



ИТ-ОБРАЗОВАНИЕ
В ПЕТЕРБУРГЕ
И УДАЛЕННО

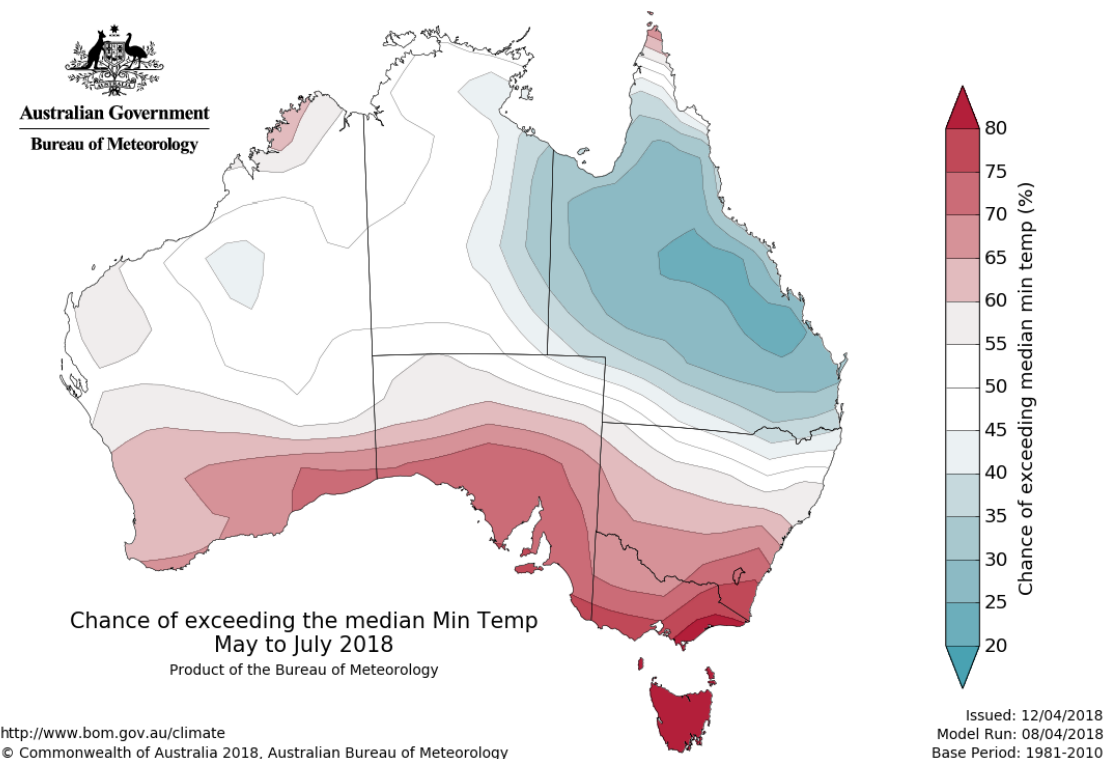
Введение в машинное обучение

Преподаватель: Зубоченко Антон

Машинное обучение

Машинное обучение — класс методов искусственного интеллекта, решающих задачу, строя алгоритм на основе размеченных данных

Применения машинного обучения



Применения машинного обучения

Специально для Вас

Аффиная, rsac, номер скрыт и другие

▶ СЛУШАТЬ ВСЕ



Приглядитесь к этим предложениям



3 295 ₽ -50 %

6 590 ₽

Кеды VANS



7 030 ₽ -35 %

10 800 ₽

Внешняя звуковая



1 875 000 ₽

Виниловый
проигрыватель Spira...



4 400 ₽ -30 %

6 290 ₽

Кеды VANS



11 790 ₽

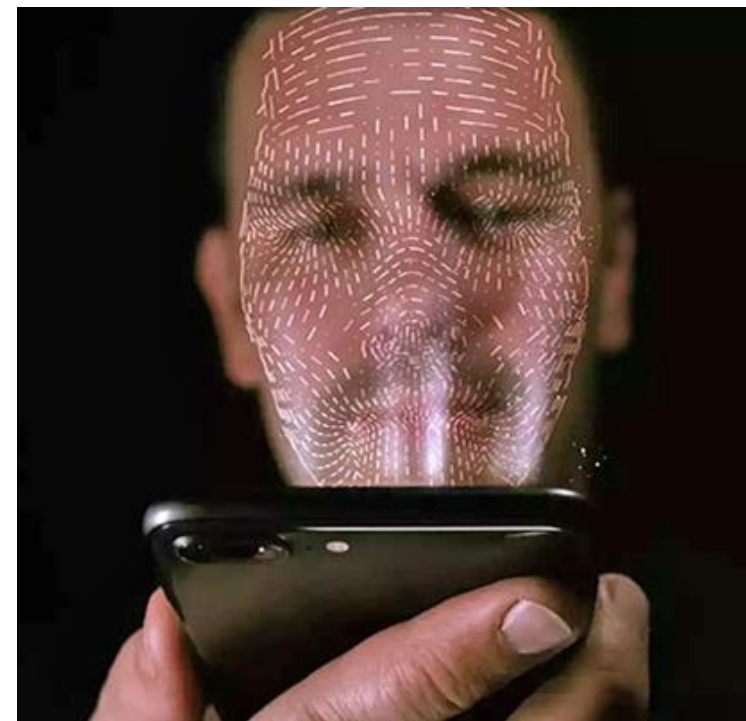
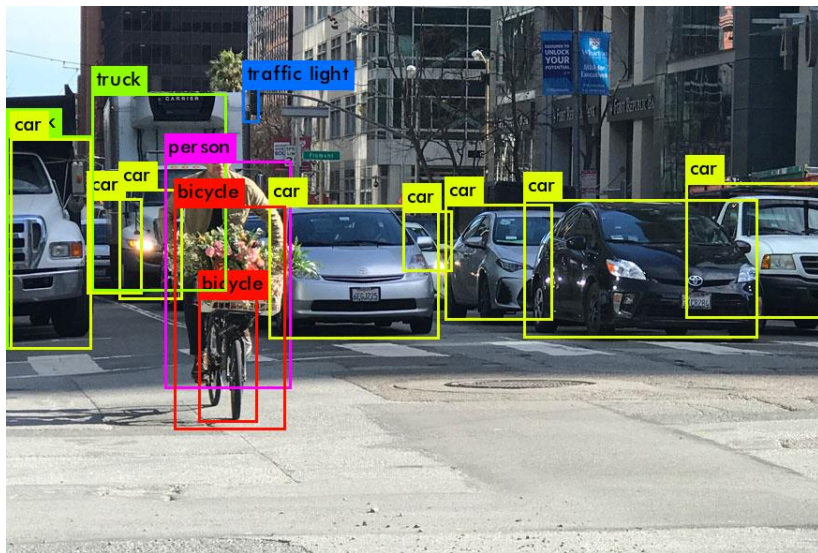
Лонгборд GoldCoast
Standard



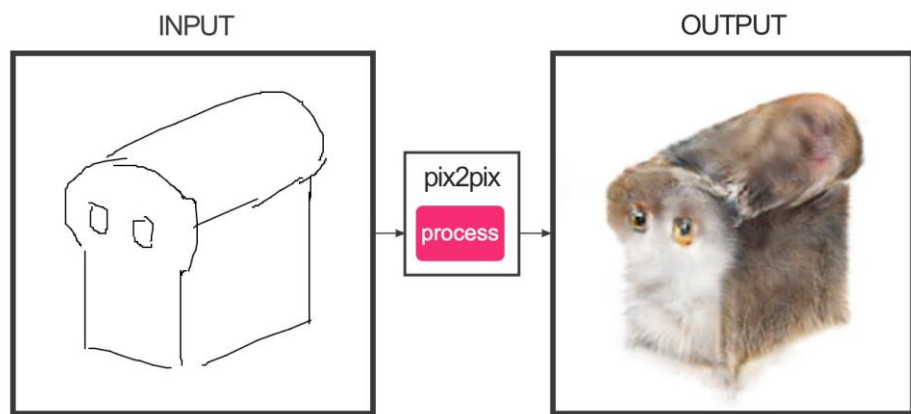
Сначала интересные



Применения машинного обучения

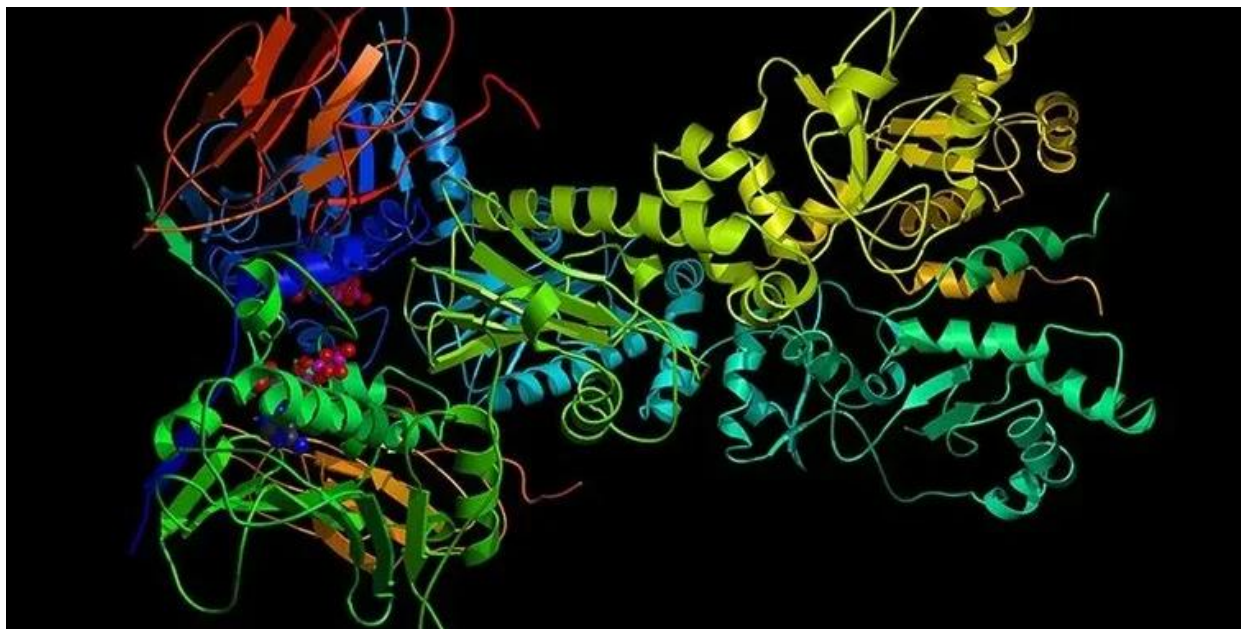


Применения машинного обучения



AlphaFold

- Нейросеть AlphaFold предсказывает структуру молекулы
- Это открывает путь к изобретению инновационных лекарств




Основные понятия и задачи машинного обучения

План

- Постановка задачи машинного обучения
 - обучение с учителем
- Табличные данные и типы признаков
- Классификация и регрессия

Постановка задачи машинного обучения

Пример: оцените стоимость ноутбука

	Кол-во ядер	RAM (ГБ)	Объем жесткого диска (ГБ)	Диагональ/ разрешение	Работа от аккумулятора	Цена (руб.)
1 	4	16	1000 (HDD)+ 128 (SSD)	17"/1920x1080 пикс.	до 6 часов	?

Пример: оцените стоимость ноутбука

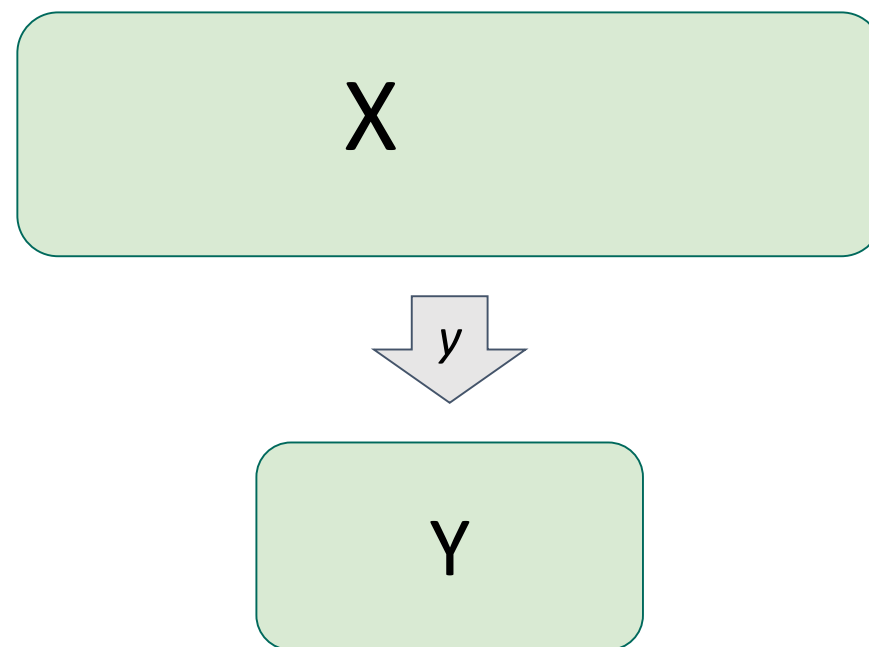
	Кол-во ядер	RAM (ГБ)	Объем жесткого диска (ГБ)	Диагональ/ разрешение	Работа от аккумулятора	Цена (руб.)
1 	4	16	1000 (HDD)+ 128 (SSD)	17"/1920x1080 пикс.	до 6 часов	?
2 	2	4	500 (HDD)	15"/1920x1080 пикс.	до 5 часов	31 490
3 	4	8	256 (SSD)	14"/1920x1080 пикс.	до 12 часов	60 990
4 	4	16	1000 (HDD)	17"/1920x1080 пикс.	до 3 часов	65 990
5 	8	16	1000 (HDD) + 256 (SSD)	17"/1920x1080 пикс.	до 11 часов	109 990

Пример: оцените стоимость ноутбука

	Кол-во ядер	RAM (ГБ)	Объем жесткого диска (ГБ)	Диагональ/ разрешение	Работа от аккумулятора	Цена (руб.)
1 	4	16	1000 (HDD)+ 128 (SSD)	17"/1920x1080 пикс.	до 6 часов	86990
2 	2	4	500 (HDD)	15"/1920x1080 пикс.	до 5 часов	31 490
3 	4	8	256 (SSD)	14"/1920x1080 пикс.	до 12 часов	60 990
4 	4	16	1000 (HDD)	17"/1920x1080 пикс.	до 3 часов	65 990
5 	8	16	1000 (HDD) + 256 (SSD)	17"/1920x1080 пикс.	до 11 часов	109 990

Постановка задачи машинного обучения

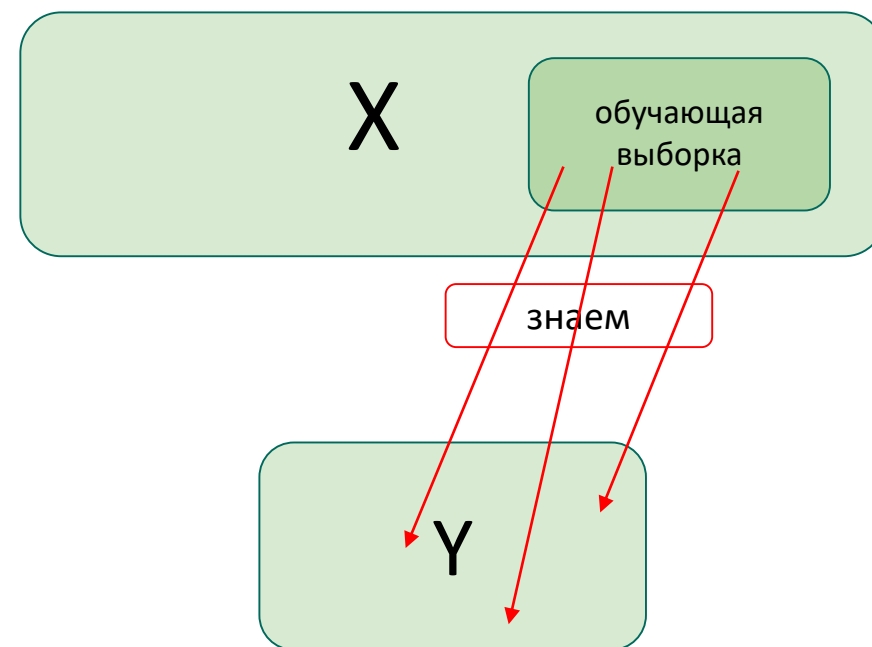
- X — множество *объектов*
- Y — множество *ответов* (например, два класса или произвольные числа)
- $y: X \rightarrow Y$ — неизвестная закономерность



Постановка задачи машинного обучения

- X — множество *объектов*
- Y — множество *ответов* (например, два класса или произвольные числа)
- $y: X \rightarrow Y$ — неизвестная закономерность

Дано: обучающая выборка, $\{x^1, x^2, \dots, x^l\}$ — подмножество множества X с известными лейблами

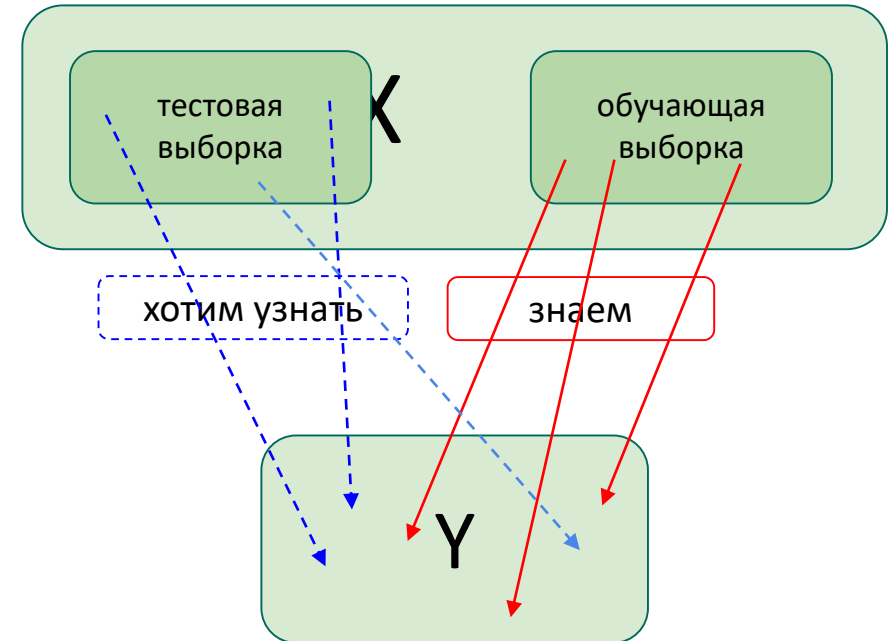


Постановка задачи машинного обучения

- X — множество *объектов*
- Y — множество *ответов* (например, два класса или произвольные числа)
- $y: X \rightarrow Y$ — неизвестная закономерность

Дано: обучающая выборка, $\{x^1, x^2, \dots, x^l\}$ — подмножество множества X с известными лейблами

Цель: подобрать алгоритм $\hat{y}: X \rightarrow Y$, приближающий функцию $y(x)$ на **тестовой выборке**



Обучение с учителем (supervised learning)

- Дан набор данных с известными ответами
- Обучаем модель на данных
- Используем модель на новых данных

Разметка — наличие ответов для имеющихся данных

Обучение с учителем — постановка задачи машинного обучения, при которой обучение происходит с использованием размеченных данных

Табличные данные и типы признаков

Какие бывают данные?

MNIST Dataset

- Изображения цифр, написанных от руки
- ~50 тысяч изображений
- Можно научить модель распознавать цифру



Как задаются объекты. Признаковое описание

Объект x задаётся *признаковым описанием*

$x = (x_1, x_2, \dots, x_k)$ — *вектор признаков* объекта x

$$\begin{pmatrix} x_1^1 & x_2^1 & \dots & x_k^1 \\ x_1^2 & x_2^2 & \dots & x_k^2 \\ \dots & \dots & \dots & \dots \\ x_1^\ell & x_2^\ell & \dots & x_k^\ell \end{pmatrix}$$

— *матрица “объекты-признаки”*
объект, пригодный для применения
алгоритмов машинного обучения

Табличные данные

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 310128
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450
6	0	3	Moran, Mr. James	male		0	0	330877
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909

- Титаник — датасет для классификации: выжил/не выжил

Терминология

- Набор данных – **выборка**
- Строка в таблице — **объект**
- Столбец в таблице — **признак**
- Признаки могут быть 3-х типов:
 - Числовые
 - Категориальные
 - Бинарные
- Столбец, который нужно предсказать — **целевая переменная**

Признаки и целевая переменная

y	features						
Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 310128
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
0	3	Allen, Mr. William Henry	male	35	0	0	373450
0	3	Moran, Mr. James	male		0	0	330877
0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463
0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909

Объект

Числовые признаки

- Обычные числа с большим количеством возможных значений
- Можно сравнивать друг с другом

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 310128
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
0	3	Allen, Mr. William Henry	male	35	0	0	373450
0	3	Moran, Mr. James	male		0	0	330877
0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463
0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909

Числовые признаки

- Обычные числа с большим количеством возможных значений
- Можно сравнивать друг с другом

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 310128
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
0	3	Allen, Mr. William Henry	male	35	0	0	373450
0	3	Moran, Mr. James	male		0	0	330877
0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463
0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909

Категориальные признаки

- Строки или натуральные числа, у которых мало возможных значений
- Даже если признак является числом, то порядок может не иметь смысла

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 310128
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
0	3	Allen, Mr. William Henry	male	35	0	0	373450
0	3	Moran, Mr. James	male		0	0	330877
0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463
0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909

Бинарные признаки

Бинарные признаки — категориальные признаки с двумя возможными значениями

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 310128
1	1	Futelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
0	3	Allen, Mr. William Henry	male	35	0	0	373450
0	3	Moran, Mr. James	male		0	0	330877
0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463
0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909

Другие признаки

Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599
1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 310128
1	1	Futelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803
0	3	Allen, Mr. William Henry	male	35	0	0	373450
0	3	Moran, Mr. James	male		0	0	330877
0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463
0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909

Виды задач машинного обучения

Классификация и регрессия

Целевая переменная, как и признаки, может быть двух типов:




- Числовая: $y \in \mathbb{R}$
- Категориальная: $y \in \{\text{class}_1, \dots, \text{class}_n\}$
- Бинарная: $y \in \{0,1\}$

Классификация и регрессия

Целевая переменная, как и признаки, может быть двух типов:

- Числовая: $y \in \mathbb{R}$ — **регрессия**
- Категориальная: $y \in \{\text{class}_1, \dots, \text{class}_n\}$ — **классификация**
- Бинарная: $y \in \{0,1\}$ — **бинарная классификация**

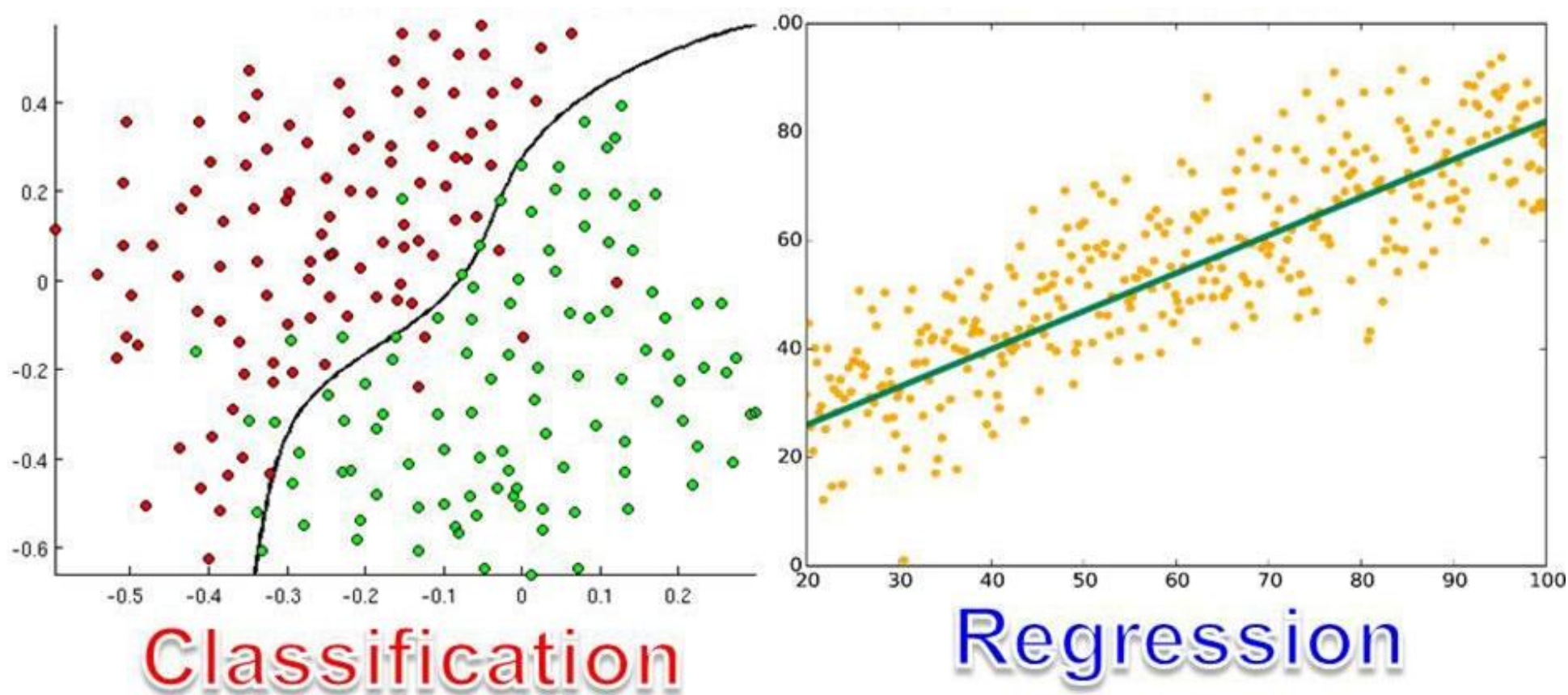
Пример: оцените стоимость ноутбука

	Кол-во ядер	RAM (Гб)	Объем жесткого диска (ГБ)	Диагональ/ разрешение	Работа от аккумулятора	Цена (руб.)
1 	2	4	500 (HDD)	15"/1920x1080 пикс.	до 5 часов	31 490
2 	4	8	256 (SSD)	14"/1920x1080 пикс.	до 12 часов	60 990
3 	4	16	1000 (HDD)	17"/1920x1080 пикс.	до 3 часов	65 990
4 	8	16	1000 (HDD) + 256 (SSD)	17"/1920x1080 пикс.	до 11 часов	109 990
5 	4	16	1000 (HDD)+ 128 (SSD)	17"/1920x1080 пикс.	до 6 часов	?

Пример: оцените стоимость ноутбука

		Кол-во ядер	RAM (ГБ)	Объем жесткого диска (ГБ)	Диагональ/ разрешение	Работа от аккумулятора	Тип ноутбука
1		2	4	500 (HDD)	15"/1920x1080 пикс.	до 5 часов	Офисный
2		4	8	256 (SSD)	14"/1920x1080 пикс.	до 12 часов	Игровой
3		4	16	1000 (HDD)	17"/1920x1080 пикс.	до 3 часов	Офисный
4		8	16	1000 (HDD) + 256 (SSD)	17"/1920x1080 пикс.	до 11 часов	Игровой
5		4	16	1000 (HDD)+ 128 (SSD)	17"/1920x1080 пикс.	до 6 часов	?

Классификация и регрессия



- Основные понятия машинного обучения
 - объекты и признаки
 - выборка
 - целевая переменная
- Представление данных для задачи машинного обучения
- Отличия регрессии и классификации

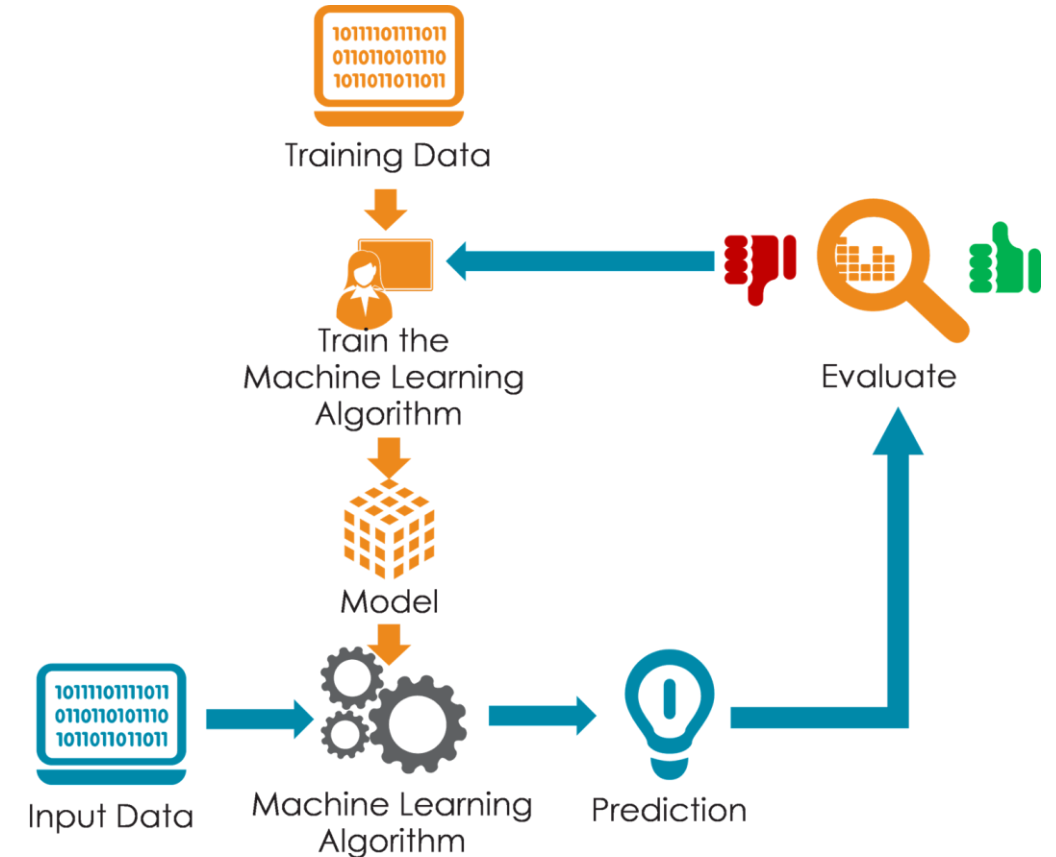
Проблема переобучения и критерии качества

План занятия

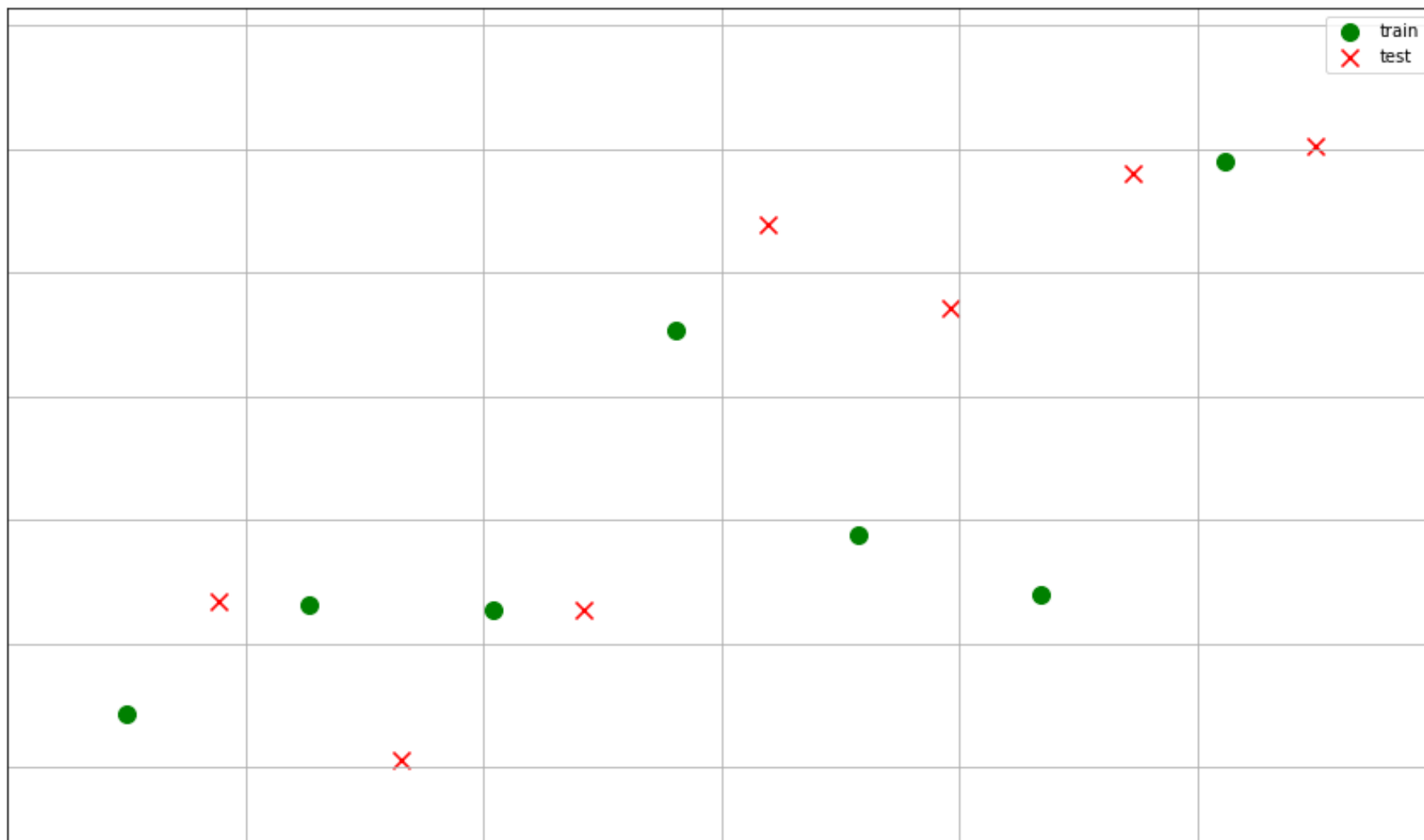
- Переобучение многочленов
- Как определять переобучение?
- Критерии качества моделей

Фреймворк машинного обучения

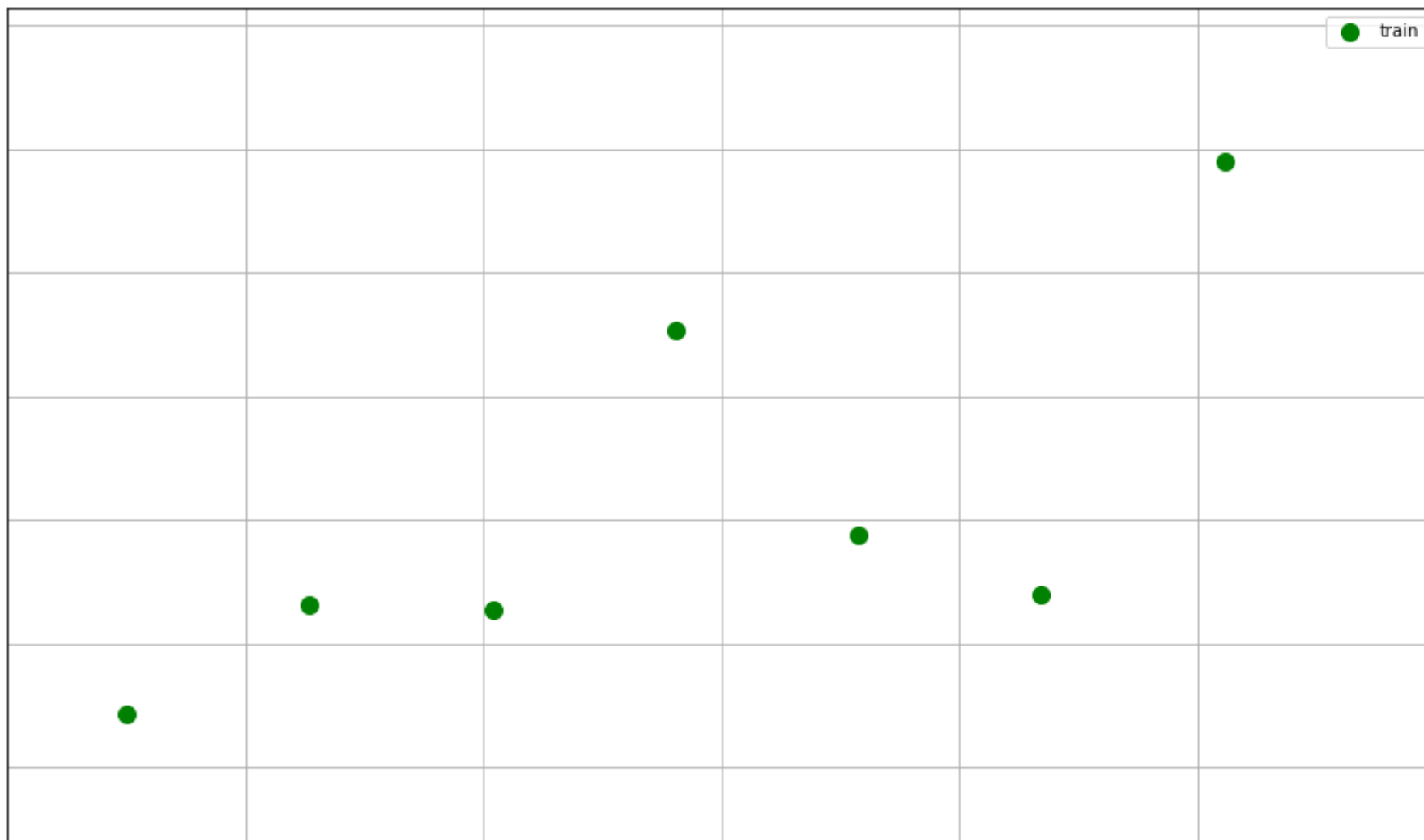
- Формируем матрицу “объекты-признаки” по размеченным данным
- Разбиваем данные на train и test
- Настраиваем алгоритм $\hat{y} : X \rightarrow Y$ так, чтобы \hat{a} приближал y на train
- Тестируем, насколько хорошо \hat{y} приближает y на test



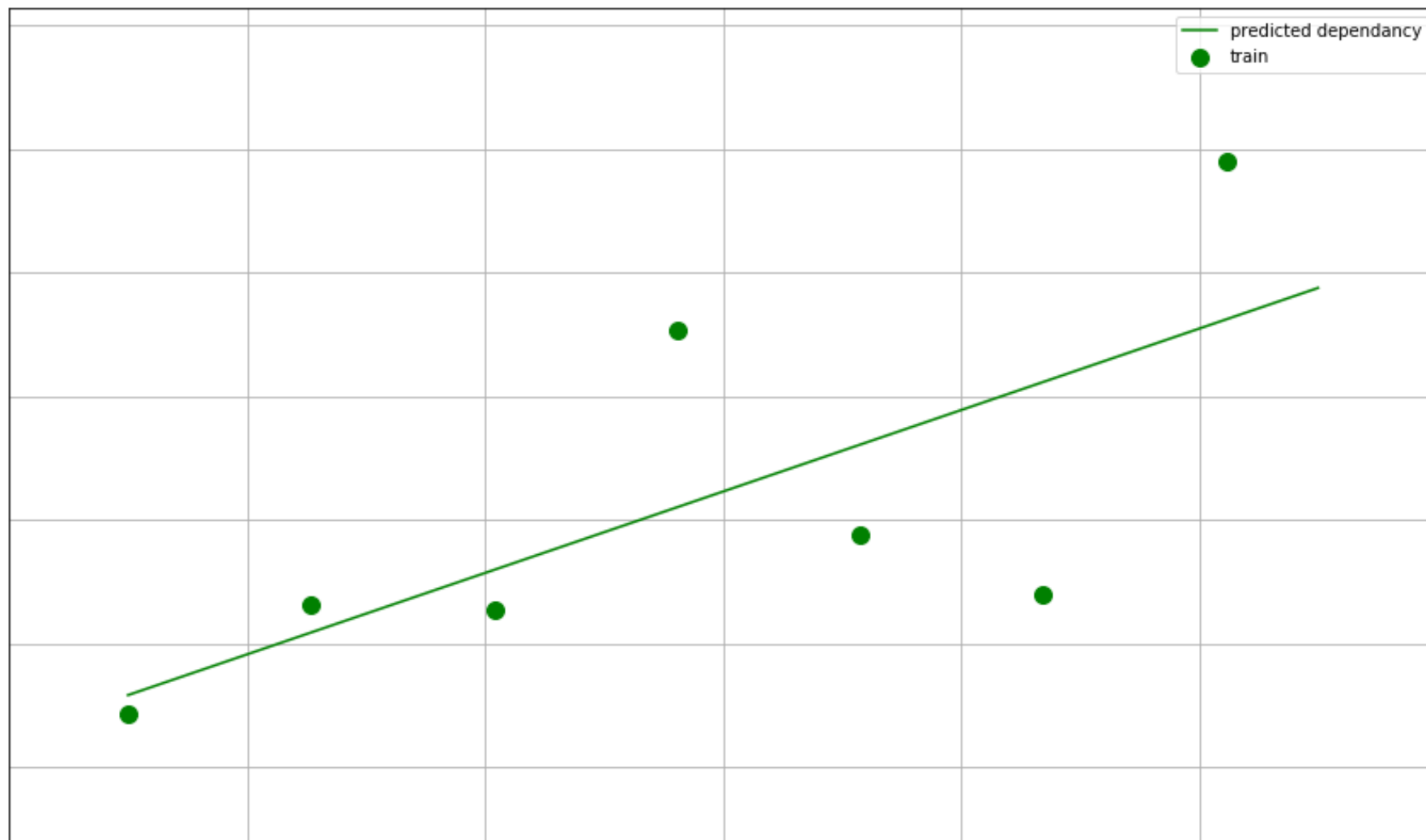
Пример: один признак



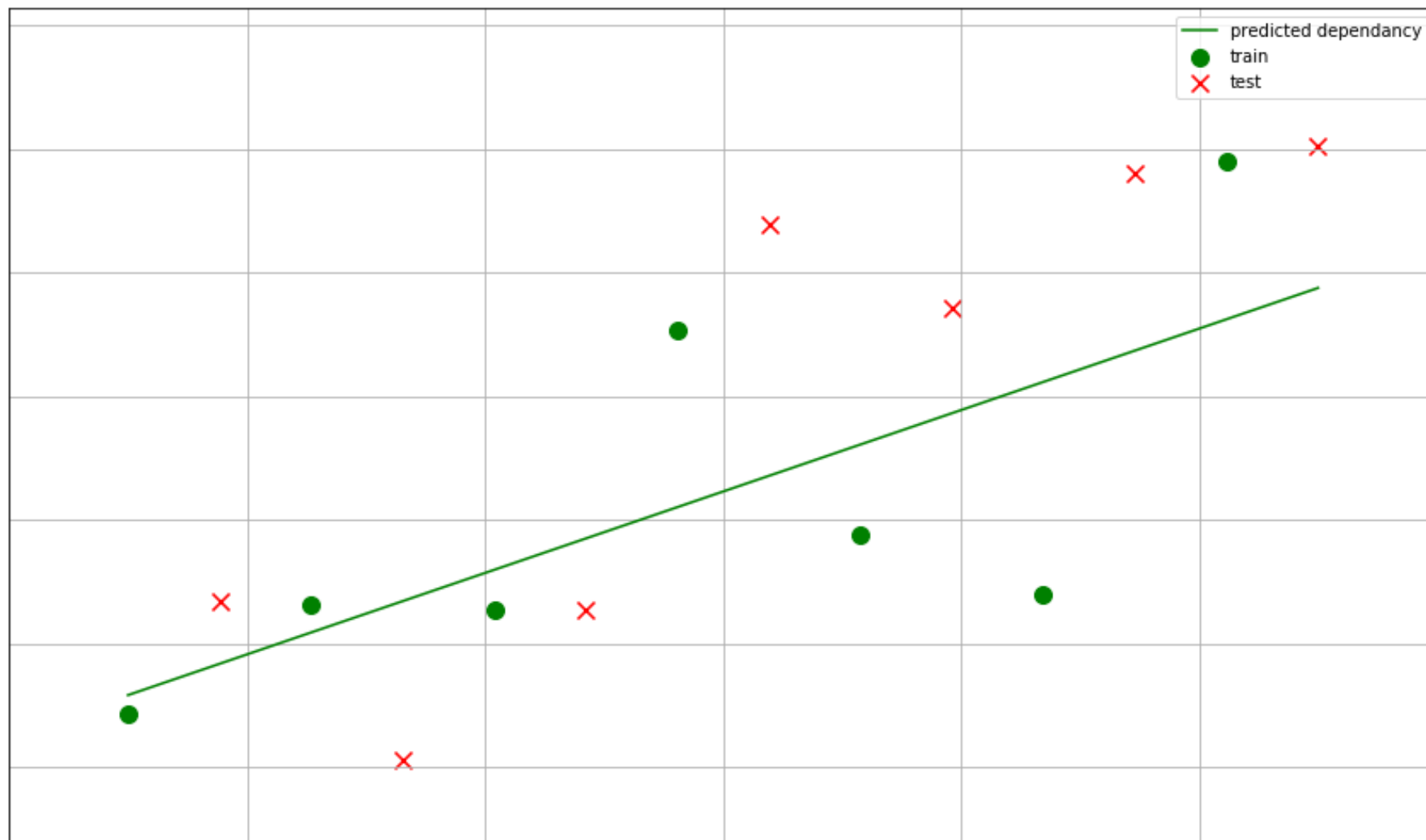
Пример: один признак



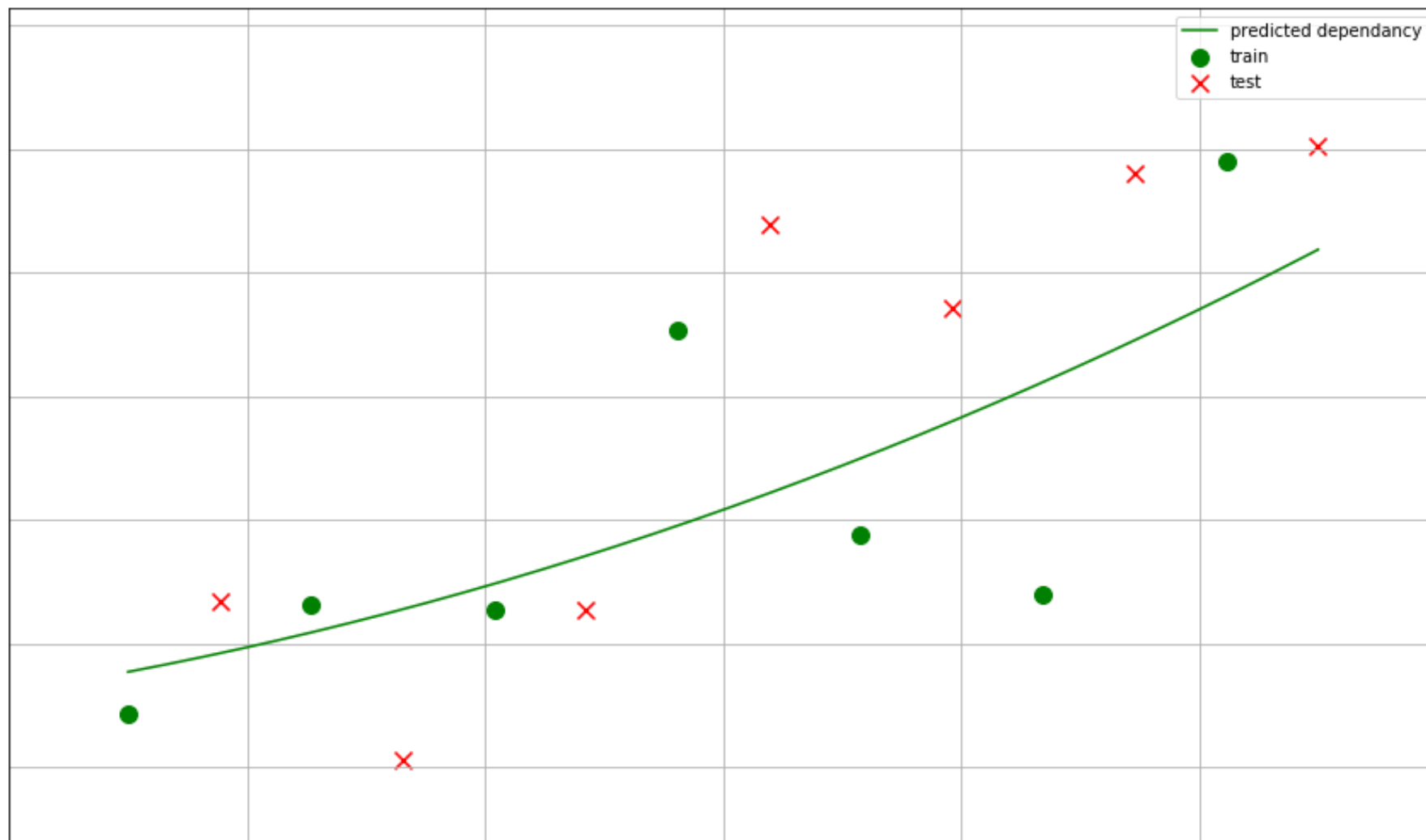
Пример: один признак



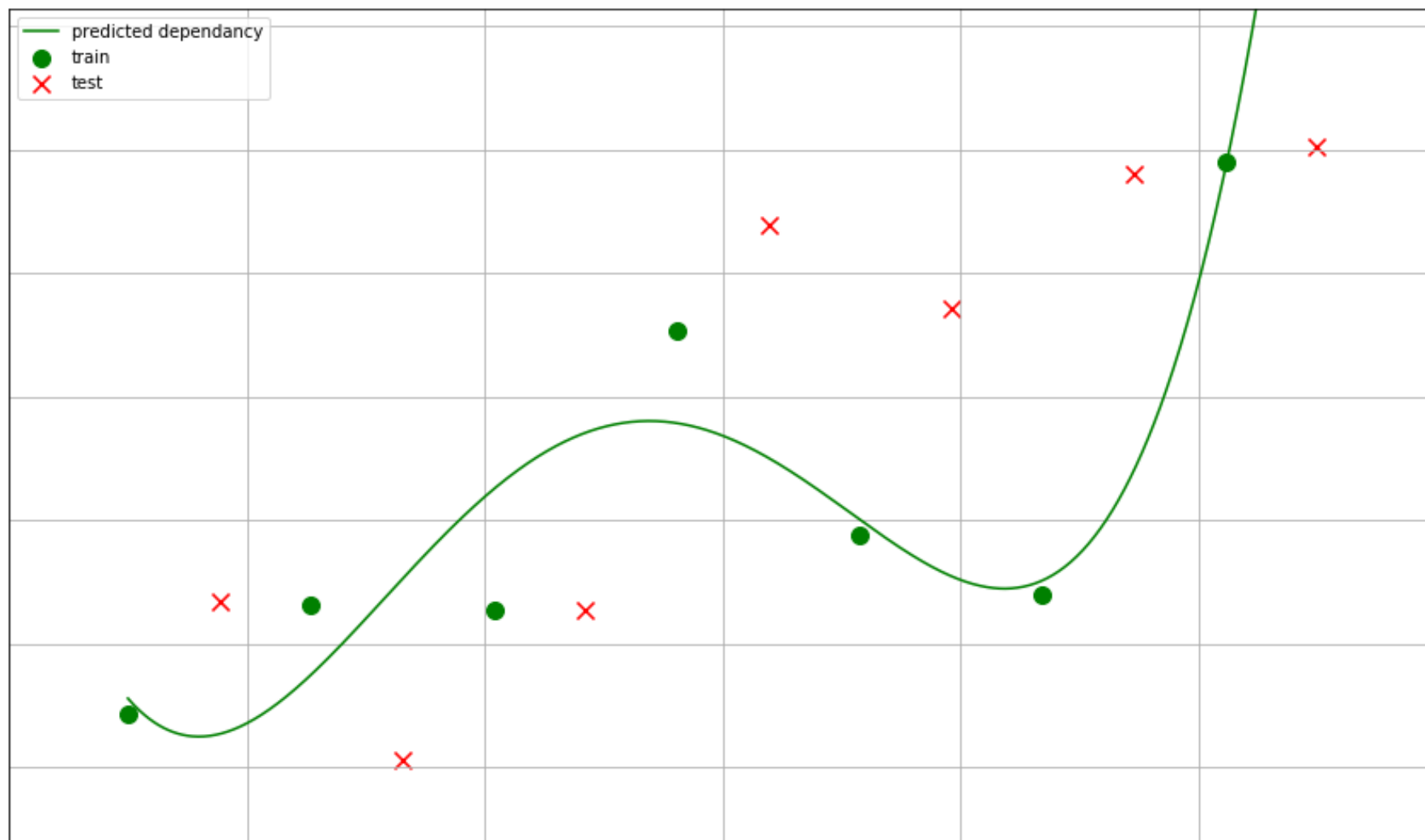
Пример: один признак



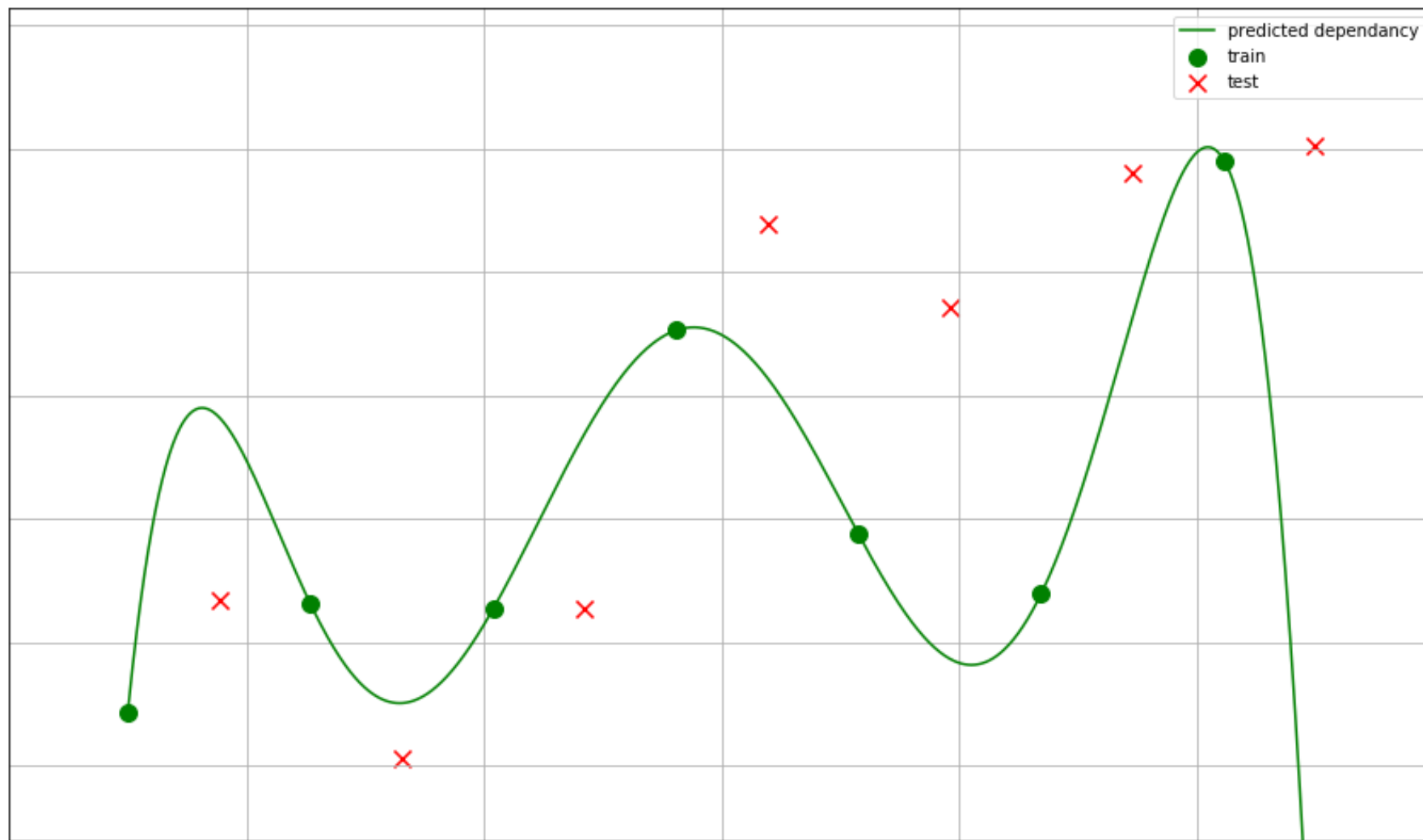
Многочлен степени 2



Многочлен степени 4



Многочлен степени 6



Переобучение

- Переобучение

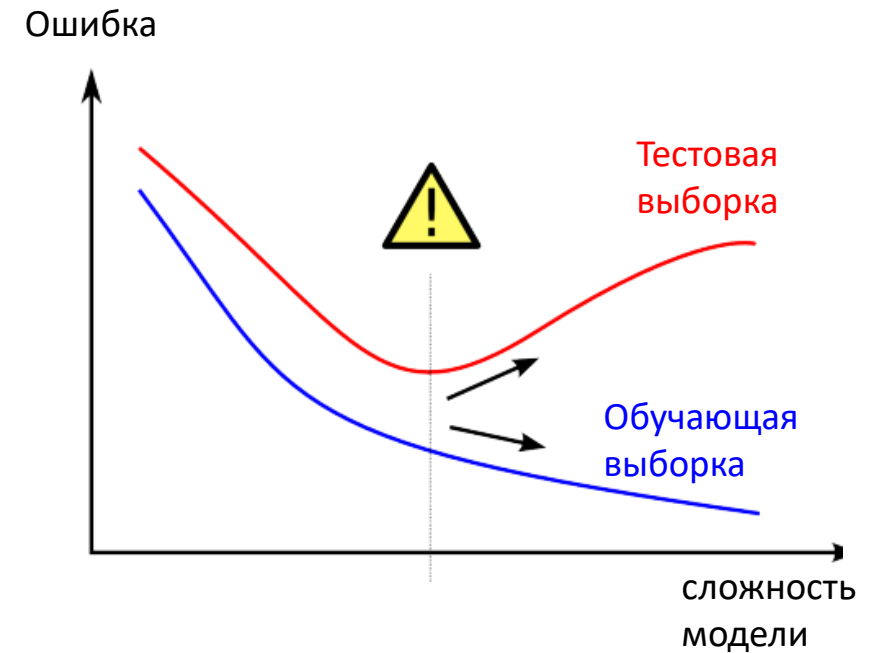
Эффект, при котором модель на тестовой выборке работает хуже, чем на обучающей

- Как обнаружить переобучение?

Разбивать данные на обучающую и тестовую выборки

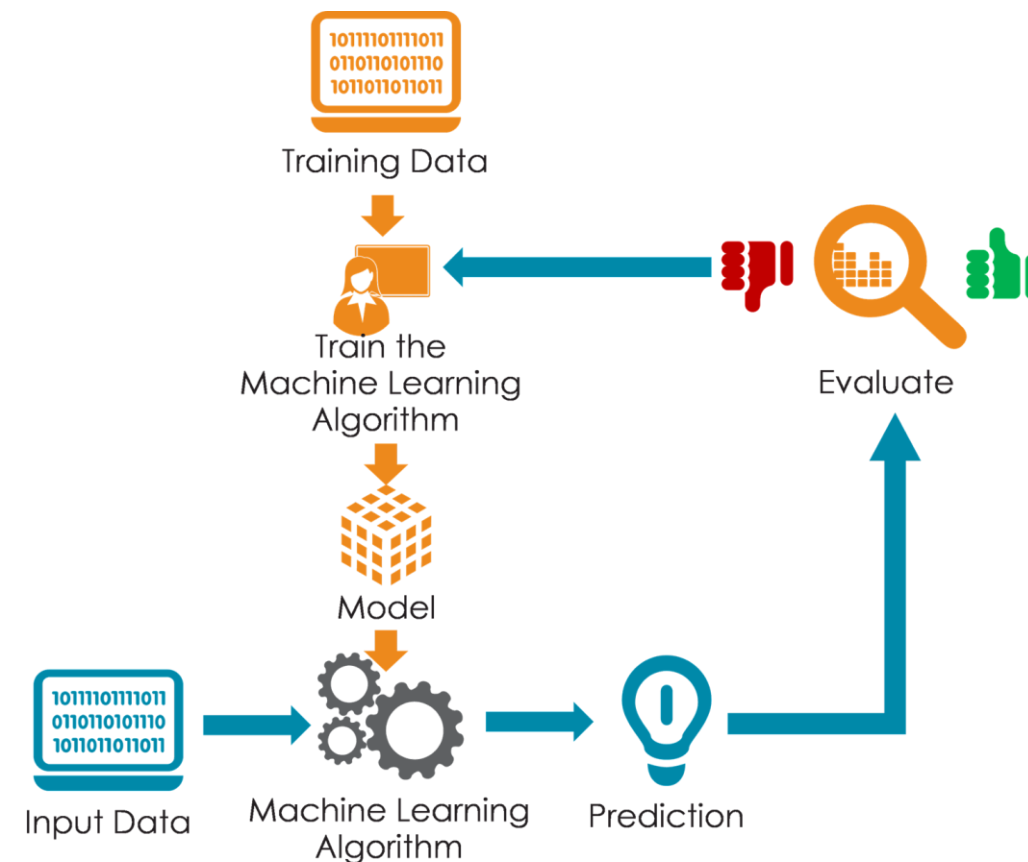
- Как нивелировать его эффект?

Экспериментально находить оптимальную модель



Фреймворк машинного обучения

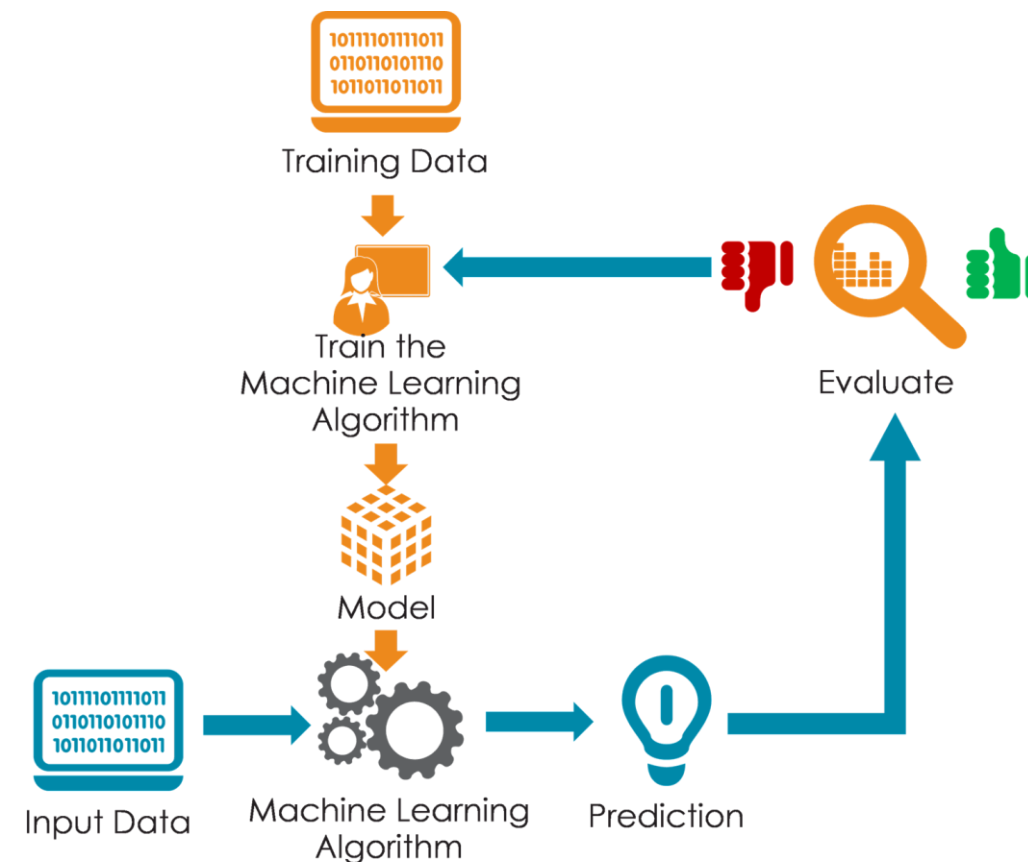
- Формируем матрицу “объекты-признаки” по размеченным данным
- Разбиваем данные на train и test
- Настраиваем алгоритм $\hat{y} : X \rightarrow Y$ так, чтобы a приближал y на train
- Тестируем, насколько хорошо a приближает y на test



Фреймворк машинного обучения

- Формируем матрицу “объекты-признаки” по размеченным данным
- Разбиваем данные на train и test
- Настраиваем алгоритм $\hat{y} : X \rightarrow Y$ так, чтобы a приближал y на train
- Тестируем, насколько хорошо a приближает y на test

Как измерять качество?



Функционалы качества

- Функционал качества (метрика, функция потерь) — численный показатель качества модели
- Для разных задач используются разные метрики

Функционалы качества

- Функционал качества (метрика, функция потерь) — численный показатель качества модели
- Для разных задач используются разные метрики
- Метрика для задачи классификации: Accuracy (доля верных ответов)

Функционалы качества

- Функционал качества (метрика, функция потерь) — численный показатель качества модели
- Для разных задач используются разные метрики
- Метрика для задачи классификации: Accuracy (доля верных ответов)
- Метрики для задачи регрессии
 - Mean Squared Error

$$MSE = \frac{1}{\ell} \sum_{i=1}^{\ell} (\hat{y}_i - y_i)^2$$

- Mean Absolute Error

$$MAE = \frac{1}{\ell} \sum_{i=1}^{\ell} |\hat{y}_i - y_i|$$

- Что такое переобучение и как его находить
- Необходимость разделения на обучающую и тестовую выборки
- Accuracy — метрика классификации, MSE и MAE — метрики регрессии

Алгоритм k ближайших соседей

- Метод ближайших соседей
- Преимущества и недостатки

Предсказание стоимости ноутбука

	Кол-во ядер	RAM (ГБ)	Объем жесткого диска (ГБ)	Диагональ/ разрешение	Работа от аккумулятора	Цена (руб.)
1 	4	16	1000 (HDD)+ 128 (SSD)	17"/1920x1080 пикс.	до 6 часов	?
2 	2	4	500 (HDD)	15"/1920x1080 пикс.	до 5 часов	31 490
3 	4	8	256 (SSD)	14"/1920x1080 пикс.	до 12 часов	60 990
4 	4	16	1000 (HDD)	17"/1920x1080 пикс.	до 3 часов	65 990
5 	8	16	1000 (HDD) + 256 (SSD)	17"/1920x1080 пикс.	до 11 часов	109 990

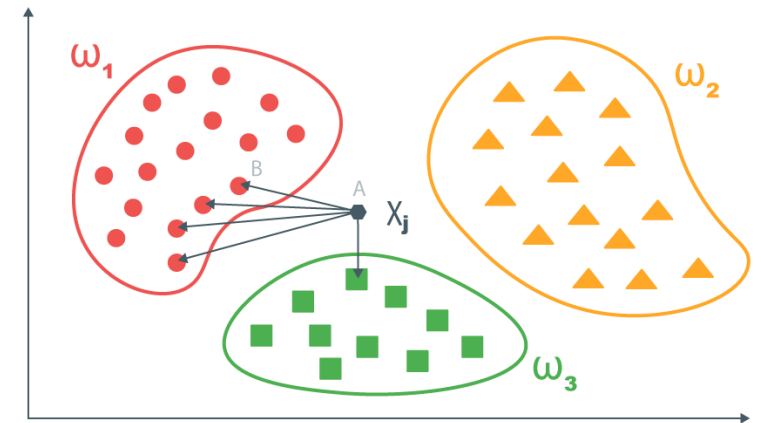
Метод ближайших соседей

Алгоритм k ближайших соседей (k Nearest Neighbors) для задачи классификации:

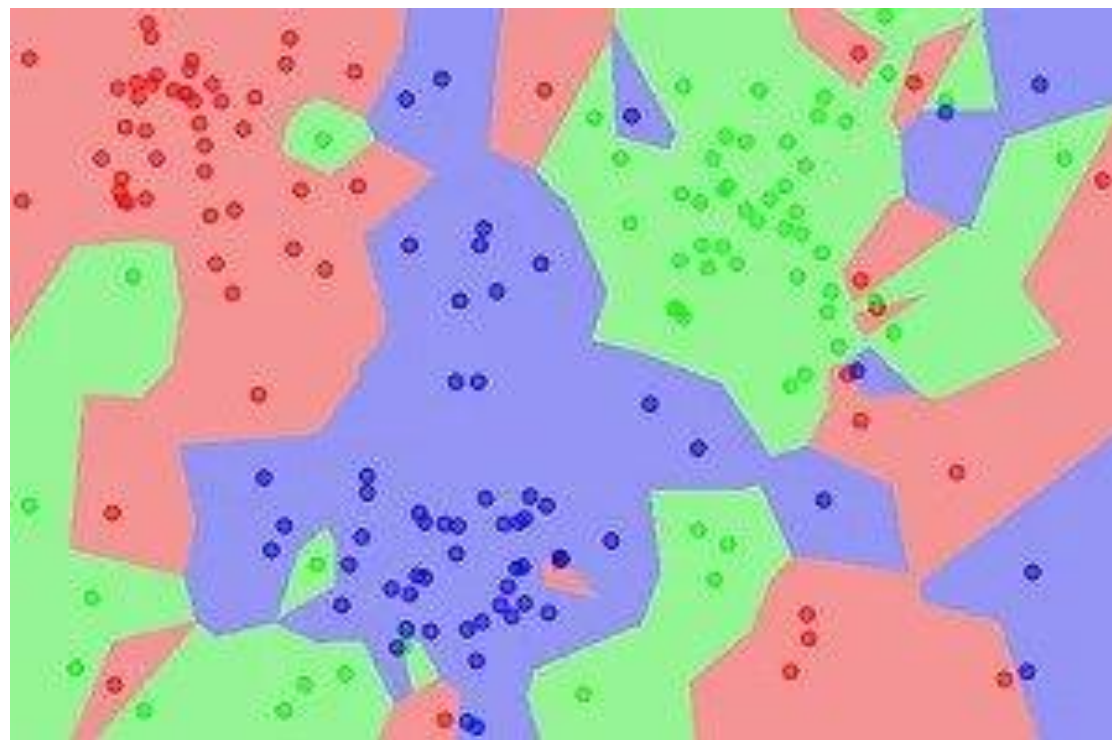
Предсказание:

- Получаем точку x , в которой надо сделать предсказание.
- Ищем k ближайших соседей.
- В качестве ответа возвращаем класс, которого больше всего среди соседей.

Обучение: Просто запоминаем обучающую выборку



Линии разбиения для $k=1$

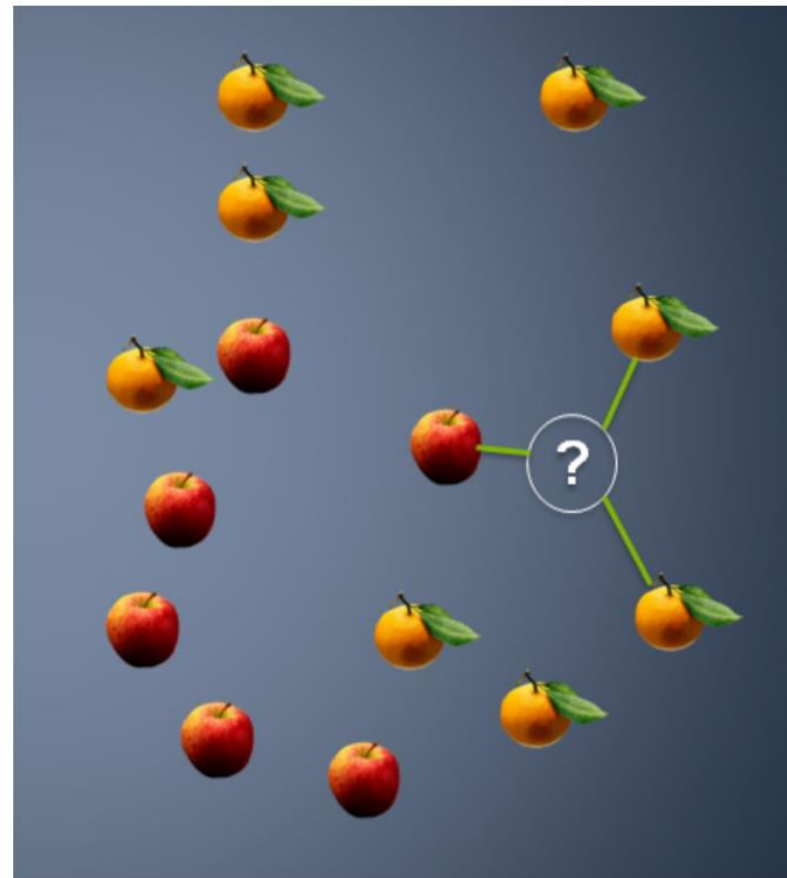


Knn для задачи регрессии

- Усредняем ответы k ближайших объектов

Метод К ближайших соседей

Идея: близким объектам соответствуют близкие ответы



Метод К ближайших соседей

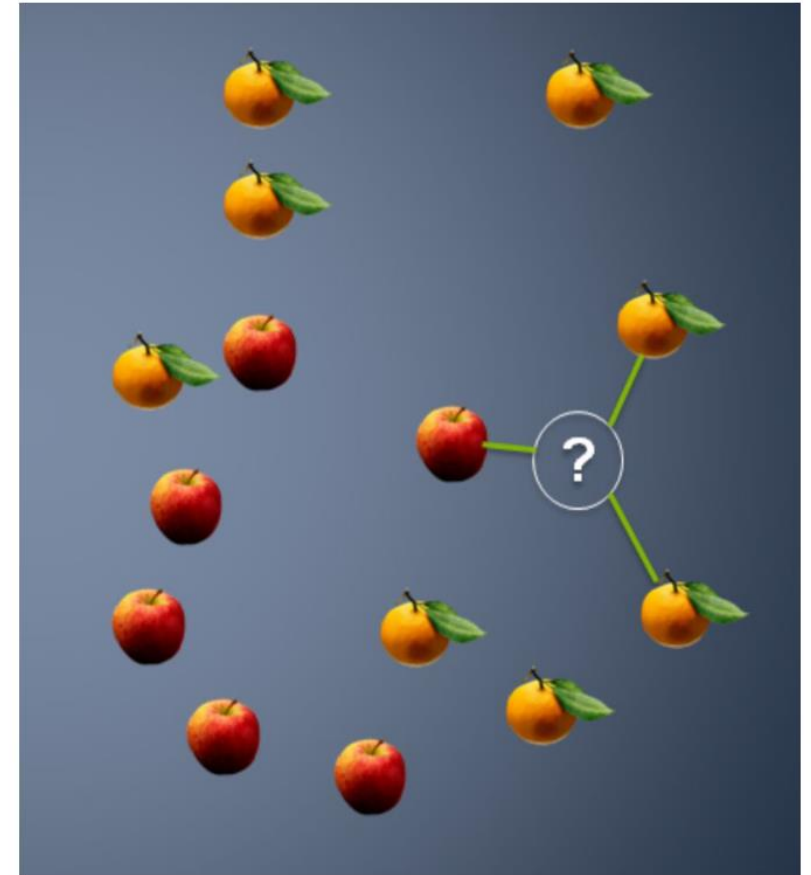
Идея: близким объектам соответствуют близкие ответы.

Формализация понятия **близости**:

Задается функция расстояния между объектами

Пример: обычное расстояние между точками

$$\rho(x, x_i) = \left(\sum_{j=1}^n |x^j - x_i^j|^2 \right)^{1/2}$$



Метод К ближайших соседей

	F1	F2	F3	F4	Y
	площадь	удаленность от центра, км	год постройки	лет после ремонта	цена квартиры
X1	25	3	2005	1	10 рублей
X2	55	10	1987	5	?
X3	50	12	1990	6	12 рублей

Метрика расстояния:

Метод К ближайших соседей

	F1	F2	F3	F4	Y
	площадь	удаленность от центра, км	год постройки	лет после ремонта	цена квартиры
X1	25	3	2005	1	10 рублей
X2	55	10	1987	5	?
X3	50	12	1990	6	12 рублей

Метрика расстояния:

$$\rho(X_1, X_2) = 1289$$

$$\rho(X_2, X_3) = 39$$

Свойства метода k ближайших соседей

- **Интерпретируемый:** Можно понять, почему модель для объекта X выдала тот или иной результат, предъявив похожие на X объекты обучающей выборки
- Требуется задания функции расстояния между объектами. Плохо работает, если функция расстояния не отражает свойства признаков
- Подвержен **проклятию размерности**: работает долго, если в датасете много объектов/много признаков

Масштабирование признаков

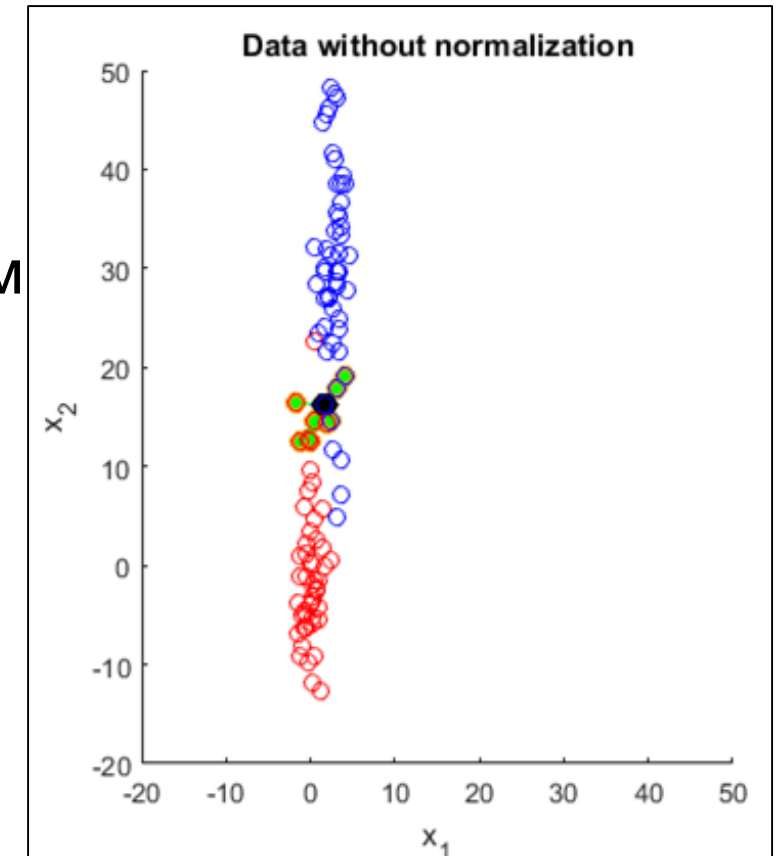
Если в качестве метрики взять обычное расстояние между векторами, то возникает проблема масштаба признаков

Пример

Задача определения стоимости дома по признакам

- Расстояние до метро в метрах
- Количество комнат

Количество комнат почти не будет влиять на предсказание



Итог занятия

- Разобрали KNN, его плюсы и минусы
- KNN подвержен проклятию размерности и чувствителен к масштабу признаков



IT-ОБРАЗОВАНИЕ
В ПЕТЕРБУРГЕ
И УДАЛЕННО

Спасибо за внимание!