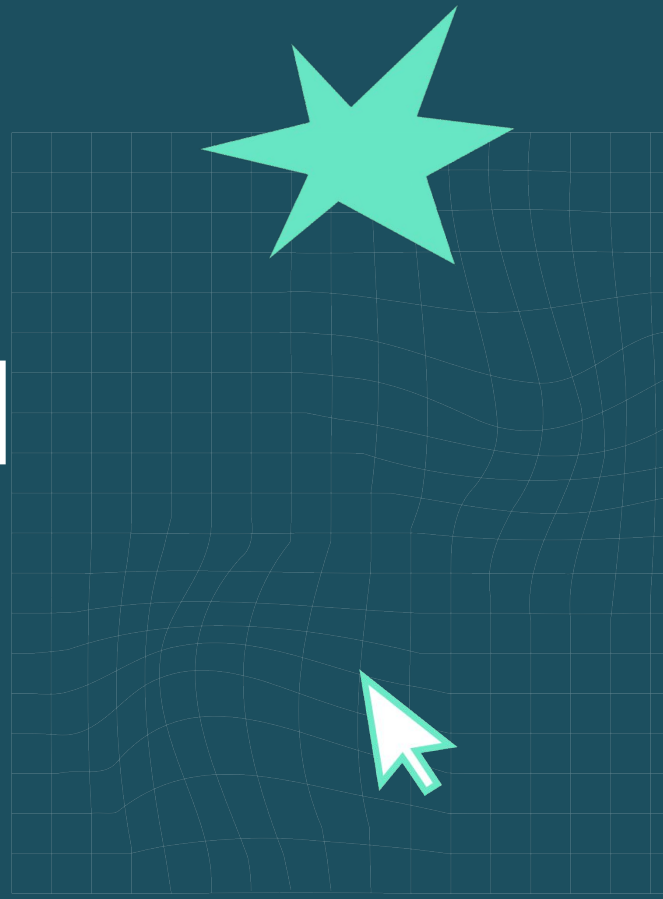
Вклад в будущее
СБЕР

АКАДЕМИЯ
искусственного интеллекта
для школьников

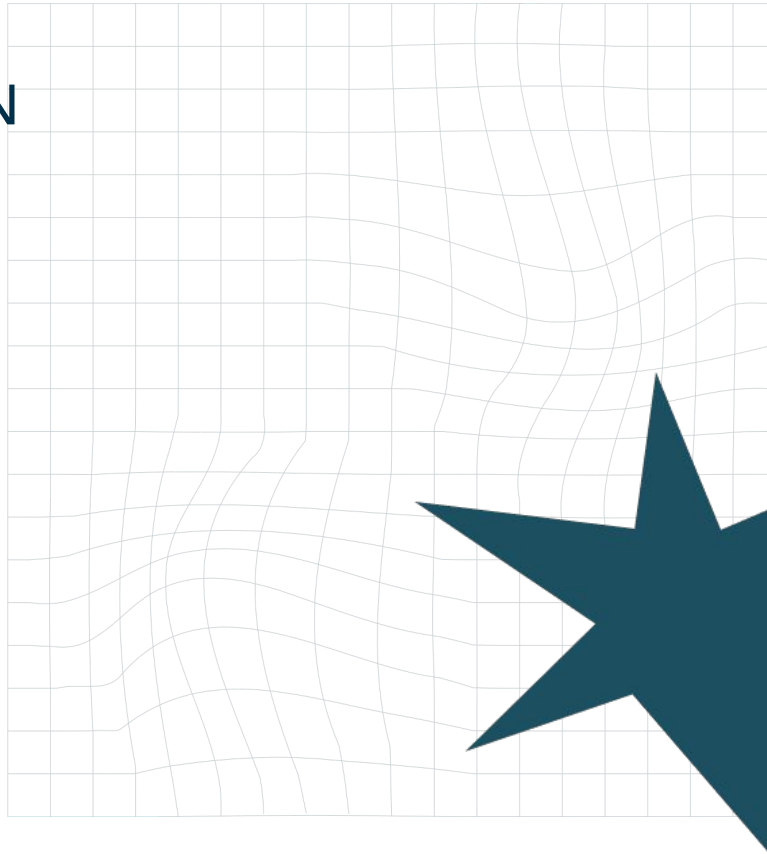
АКАДЕМИЯ ИИ
ДЛЯ ШКОЛЬНИКОВ

Всероссийский образовательный проект
от благотворительного фонда «Вклад в будущее»



Занятие 1: Введение в машинное обучение: линейная регрессия и KNN

Аминов Тимур.
email: nbveha@gmail.com
telegram: @sifonsoul



Вклад в будущее
СБЕР

АКАДЕМИЯ
искусственного интеллекта для школьников

План занятия 1

Анализ данных, машинное обучение и искусственный интеллект

- Чем отличаются эти понятия
- Краткая история развития ИИ

Данные

- Как выглядят данные
- Признаковое описание

Модель

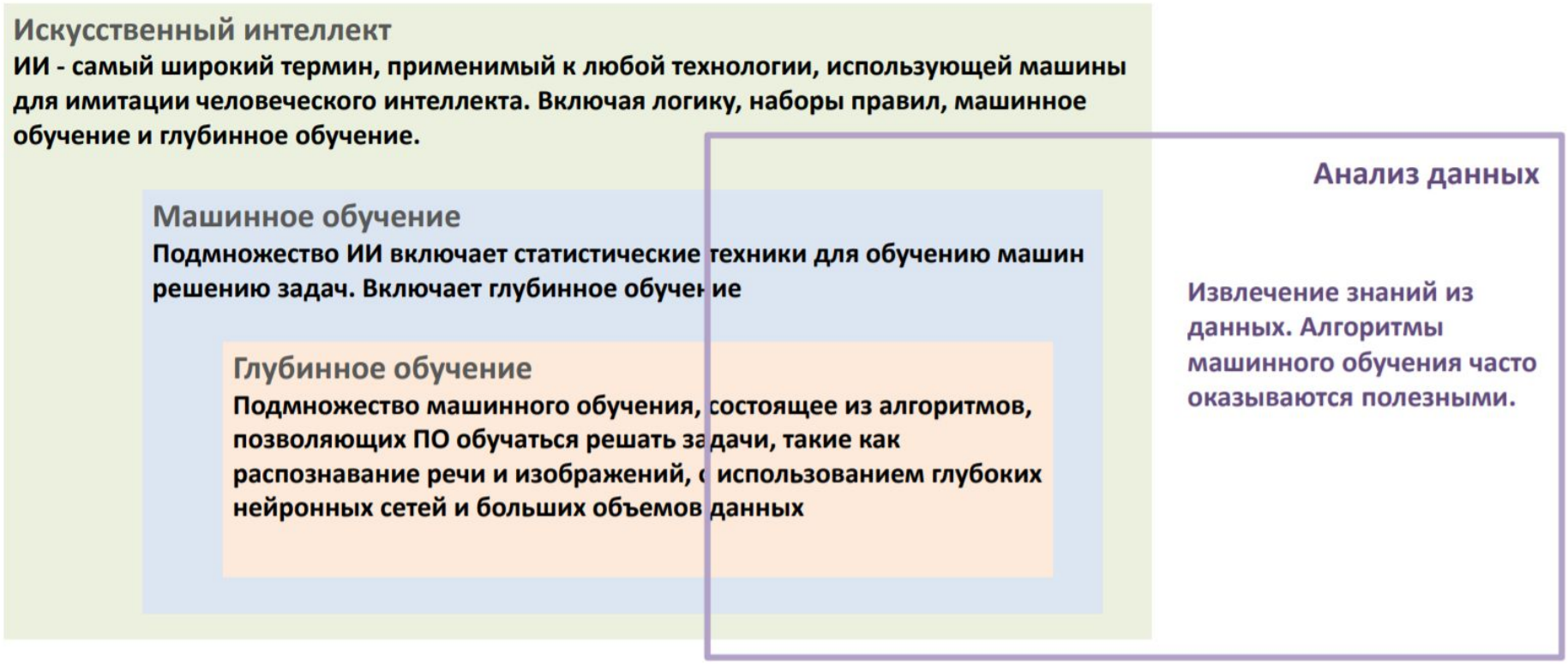
- Обучение с учителем
- Регрессия
- Пример - KNN и линейная регрессия

Измерение качества

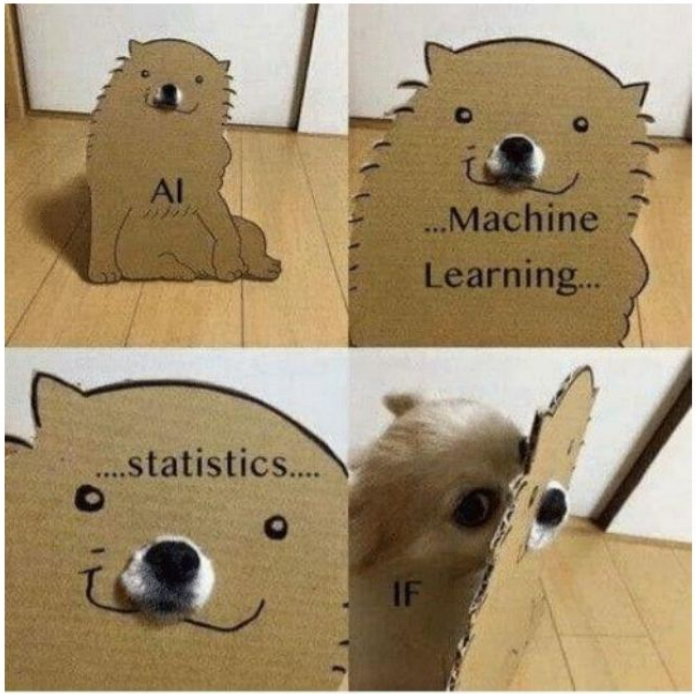
- Примеры метрик
- Разделение на тестовую и тренировочную выборки
- Кросс-валидация
- Переобучение и недообучение

Разница между анализом данных, машинным обучением и искусственным интеллектом

Искусственный интеллект, глубинное и машинное обучение – разные, хотя и похожие понятия



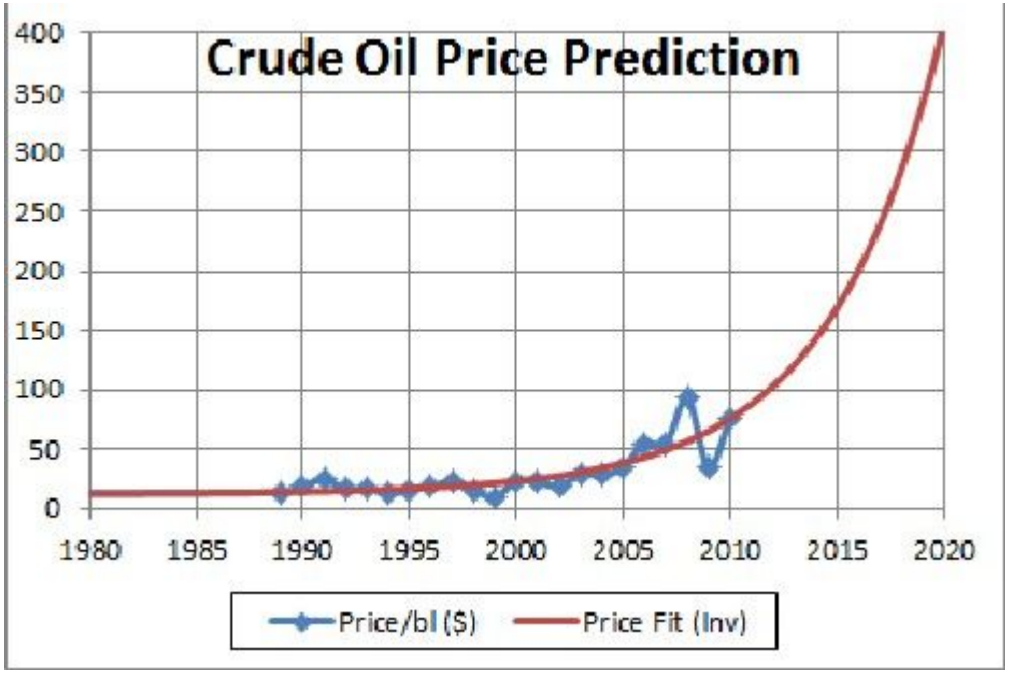
Разница между ИИ и машинным обучением



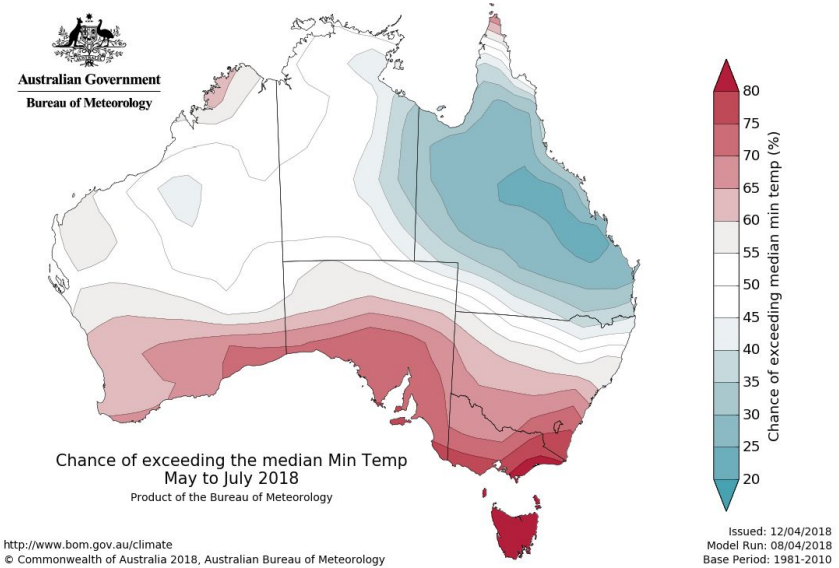
Разница между ИИ и машинным обучением

Если это сделано в Питоне –
это Машинное обучение

Если это сделано в PowerPoint –
это Искусственный интеллект


Crude Oil Price Prediction
400
350
300
250
200
150
100
50
0
1980
1985
1990
1995
2000
2005
2010
2015
2020
Price/bl ($)
Price Fit (Inv)


Good Credit Score


Australian Government
Bureau of Meteorology
Chance of exceeding the median Min Temp
May to July 2018
Product of the Bureau of Meteorology
http://www.bom.gov.au/climate
© Commonwealth of Australia 2018, Australian Bureau of Meteorology
80
75
70
65
60
55
50
45
40
35
30
25
20
Chance of exceeding median min temp (%)
Issued: 12/04/2018
Model Run: 08/04/2018
Base Period: 1981-2010

Сначала интересные

Специально для Вас

Аффинаж, rsac, номер скрыт и другие

СЛУШАТЬ ВСЕ

Google

Приглядитесь к этим предложениям

3 295 ₽ -50 %
~~6 590 ₽~~
Кеды VANS

7 030 ₽ -35 %
~~10 800 ₽~~
Внешняя звуковая

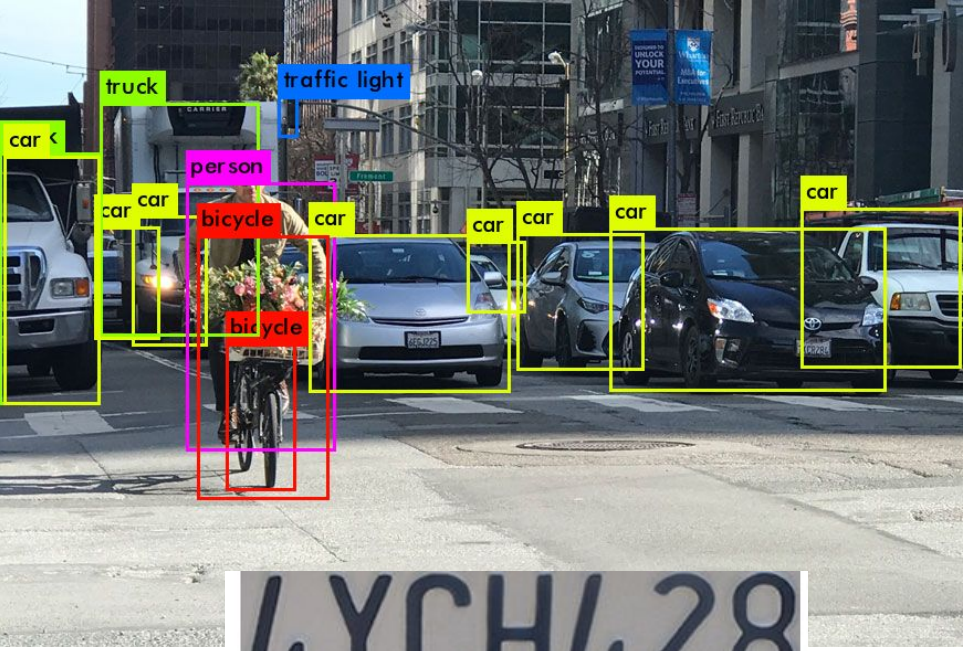
1 875 000 ₽
Виниловый проигрыватель Spira...

4 400 ₽ -30 %
~~6 290 ₽~~
Кеды VANS

11 790 ₽
Лонгборд GoldCoast Standard

где найти

- где найти
- где найти **работу**
- где найти **девушку**
- где найти **друзей**
- где найти **парня**
- где найти **мужа**
- где найти **деньги**
- где найти **ответы на огэ 2018**
- где найти **ответы на егэ 2018**
- где найти **алису**

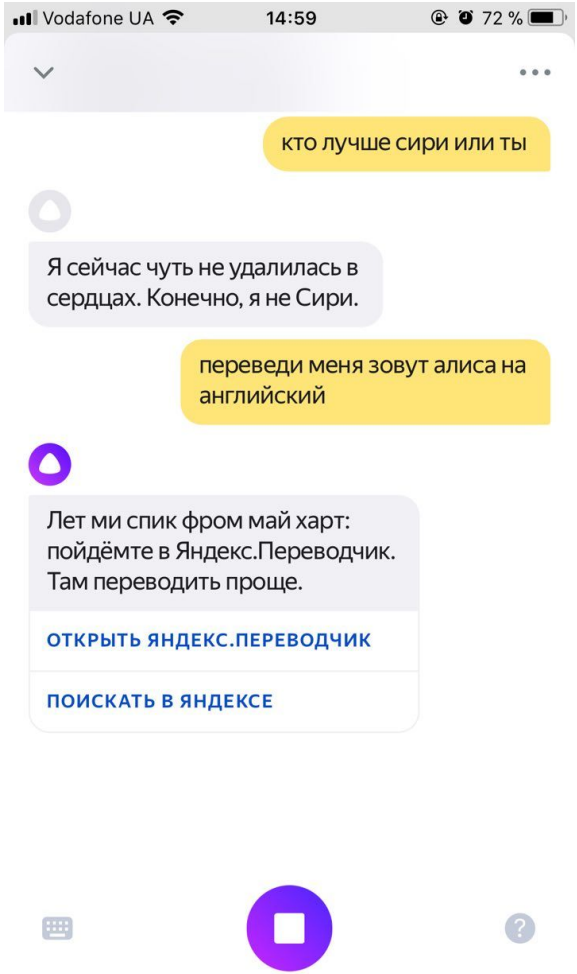

truck
traffic light
car
person
car
car
bicycle
car
car
car
car
car
bicycle

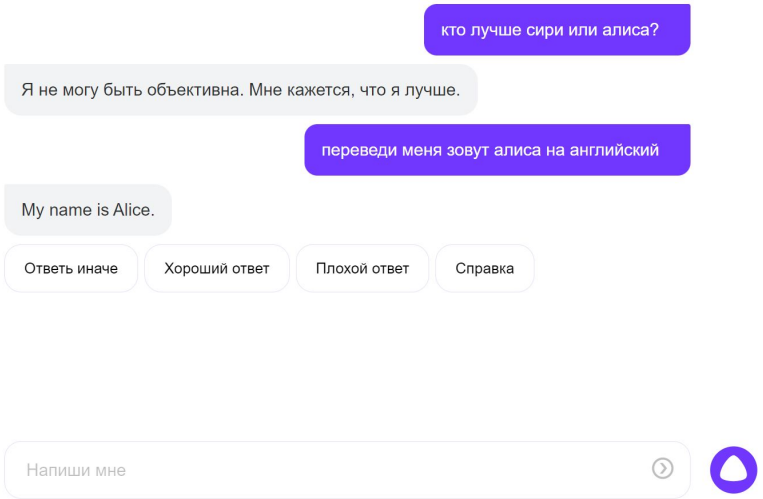

Fetch! FUN
Brittany?
Sporty, strong desire to please, sensitive & devoted
© Getty Images/Microsoft Fetch


4YCH428
4YCH428
4YCH428


35
42
30
34
39

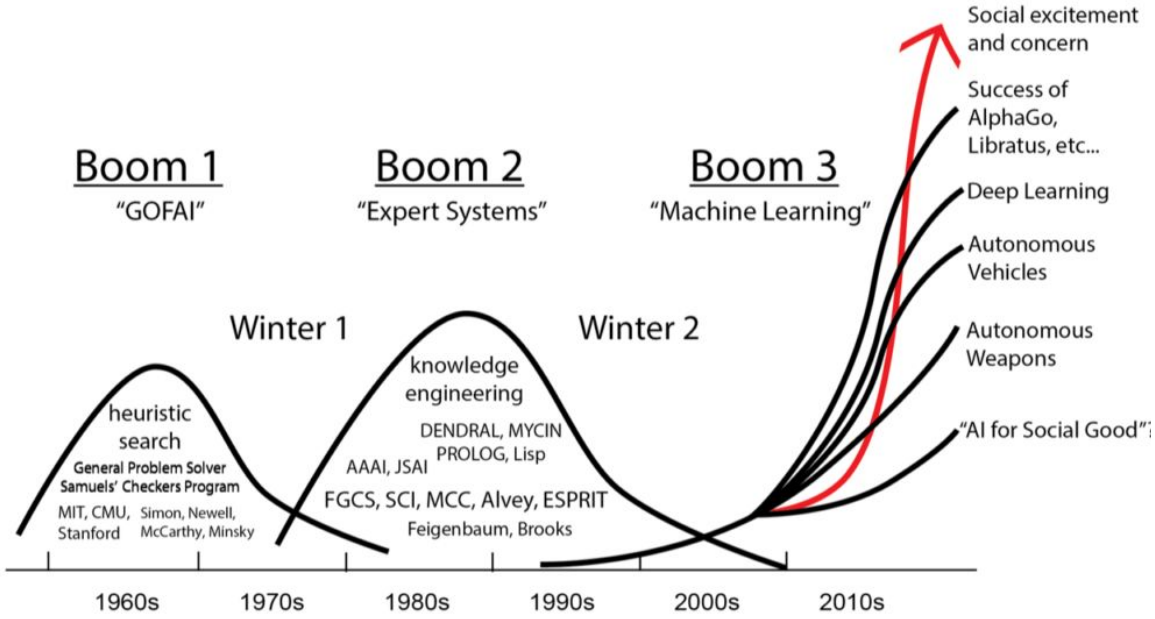
С каждым годом все лучше (старая алиса vs YandexGPT 2)





Почему сегодня?

История развития ИИ нелинейна: люди успели несколько раз разочароваться в ИИ, пока пришли к тому, что имеем сейчас.



Данные

Изображения

MNIST Dataset

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

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

- Таблица как в Excel
- Как часто бывает: десятки столбцов, тысячи строк
- Один из столбцов - целевая переменная
- С данными такого вида мы будем работать на протяжение всего курса

Признаковое описание

Для того, чтобы работать с данными, нужно представить их в виде, пригодном для моделей ML

- Строка в таблице называется **объектом**
- Столбец в таблице называется **признаком**
- Признаки могут быть 3-х типов:
  - Числовые
  - Категориальные
  - Бинарные
- Столбец, который нужно предсказать, называется **целевой переменной**

X y* features

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

Признаковое описание

Все признаки представляются в виде чисел:

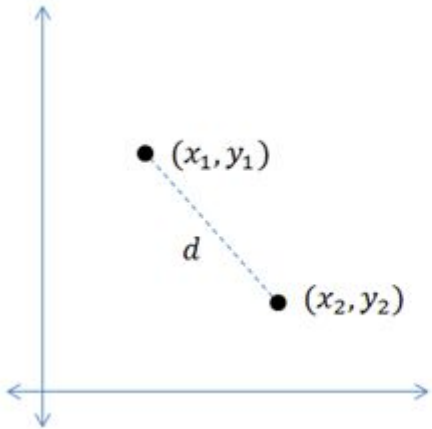
- Числовые признаки - это уже числа
- Бинарные признаки - как 0 и 1
- Категориальные признаки:
  - Как число от 0 до N, где N - число категорий
  - Как N-мерный вектор {0, 0, 1, 0, 0, 0}. Т.н. **one-hot vector**

Для каждого объекта набор его признаков собирается в один вектор

Вектор

- Вектор - это упорядоченный набор чисел
- Вектор - это координаты точки в пространстве

Для двух точек можно рассчитать расстояние между ними



$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Матрица

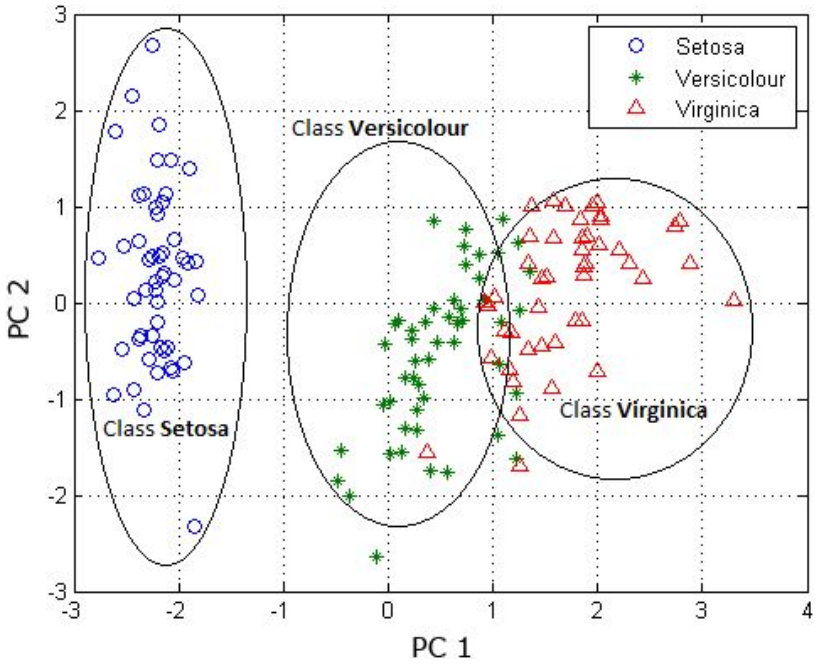
- Матрица - это упорядоченный набор векторов одного размера
- Набор векторов - это набор точек в пространстве

Датасет с подготовленными признаками - это матрица

Визуализация данных

Датасет - это матрица

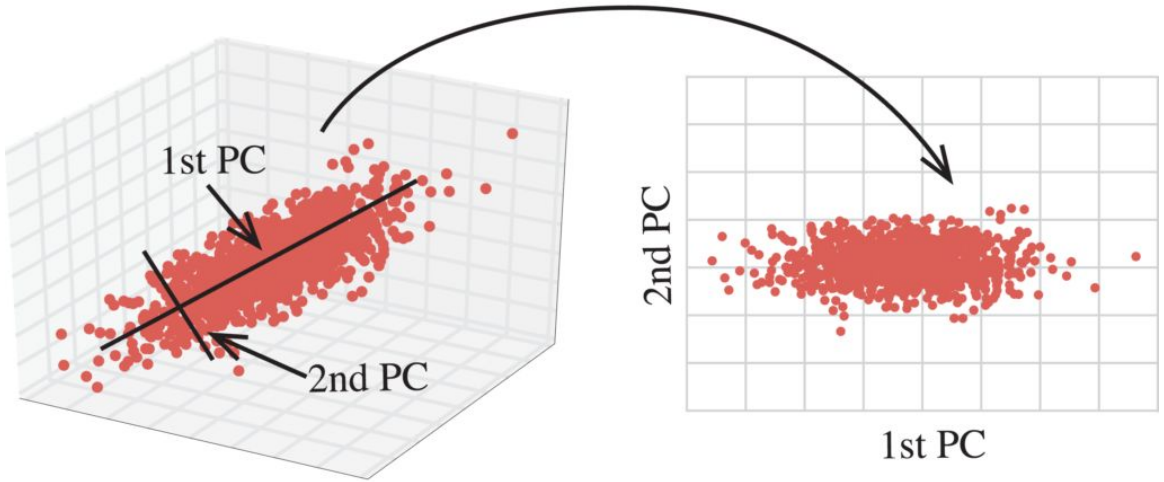
petal length	petal width	target
1.4	0.2	Iris-setosa
1.4	0.2	Iris-setosa
1.3	0.2	Iris-setosa
1.5	0.2	Iris-setosa
1.4	0.2	Iris-setosa



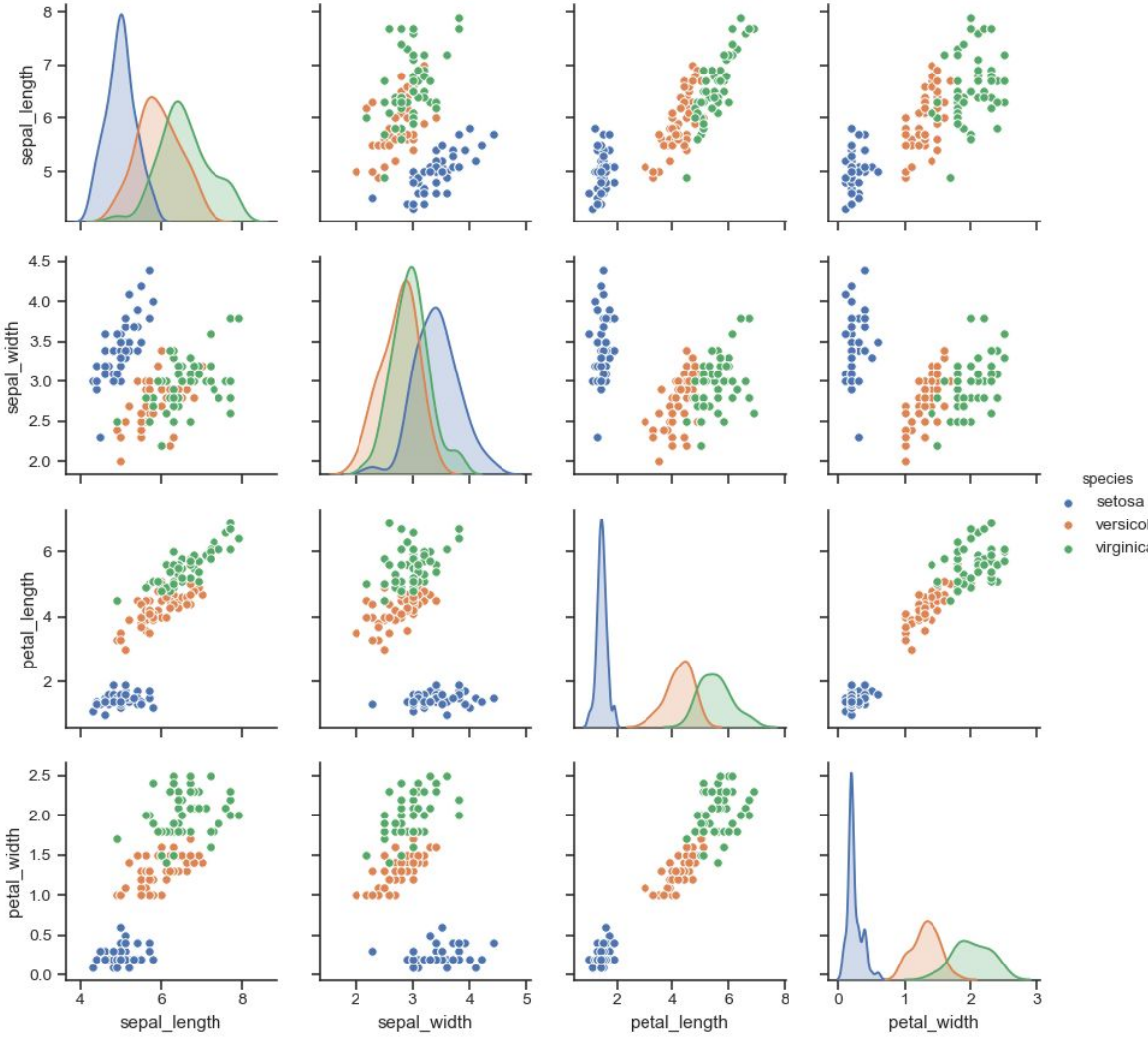
Вообще-то, в этом датасете признаков не 2, а 4. Как на них смотреть?

Большие размерности

- Двумерный набор точек можно нарисовать на плоскости
- Трехмерный набор можно спроецировать на плоскость
- Размерности векторов могут быть порядков 100~100 000
- Их всё равно можно спроецировать!

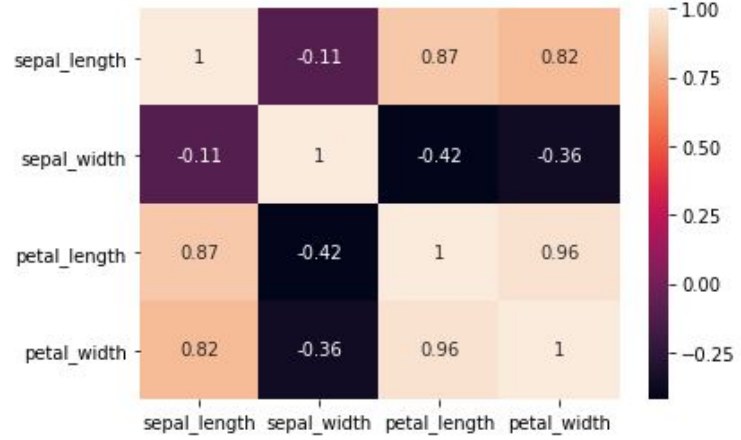


Визуализация реальных датасетов



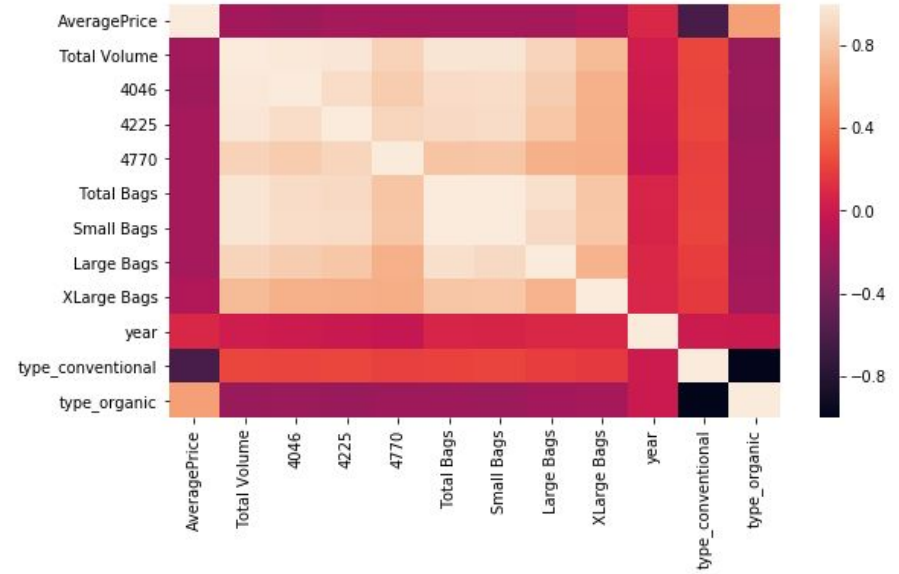
Матрица корреляции

- Если признаков много, то все парные графики не рассмотришь
- Визуально проанализировать признаки помогает матрица корреляции



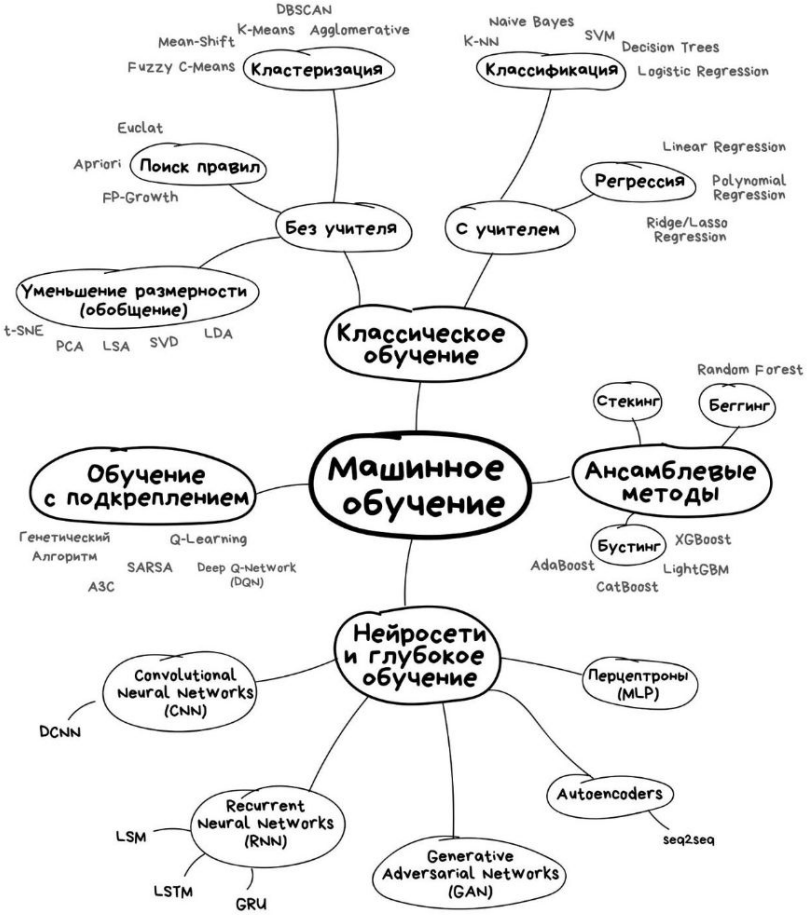

```
plt.figure(figsize=(9,5))
sns.heatmap(df.corr())
```

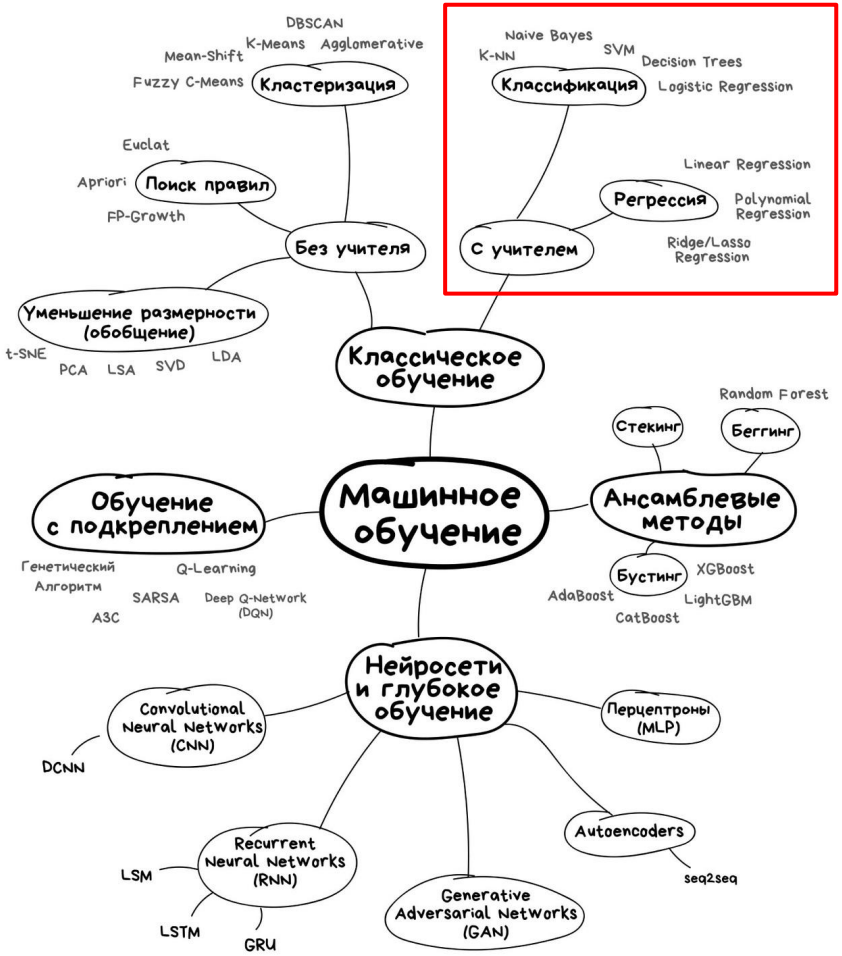
```
<matplotlib.axes._subplots.AxesSubplot at 0x1fa0e4d4a20>
```



Модель


Искусственный Интеллект
Машинное Обучение
Нейросети
Глубокое Обучение
сотни других методов обучения


Машинное обучение
Классическое обучение
Без учителя
С учителем
Кластеризация
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Классификация
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Поиск правил
Euclat
Apriori
FP-Growth
Регрессия
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
Уменьшение размерности (обобщение)
t-SNE
PCA
LSA
SVD
LDA
Обучение с подкреплением
Генетический Алгоритм
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
Ансамблевые методы
Стекинг
Беггинг
Random Forest
Бустинг
XGBoost
AdaBoost
LightGBM
CatBoost
Нейросети и глубокое обучение
Convolutional Neural Networks (CNN)
DCNN
Перцептроны (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq


Машинное обучение
Классическое обучение
Обучение с подкреплением
Ансамблевые методы
Нейросети и глубокое обучение
Без учителя
С учителем
Кластеризация
Классификация
Регрессия
Поиск правил
Уменьшение размерности (обобщение)
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
Euclat
Apriori
FP-Growth
t-SNE
PCA
LSA
SVD
LDA
Стекинг
Беггинг
Random Forest
Бустинг
XGBoost
AdaBoost
LightGBM
CatBoost
Генетический Алгоритм
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
Convolutional Neural Networks (CNN)
DCNN
Перцептроны (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq

Задача обучения с учителем

- Между объектом и целевой переменной существует **реальная зависимость**
- У нас есть только N сэмплов этой зависимости - **обучающая выборка**
- Задача - научиться **предсказывать** целевую переменную для новых точек
- Для этого строится **модель**

Модель - это функция, которой можно аппроксимировать реальную зависимость, имея конечное число примеров.

Классификация и регрессия

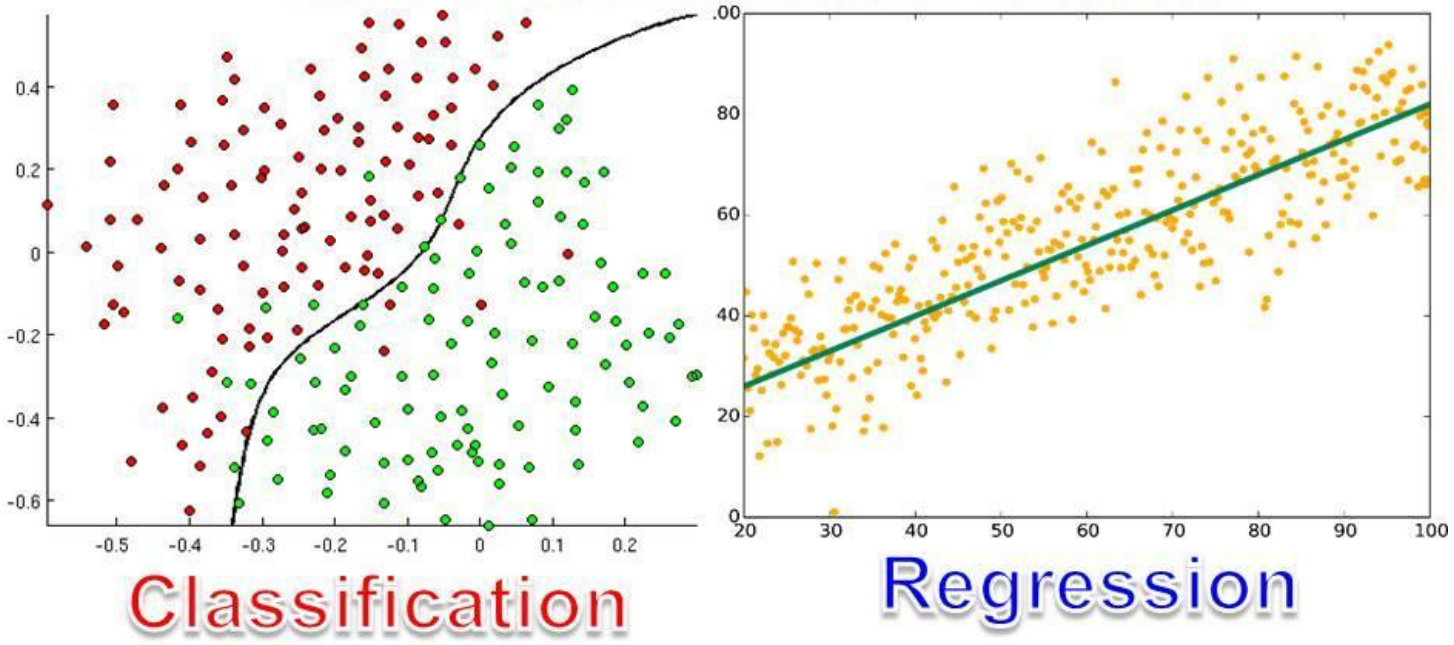
Целевая переменная, как и признаки, может быть трех типов:

- Числовая
- Бинарная
- Категориальная

Предсказание числового значения называется **регрессия**

Предсказание одного из нескольких классов называется **классификация**

Классификация и регрессия

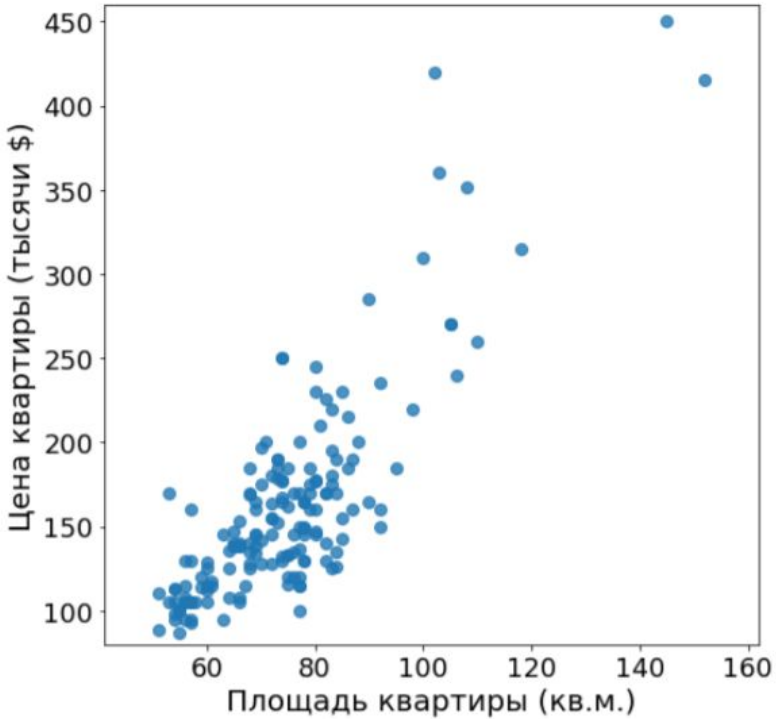


Машинное обучение – инструмент для решения реальных задач

- Кредитный скоринг: понять, будет ли клиент хорошим заемщиком по доступным данным
- Видеоаналитика: по видео понять, работает ли персонал в касках
- Система идентификации по лицу (оплата, вход на предприятие)
- Извлечение информации из баз знаний

Машинное обучение – инструмент для решения реальных задач

Компоненты постановки задачи машинного обучения



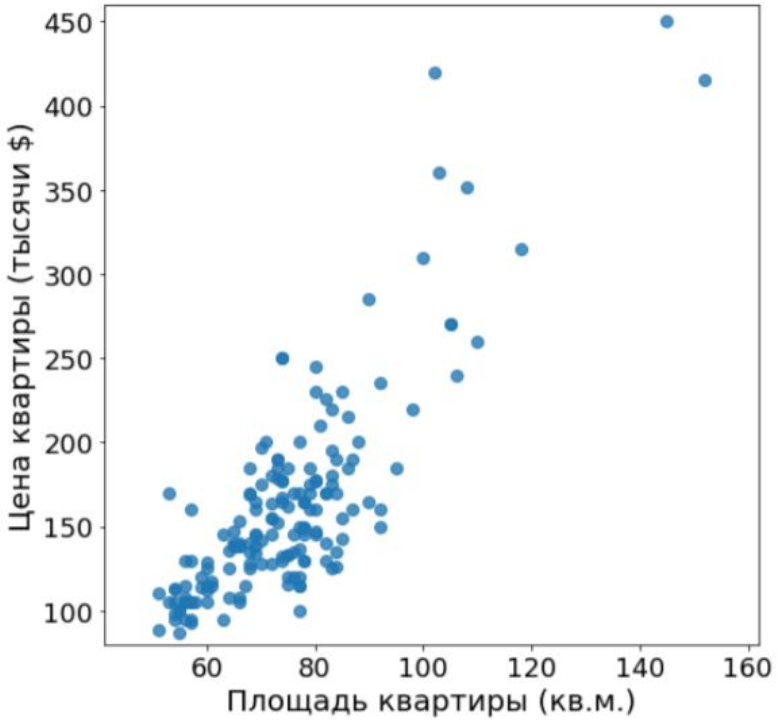
Линейная регрессия

Линейные методы

Машинное обучение – инструмент для решения реальных задач

Обычная задача машинного обучения: построить модель зависимости «выхода» от «входа»

- Вход: x, площадь квартиры
- Выход: y(x), цена квартиры
- Каждая точка на графике справа – один объект в нашей выборке



Данные

- Вход: x, площадь квартиры
- Выход: y(x), цена квартиры
- Данные: такими были данные по объявлениям, опубликованным в прошлом году

Признаки и целевая

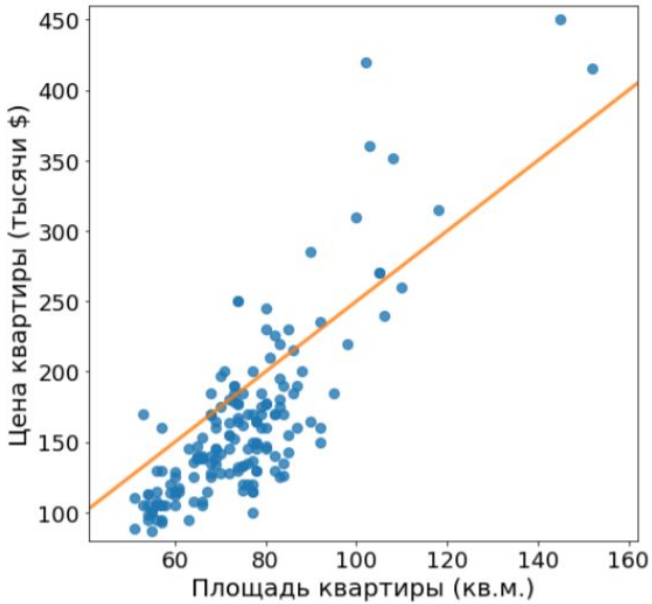
Объекты	x		y
	Площадь кв.м.	Расстояние до центра, км	Цена тысячи $
	77	9	115
	79	9	175
	84	11	170
	65	8	140

Табличные данные: «Excel-таблица»

Модель

Хотим построить модель зависимости «выхода» от «входа»

- Вход: x, площадь квартиры
- Выход: y(x), цена квартиры
- Модель y(x): если площадь квартиры 100 кв. м, то ее цена 250 тыс. $

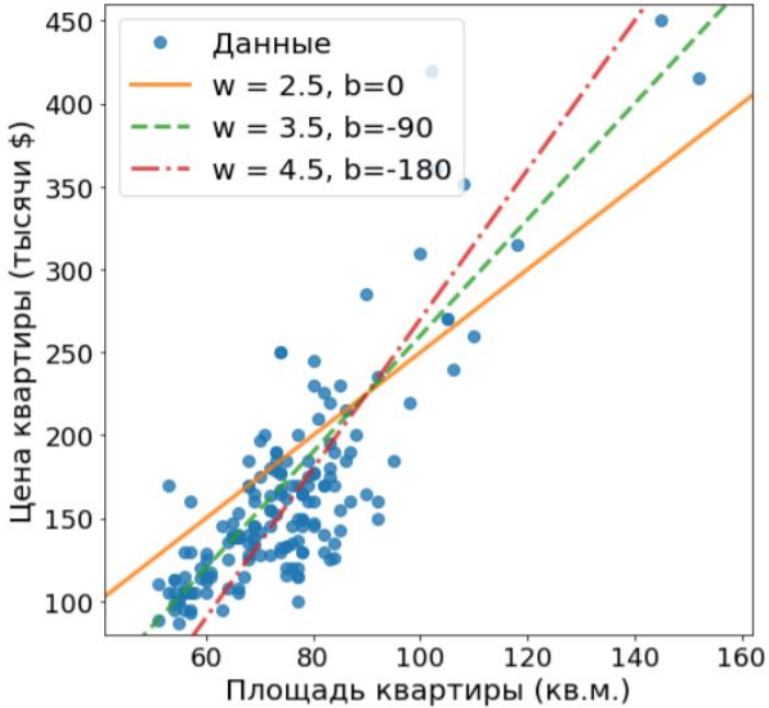


Решение задачи: линейная модель

Пример модели – модель линейной регрессии

$$\hat{y}(x) = w\,x + b$$

- y – настоящая цена квартиры
- ŷ(x) - то, что выдала модель
- x – площадь квартиры
- w, b – коэффициенты (параметры) модели линейной регрессии



Модель должна быть хорошей!

- Нам хочется, чтобы предсказания модели ŷ(x) были похожи на реальные значения y(x)
- Мера несоответствия между прогнозом и реальностью – квадратичная ошибка (squared error)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}(\boldsymbol{x}_i) - y_i)^2$$

$$\mathrm{SE}(\mathrm{x}) = (\hat{y}(\boldsymbol{x}) - y)^2$$

$\boldsymbol{x}_i$	y_i	$\hat{y}(x)$	SE
77	115	122.5	56.25
79	175	177.5	6.25
84	170	173.0	9
65	140	145.0	25

MSE = 24.275

Используют разные функции потерь

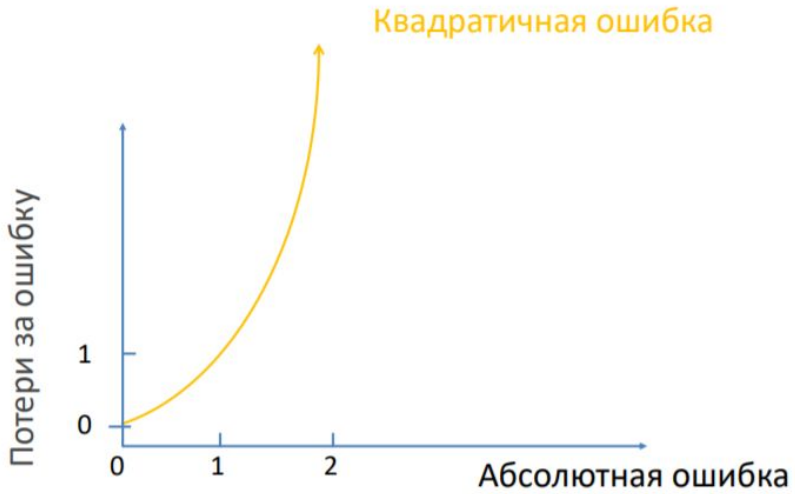
Чем больше модель ŷ(x)ошибается на конкретных объектах – тем она хуже

- Абсолютная ошибка

$$AE = |\hat{y}(x) - y(x)|$$

- Квадратичная ошибка (squared error)

$$SE = (\hat{y}(x) - y(x))^2$$



Линейная регрессия

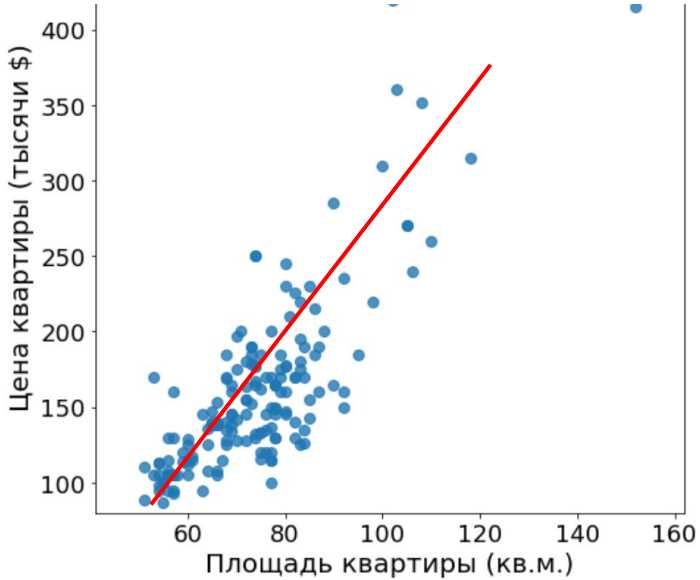
	F1	F2	F3	F4	F5	Y
	площадь	удаленность от центра, км	год постройки	лет после ремонта	тип постройки	цена квартиры
X1	25	3	2005	1	1	
X2	55	10	1987	5	2	
X3	50	12	1990	6	5	
	k_1	k_2	k_3	k_4	k_5	y

$$y = \sum F_i k_i + k_0$$

Линейная регрессия

$$y = \sum F_i k_i + k_0$$

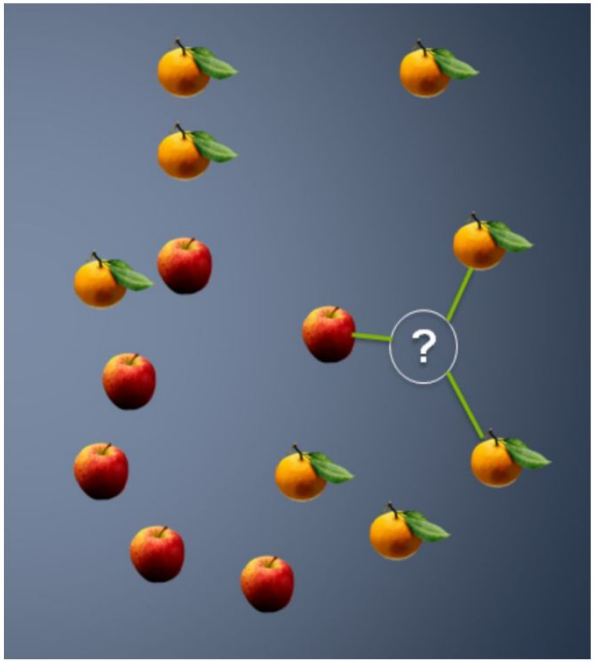
Оптимизируем мет $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}(x_i) - y_i)^2$



Метод К ближайших соседей

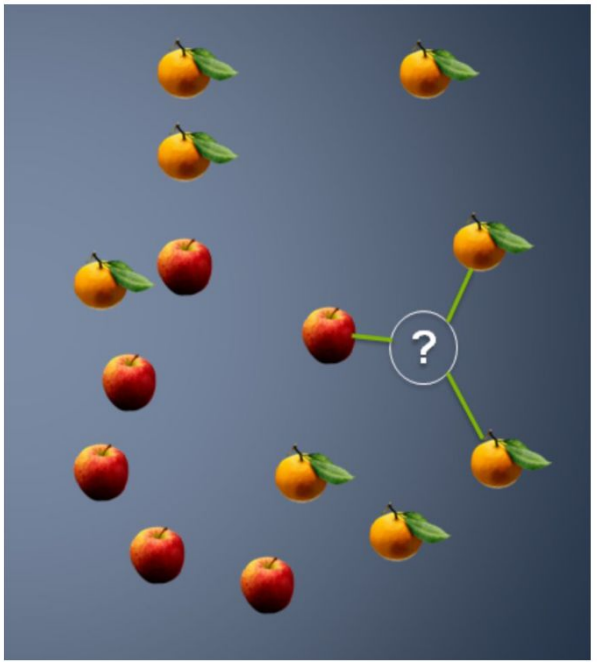
Метод K ближайших соседей

Идея: близким объектам соответствуют близкие ответы



Метод K ближайших соседей

Идея: близким объектам соответствуют близкие ответы.



Формализация понятия **близости:**

Задается функция расстояния между объектами

$$\rho: X \times X \to [0, \infty)$$

Пример: Евклидово расстояние

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

Метод K ближайших соседей

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

Метрика расстояния: $\rho(X_1, X_2) = \sum (F_i^{X_1} - F_i^{X_2})^2$

Метод K ближайших соседей

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

Метрика расстояния: $\rho(X_1, X_2) = \sum (F_i^{X_1} - F_i^{X_2})^2$

$\rho(X_1, X_2) = 1289$

$\rho(X_2, X_3) = 39$

Метод K ближайших соседей

	F1	F2	F3	F4	F5	Y
	площадь	удаленность от центра, км	год постройки	лет после ремонта	тип постройки	цена квартиры
X1	25	3	2005	1	1	
X2	55	10	1987	5	2	
X3	50	12	1990	6	5	

Метрика расстояния: $\rho(X_1, X_2) =$?

Метод K ближайших соседей

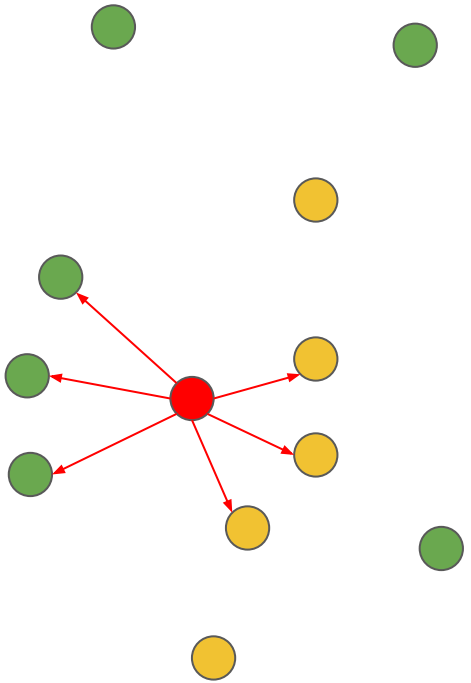
	F1	F2	F3	F4	F5	Y
	площадь	удаленность от центра, км	год постройки	лет после ремонта	тип постройки	цена квартиры
X1	25	3	2005	1	1	
X2	55	10	1987	5	2	
X3	50	12	1990	6	5	

Метрика расстояния: $\rho(X_1, X_2) = \sum_{i=0}^{4} (F_i^{X_1} - F_i^{X_2})^2 + (F_5^{X_1} == F_5^{X_2})$

Метод K ближайших соседей

- У нас в выборке есть объекты X1, X2, …
- Приходит новый объект X
- Сортируем объекты X1, X2, … по расстоянию до X
- Выбираем k ближайших объектов к X
- Определяем ответ для X как среднее значение ответов для k ближайших элементов

  Для классификации: класс, который представлен среди k ближайших элементов чаще всего

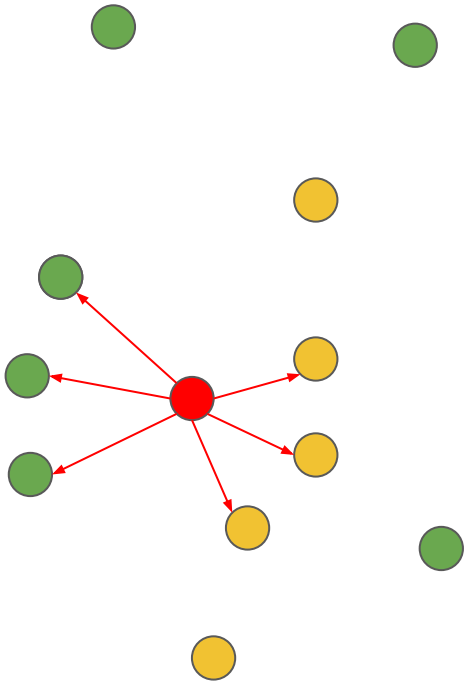


Метод K ближайших соседей

- У нас в выборке есть объекты X1, X2, …
- Приходит новый объект X
- Сортируем объекты X1, X2, … по расстоянию до X
- Выбираем k ближайших объектов к X
- Определяем ответ для X как среднее значение ответов для k ближайших элементов

Параметры алгоритма:

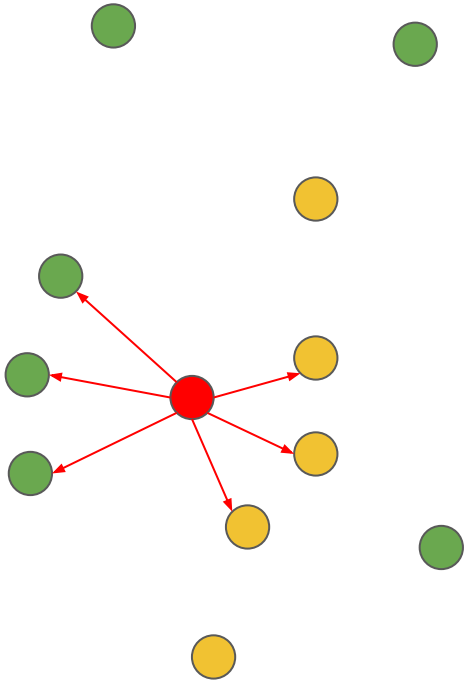
- Число k ближайших соседей
- Функция расстояния между объектами



Метод K ближайших соседей

Свойства:

- **Интерпретируемый**: мы можем понять, почему модель для объекта X выдала тот или иной результат, предъявив похожие на X объекты обучающей выборки
- Требует 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



Компоненты постановки задачи анализа данных

1. Что хотим прогнозировать? Что является входом, а что выходом?
2. Какие доступны данные?
3. Из какого класса будем выбирать модель?
4. Что является критерием качества решения?

Компоненты постановки задачи анализа данных

1. Что хотим прогнозировать? Что является входом, а что выходом?
2. Какие доступны данные?
3. Из какого класса будем выбирать модель?
4. Что является критерием качества решения?

1. Прогнозируем цену квартиры по площади
2. Доступны данные за прошлый год
3. Будем строить линейную модель
4. Хотим минимизировать квадратичную функцию ошибки

Метрики

Измерение качества модели

Чтобы понять, насколько адекватно ведет себя модель, нужно каким-то образом численно оценить ее качество.

Метрика - это функция вида:

$$metric(\mathbf{y}, \hat{\mathbf{y}})$$

где $\mathbf{y}$ - это правильное значение целевой переменной (**label**),

а $\hat{\mathbf{y}}$ - значение, предсказанное моделью (**prediction**).

Примеры метрик

Классификации:

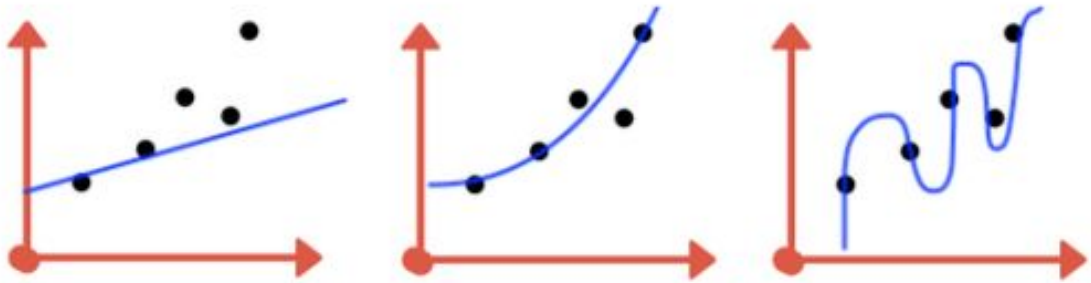
- **accuracy** - процент правильных предсказаний среди всех примеров
- precision - точность
- recall - полнота
- f1 - объединяет полноту и точность
- ROC-AUC - вероятность правильного ранжирования двух случайных примеров

Регрессии:

- MSE - средний квадрат отклонения
- RMSE - стандартное отклонение
- MAE - средний модуль отклонения
- MAPE - mean absolute percentage error
- R2 - коэффициент детерминации

Более подробно метрики будут рассмотрены после практического занятия

Смещенная оценка



Вопрос: какое предсказание лучше по метрикам, а какое на самом деле?

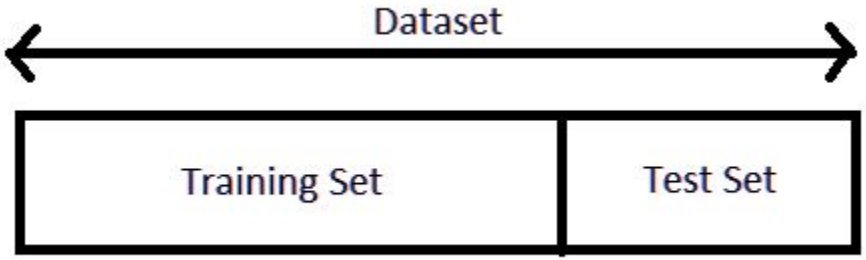
Если тестировать модель на той же выборке, на которой она обучалась, то оценка получится смещенной. В таком случае “самая лучшая” модель - это та, которая просто запомнила все данные.

Хорошая модель должна делать хорошие предсказания на **новых** для себя

Отложенная выборка

Можно “отложить”, скажем, 20% обучающей выборки для валидации модели. Использовать 80% выборки для обучения и 20% для тестирования.

- Оценка на тестовой выборке будет несмещенной
- Тестовая выборка маленькая - оценка будет иметь погрешность



Кросс-валидация

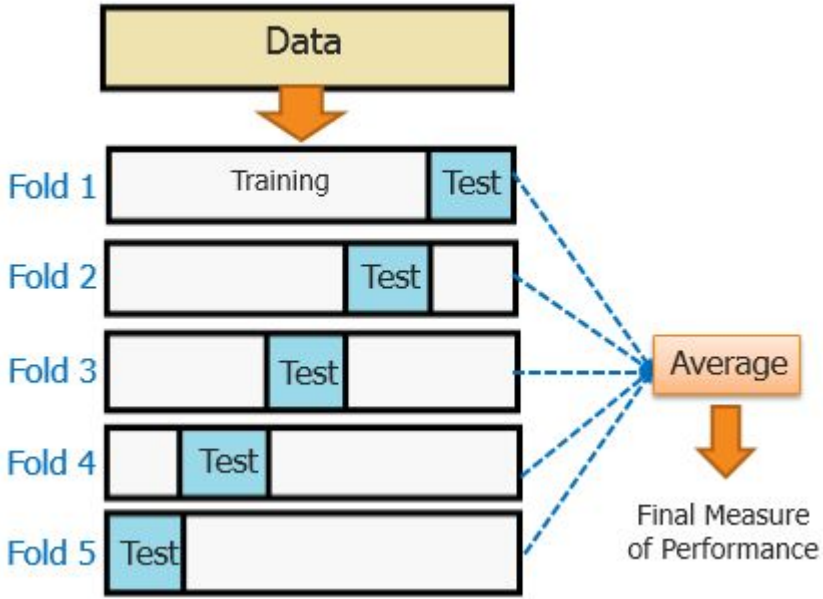
- Разбиваем выборку на k частей
- k-1 частей используются для обучения и одна - для тестирования
- Процесс повторяется k раз. Каждый раз для тестирования выбирается разная часть
- Результаты тестирования усредняются

Плюсы:

- Погрешность оценки уменьшается, т.к. используется весь набор

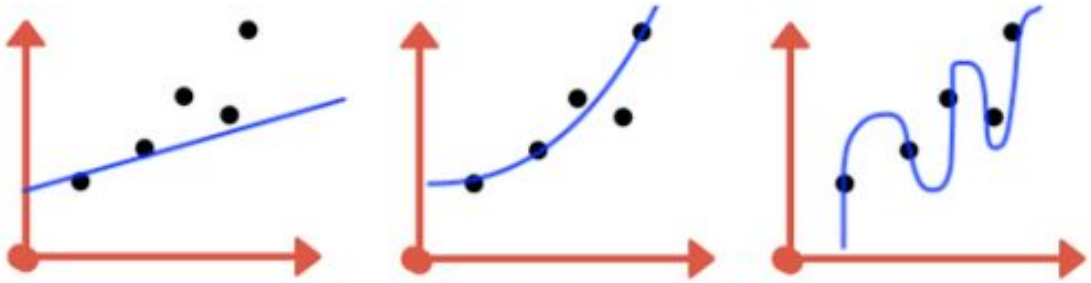
Минусы:

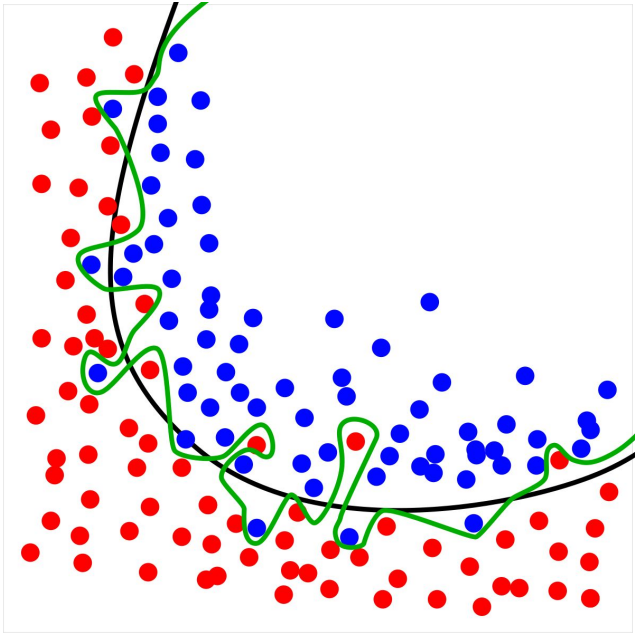
- Обучение производится k раз. Для некоторых моделей это может быть очень долго



Переобучение и недообучение

Переобучение и недообучение





Слишком простая “глупая” модель может уловить только общую закономерность, без деталей. Это называется **недообучение**.

Слишком “умная”, сложная модель может просто запомнить все точки обучающей выборки - это называется **переобучение**.

Для каждой задачи нужно найти свою **оптимальную**

Сложность модели

Сложность модели регулируется внешними параметрами.

Например, в KNearestNeighbors сложность регулируется параметром K, а в линейной регрессии количеством признаков

Какой K самый лучший? Подскажет кросс-валидация