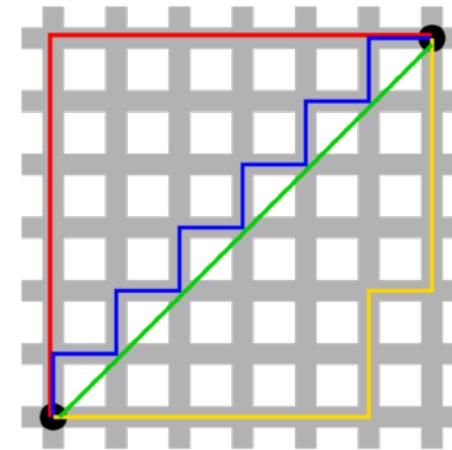
Расстояние Минковского порядка p между двумя точками определяется как:

$$\rho(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}.$$

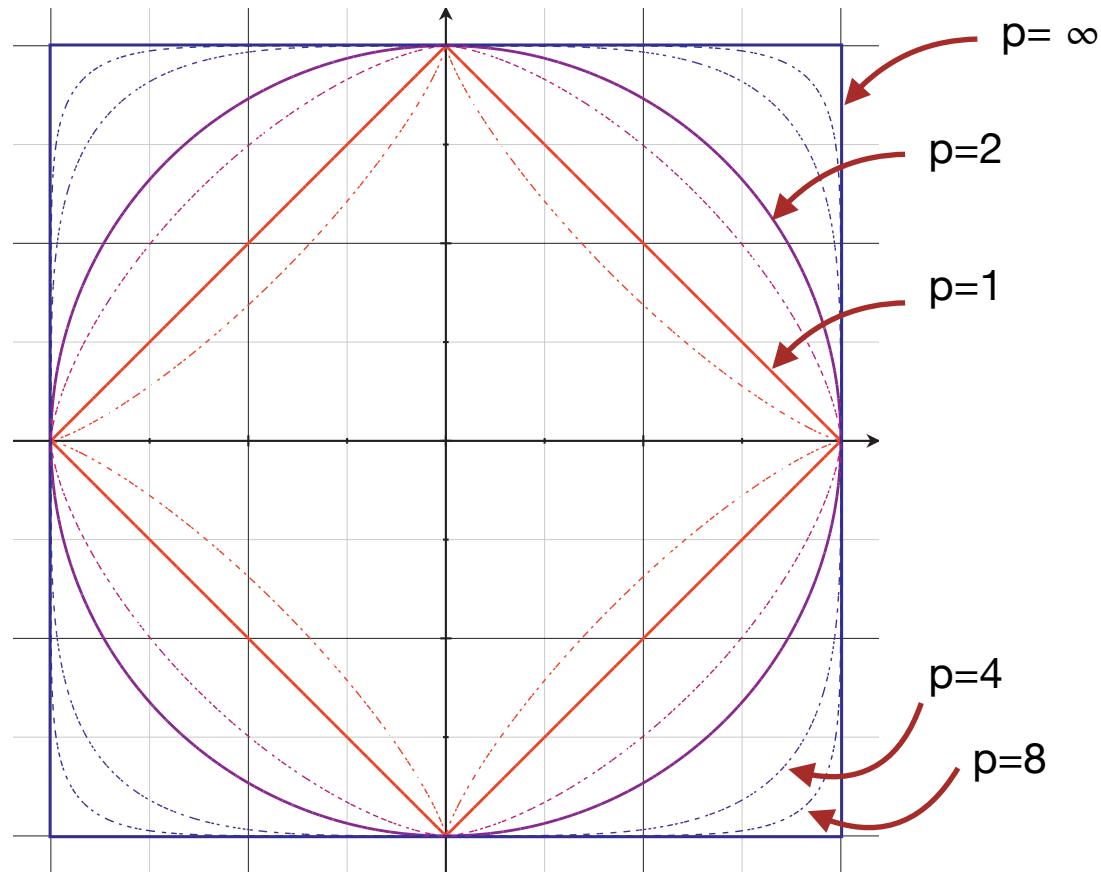
Для $p \geq 1$ расстояние Минковского является метрикой вследствие неравенства Минковского.

Для $p < 1$ расстояние не является метрикой, поскольку нарушается неравенство треугольника.

При $p = \infty$ метрика обращается в расстояние Чебышёва.

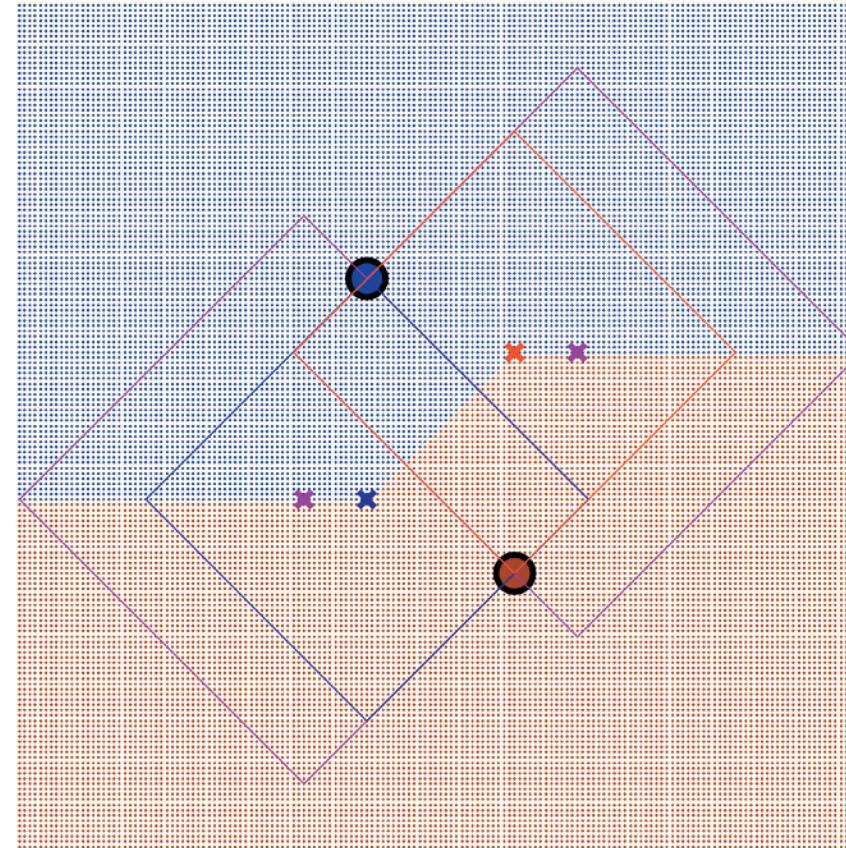
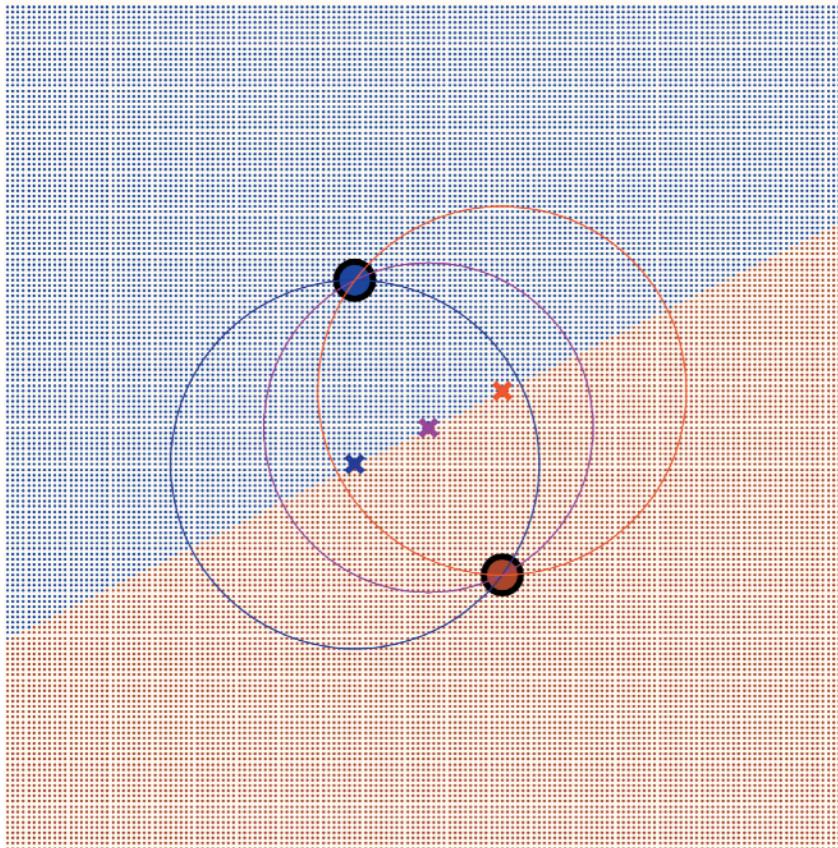


Расстояние Миньковского

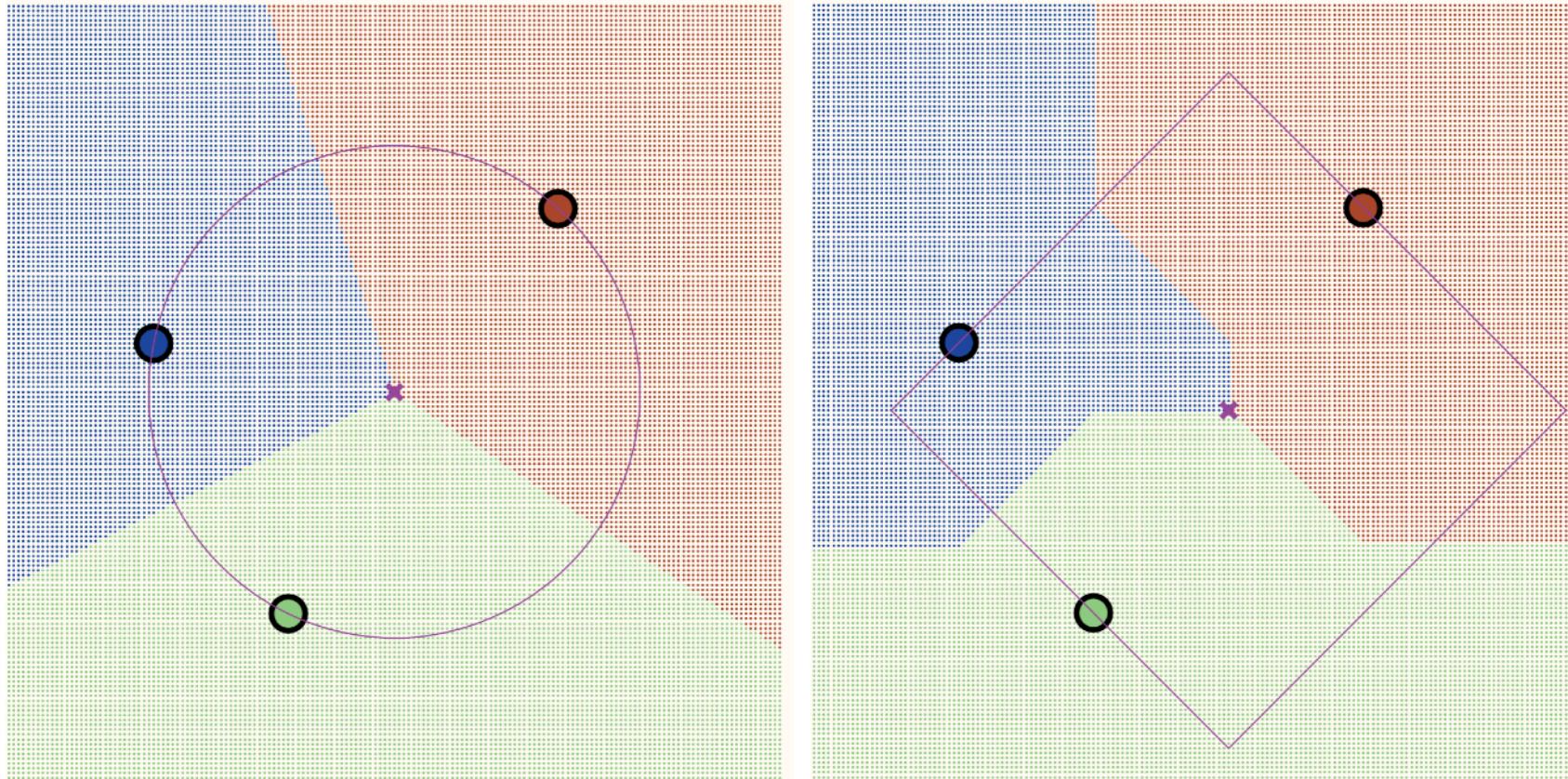


$$\rho(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}.$$

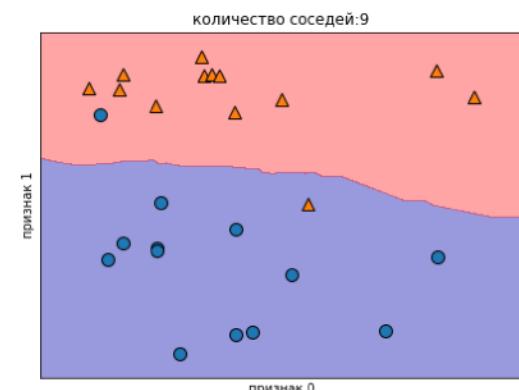
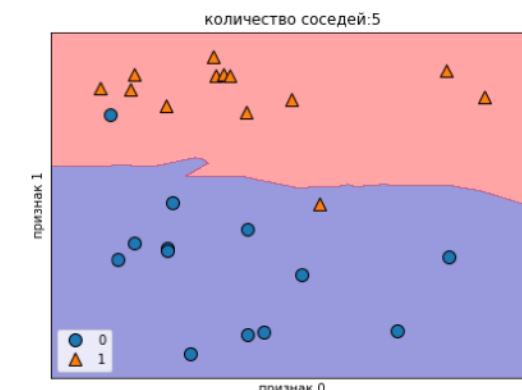
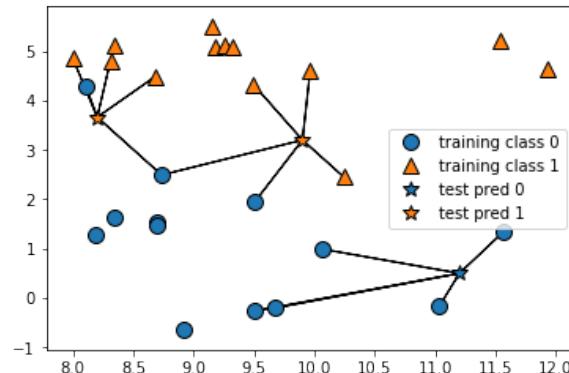
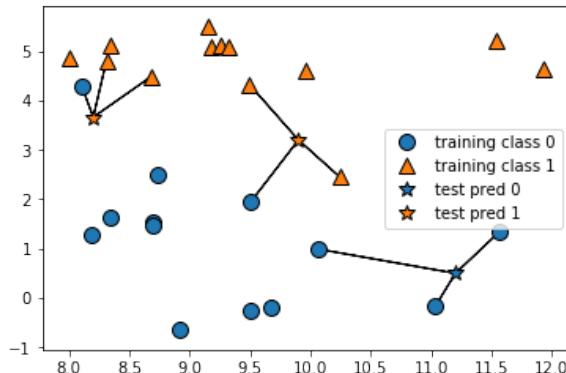
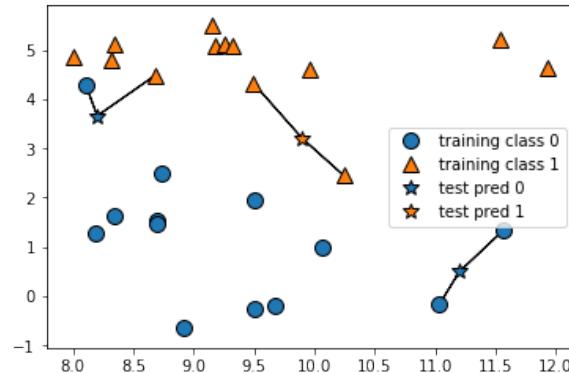
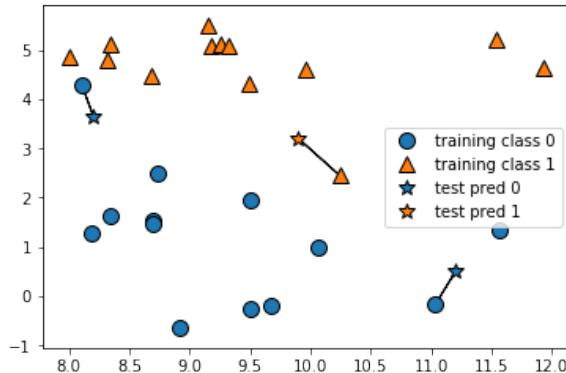
Решающая граница (2 класса)



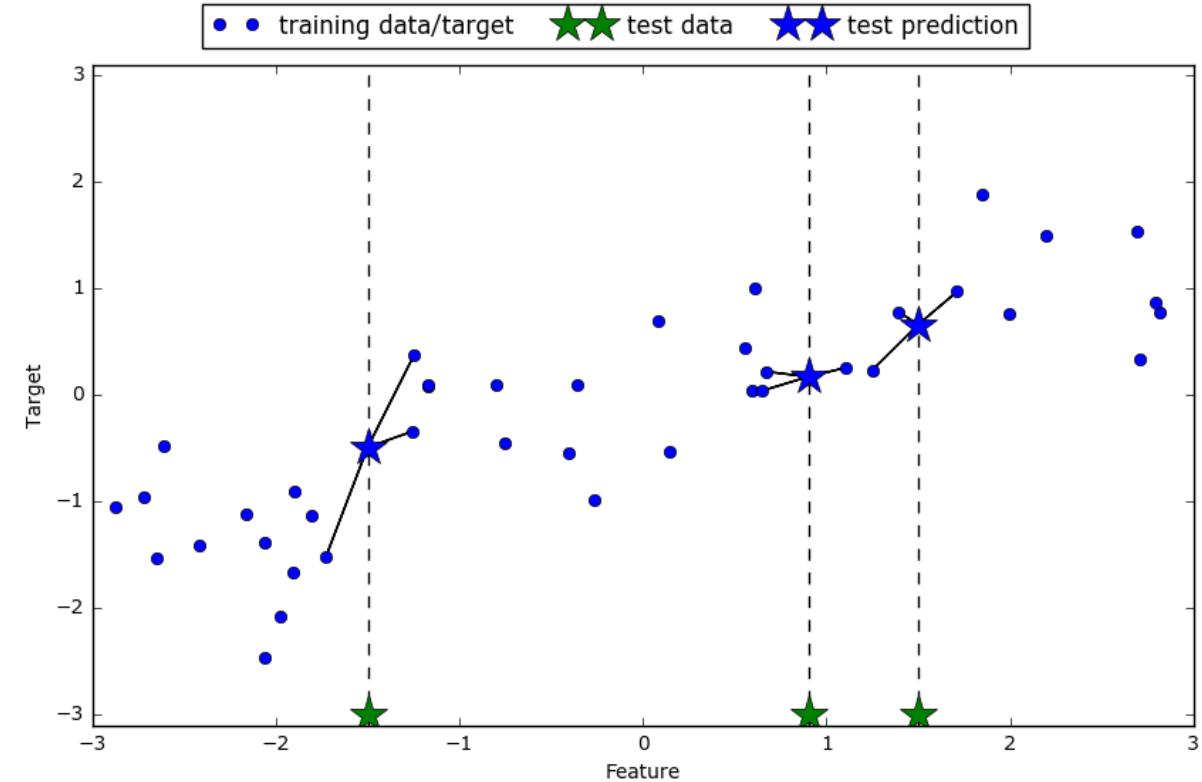
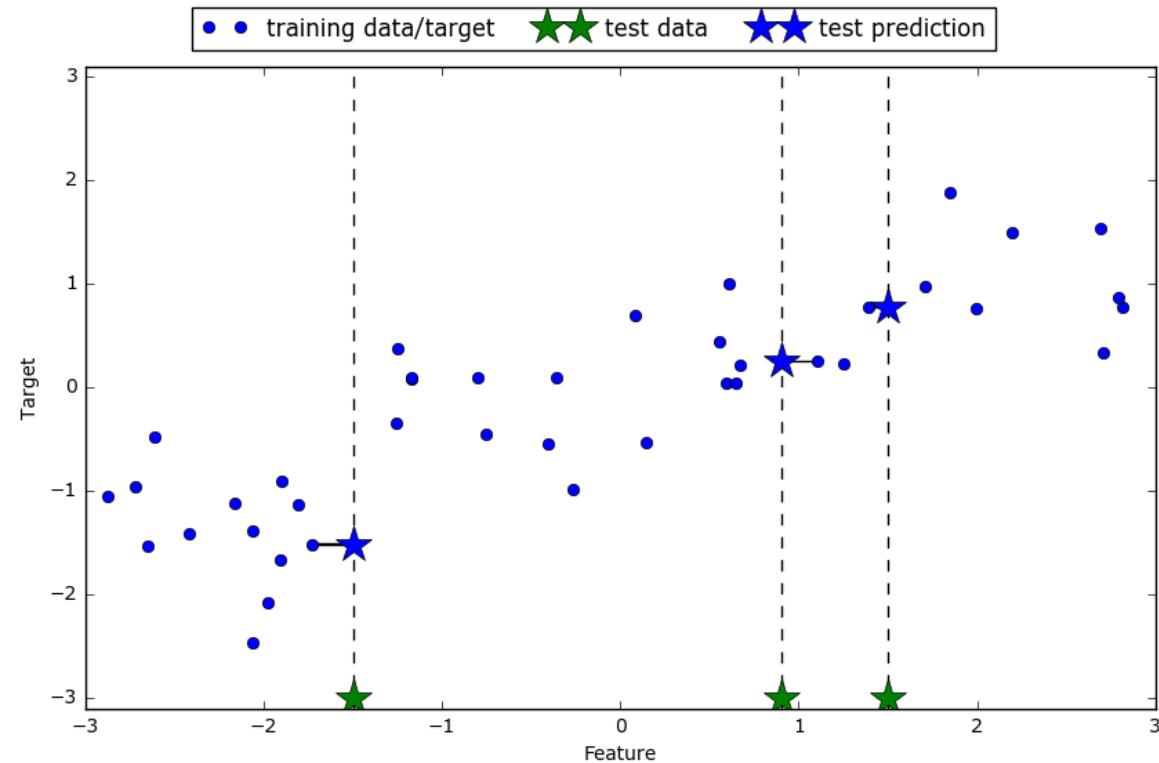
Решающая граница (3 класса)



Классификация к ближайших соседей



Регрессия к ближайших соседей



Ирисы Фишера

Набор данных «Ирисы Фишера» состоит из данных о 150 экземплярах ириса, по 50 экземпляров из трёх видов:

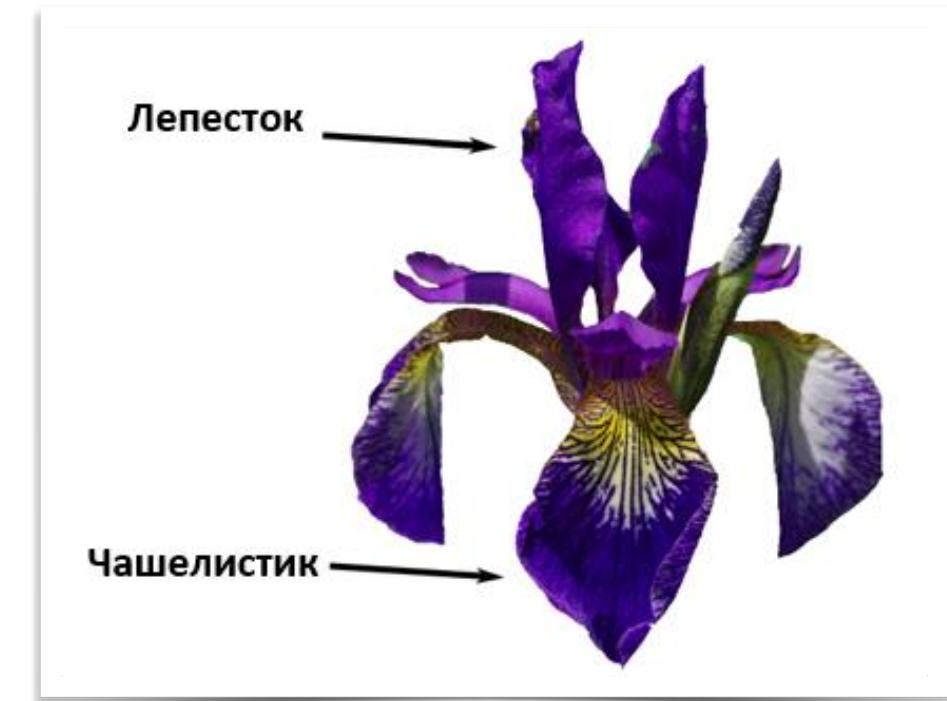
- Ирис щетинистый (*Iris setosa*),
- Ирис виргинский (*Iris virginica*)
- Ирис разноцветный (*Iris versicolor*).

Для каждого экземпляра измерялись четыре характеристики (в сантиметрах):

- Длина лепестка (англ. *sepal length*);
- Ширина лепестка (англ. *sepal width*);
- Длина чашелистика (англ. *petal length*);
- Ширина чашелистика (англ. *petal width*).

На основании этого набора данных требуется построить правило классификации, определяющее вид растения по данным измерений.

Это задача многоклассовой классификации, так как имеется три класса — три вида ириса.





ПРЕЗИДЕНТСКАЯ
АКАДЕМИЯ

СПАСИБО ЗА ВНИМАНИЕ!

Москва

РАНХиГС

2024