

Классическое машинное обучение

Введение. Основные задачи. Линейная регрессия

Что такое машинное обучение


- Машинное обучение – набор способов воспроизведения связей между событиями и результатом.
- Машинное обучение – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться.
- Machine learning – the field of study that gives computers the ability to learn without being explicitly programmed.

Пример: задача о квартирах

У нас есть:

- база данных по уже проданным квартирам с их характеристиками и ценами
- хотим оценить новую квартиру, о которой известны ее характеристики
- цена должна быть адекватной: не слишком низкой и не слишком высокой

Новостройка от застройщика






1-комн. кв., 39,6 м², 4/35 этаж
Секция 1 • Сдача корпуса 2 кв. 2026

[Только на Циан](#)

ЖК «Nagatino i-Land»
сдача ГК: 2 кв. 2026 года
Технопарк • 6 минут пешком
Москва, ЮАО, р-н Даниловский, м. Технопарк, Нагатино Ай-Ленд ЖК


15 804 901 Р
399 114 Р/м²

Nagatino i-Land - масштабный жилой квартал бизнес-класса от девелопера "Группа "Эталон" и британского архитектурного бюро AHR. Nagatino i-Land - уникальный жилой остров, который предлагает своим жителям соседство с Москвой-рекой, ее благоустроенными набережными, новыми причалами и мостами. Те, кто любит прогулки, найдут рядом сразу...



Еще фото

Новостройка от застройщика






1-комн. кв., 65,1 м², 3/8 этаж
Секция 1 • Сдача корпуса 2 кв. 2024

ЖК «Дом Лаврушинский»
сдача ГК: 2 кв. 2024 года
Третьяковская • 5 минут пешком
Москва, ЦАО, р-н Якиманка, м. Третьяковская, Лаврушинский ЖК

154 070 000 Р
2 366 667 Р/м²


Квартира с одной спальней, обращенная окнами в зелёный благоустроенный двор-сад. К кухне-гостиной, которую лучше расположить сразу у входа, примыкает балкон, откуда особенно приятно любоваться оригинальным придомовым ландшафтом с вечерней подсветкой.



Еще фото

Новостройка от застройщика

RIVER PARK TOWERS
КУТУЗОВСКИЙ



Площадь
42,8 м²
Потолки
3,1 м
Этаж
2
Отделка
нет
Вид
на Москву-реку и Москва-Сити

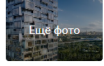


1-комн. кв., 42,8 м², 2/45 этаж
Сдача корпуса 3 кв. 2026

[Только на Циан](#)

ЖК «River Park Towers Кутузовский»
Фили • 5 минут пешком
Москва, ЗАО, р-н Дорогомилово, м. Фили, Кутузовский проезд, 16А/1

21 808 632 Р
509 547 Р/м²

В продаже квартира с 1 спальней площадью 42,80 кв. м на 2 этаже в жилом комплексе премиум-класса River Park Towers Кутузовский в Башне Amber Tower. River Park Towers Кутузовский строящийся жилой комплекс небоскребов премиум-класса на берегу Москвы-реки, с потрясающими видами на Москва-Сити и Парк Победы. В комплексе будет создана...



Еще фото

Формализация задачи

X – множество объектов

Y – множество ответов

$a: X \rightarrow Y$ – неизвестная зависимость


Дано:

$\{x_1, x_2, \dots, x_n\} \in X$ – обучающая выборка

$\{y_1, y_2, \dots, y_n\} \in Y$ – известные ответы

Найти:

$a: X \rightarrow Y$ – алгоритм (решающую функцию),
приближающую y на всем множестве X



Новостройка от застройщика

1-комн. кв., 39,6 м², 4/35 этаж
Секция 1 • Сдача корпуса 2 кв. 2026


[Только на Циан](#)

ЖК «Nagatino i-Land»
сдача ГК: 2 кв. 2026 года
Технопарк • 6 минут пешком
Москва, ЮАО, р-н Даниловский, м. Технопарк, Нагатино Ай-Ленд ЖК

15 804 901 Р
399 114 Р/м²

1/28

Еще фото




Новостройка от застройщика

1-комн. кв., 65,1 м², 3/8 этаж
Секция 1 • Сдача корпуса 2 кв. 2024

ЖК «Дом Лаврушинский»
сдача ГК: 2 кв. 2024 года
Третьяковская • 5 минут пешком
Москва, ЦАО, р-н Якиманка, м. Третьяковская, Лаврушинский ЖК

154 070 000 Р
2 366 667 Р/м²

Еще фото



Новостройка от застройщика

1-комн. кв., 42,8 м², 2/45 этаж
Сдача корпуса 3 кв. 2026

[Только на Циан](#)

ЖК «River Park Towers Кутузовский»
Фили • 5 минут пешком
Москва, ЗАО, р-н Дорогомилово, м. Фили, Кутузовский проезд, 16А/1

21 808 632 Р
509 547 Р/м²

Еще фото

Признаковое описание объектов

- Признаки объекта x можно записать в виде вектора

$$(f_1(x), f_2(x), \dots, f_n(x))$$

- Матрица объекты-признаки:

$$\begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_m) & \cdots & f_n(x_m) \end{pmatrix}$$

Стандартная постановка задач машинного обучения

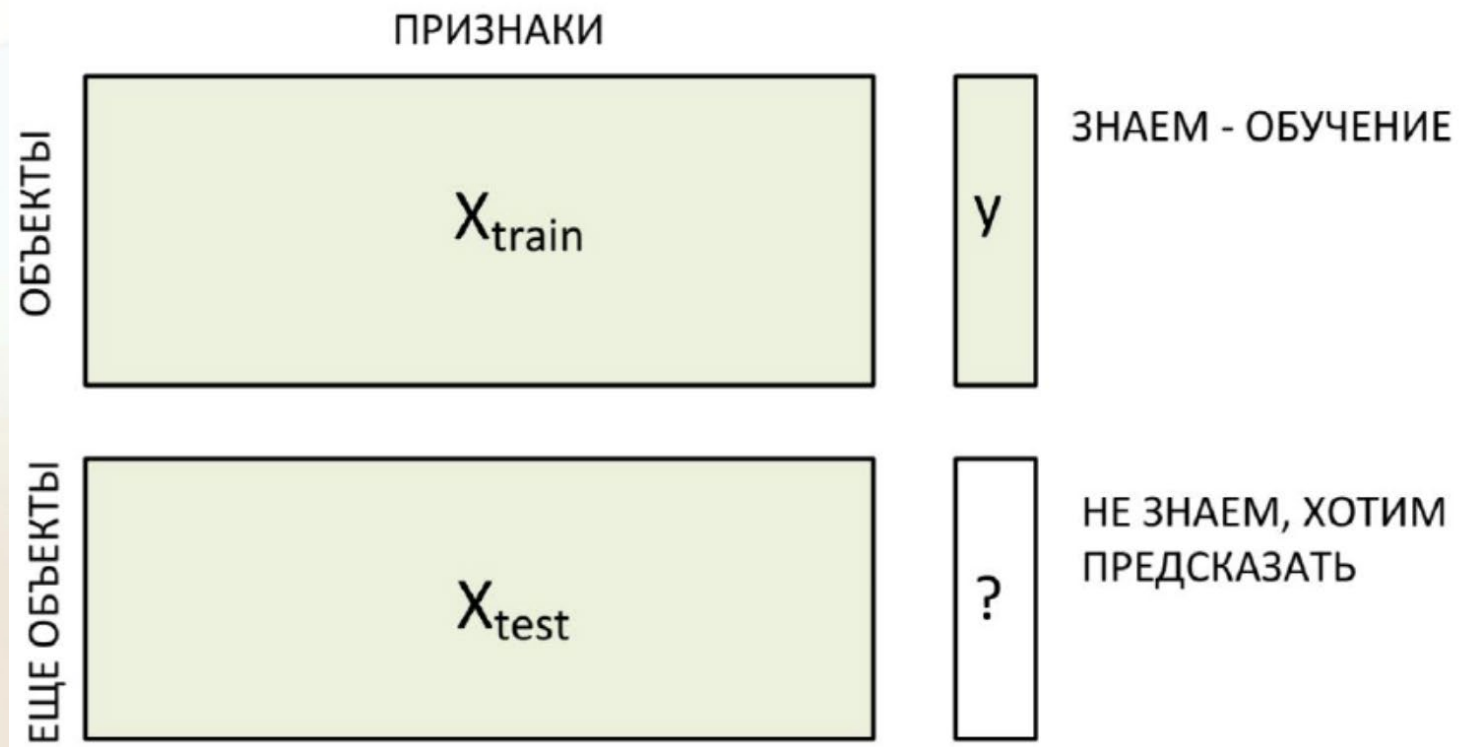


Схема получения предсказания

В задачах обучения с известными классами (обучение по прецедентам) всегда есть два этапа:

- Этап обучения (training): по выборке $X = \{(x_i, y_i)\}$ строим алгоритм a
- Этап применения (testing): алгоритм a для новых объектов выдает ответы $a(x)$

Определения

- **Признаки, факторы (features)** – количественные характеристики объекта
- **Обучающая выборка (training set)** – конечный набор объектов, для которых известны значения целевой переменной
- **Объекты** – абстрактные сущности (но компьютеры работают только с числами)
- **Признаки** описывают объекты с помощью чисел. Для их формирования обычно нужны эксперты (в нашем случае – профессиональные лингвисты)

Виды признаков

- Числовые
- Бинарные (0/1)
- Категориальные (название города, марка машины)
- Признаки со сложной внутренней структурой (изображение, текст)

Типы задач в зависимости от целевой переменной

- Классификация:
 - $Y = \{0,1\}$ - бинарная классификация
 - $Y = \{1, \dots, M\}$ - мультиклассовая (мультиномиальная) классификация
- Регрессия
 - $Y = R$ или $Y = R^n$
- Кластеризация
 - Целевая переменная отсутствует, нужно поделить данные по схожести
- Понижение размерности, визуализация данных

Виды обучения

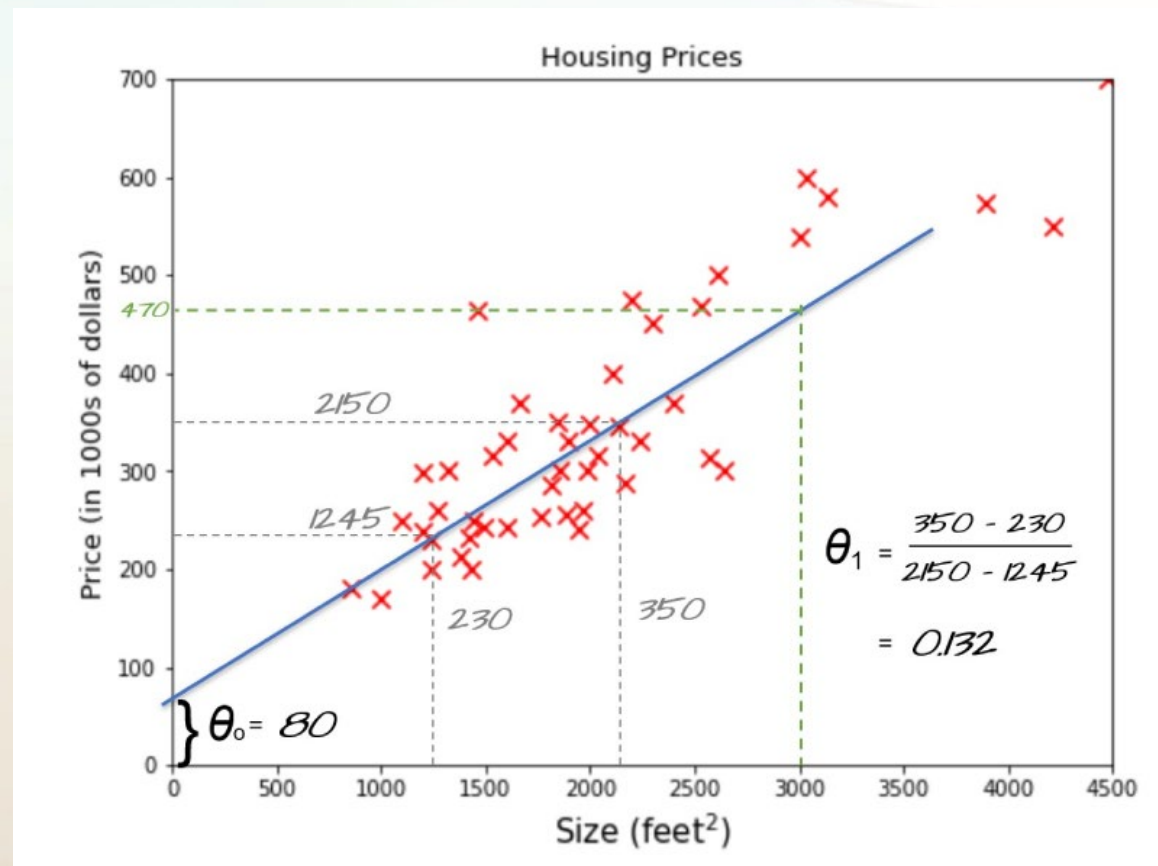
- Обучение с учителем (supervised):
если нам известны значения целевой переменной
- Обучение без учителя (unsupervised):
если значений целевой переменной нет
- Обучение с подкреплением (semi-supervised):
если невозможно заранее задать все правильные значения целевой переменной (обучаем робот-пылесос, программу - игрока в го и т.п.)

Линейная регрессия

Основные понятия машинного обучения. Обучающая и валидационная выборка. Целевая переменная. Метрики, оценка качества. Функционал ошибки. Градиентный спуск.

Задача линейной регрессии

- Наши **признаки**:
 - площадь квартиры
 - расстояние до метро
 - этаж
 - ...
- Наша **целевая переменная**:
 - цена квартиры
- Наша задача:
 - построить прямую так, чтобы для наших x ее y был максимально похож на правду



Линейная регрессия: веса

- Цель: подобрать такие коэффициенты уравнения прямой, чтобы по нашим признакам можно было угадать примерный ответ (целевую переменную):

$$w_1x_1 + w_2x_2 + \dots + w_0 = y$$

- x_1, x_2, \dots - это наши признаки (площадь квартиры, время до метро...)
- y – это целевая переменная (цена квартиры)
- w_1, w_2, \dots - это веса, или коэффициенты
- w_0 - это свободный коэффициент (шум)

Как будем учить?

1. Возьмем случайные веса
2. Посчитаем предсказанные игреки для всех известных объектов
3. Сравним с правильными ответами
4. Поправим веса, чтобы наши игреки стремились к правильным ответам
5. Вернемся к пункту 2
6. ???
7. PROFIT!

Как сравнивать игреки?

- Очевидно, нужно узнать, на сколько в среднем ошибается алгоритм, то есть:

$$\frac{1}{n} \sum_{i=1}^n (y_{pred} - y_{true})$$

- Что в такой формуле не нравится?

Метрики оценки качества

- MSE: $\frac{1}{n} \sum_{i=1}^n (y_{pred} - y_{true})^2$ и RMSE: $\sqrt{\frac{1}{n} \sum_{i=1}^n (y_{pred} - y_{true})^2}$
- MAE: $\frac{1}{n} \sum_{i=1}^n |y_{pred} - y_{true}|$
- Коэффициент детерминации: $R^2(a, X) = 1 - \frac{\sum_{i=1}^l (a(x_i) - y_i)^2}{\sum_{i=1}^l (y_i - \hat{y})^2}$
- MSLE
- MAPE
- SMAPE
- ...

Как подобрать веса?

- Возьмем, например, MSE: очевидно, что мы хотим, чтобы он был поменьше (чем меньше MSE, тем меньше ошибка модели)
- Следовательно, нам нужно **минимизировать функцию ошибки**
- То есть, уравнение, которое нам нужно решить (в матричной форме):

$$\frac{1}{n} ||X_w - y||^2 \rightarrow \min_w$$

- Это называется метод наименьших квадратов

Аналитическое решение МНК

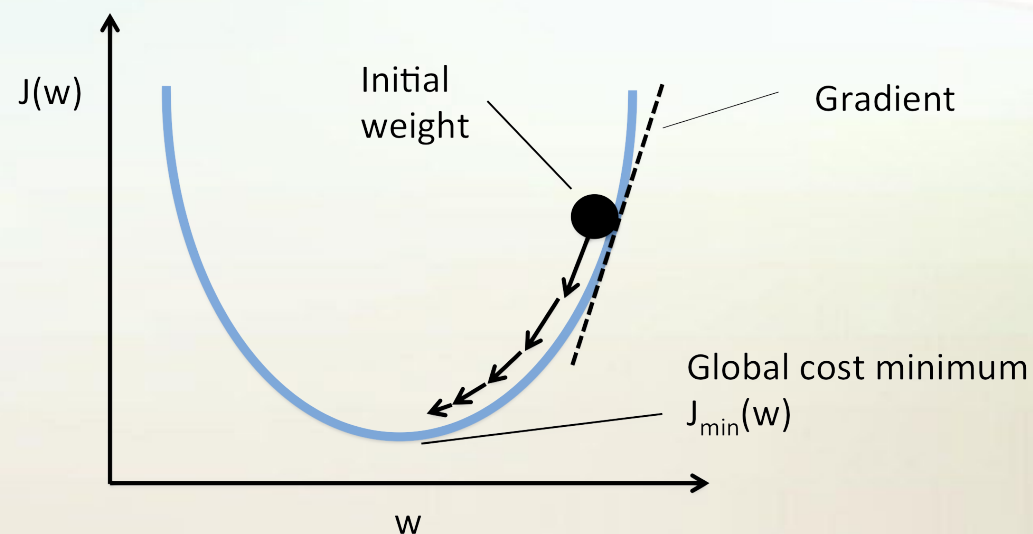
$$w = (X^T X)^{-1} X^T y$$

Недостатки:

- Обращение матрицы – сложная операция ($O(n^3)$ от числа признаков)
- Матрица $X^T X$ может быть вырожденной или плохо обусловленной
- Если функционал ошибки будет другим, можем вообще не решить задачу

Градиентный спуск

- Градиент – вектор, в направлении которого функция растет
- Антиградиент – вектор, противоположный градиенту
- Если будем двигаться в направлении антиградиента, найдем минимум
- *(Вспоминайте Лагутина)*



Градиентный спуск

- Пусть у нас только один вес w (для простоты)
- Инициализируем вес случайным числом: $w^{(0)}$
- При добавлении к весу антиградиента $-\frac{\partial Q}{\partial w}$ функция $Q(w)$ убывает.
- Вычисляем конкретные значения производной для каждого объекта в выборке
- Считаем среднее арифметическое из них
- Вычитаем из веса $w^{(0)}$
- Повторяем с начала

Градиентный спуск

- Если у нас несколько весов, то делаем это для каждого из них.
- Общая формула изменения веса:

$$w^{(k)} = w^{(k-1)} - \nabla Q(w^{(k-1)})$$

- Обычно еще добавляют коэффициент, чтобы сразу весь градиент не вычитался:

$$w^{(k)} = w^{(k-1)} - \eta \nabla Q(w^{(k-1)})$$

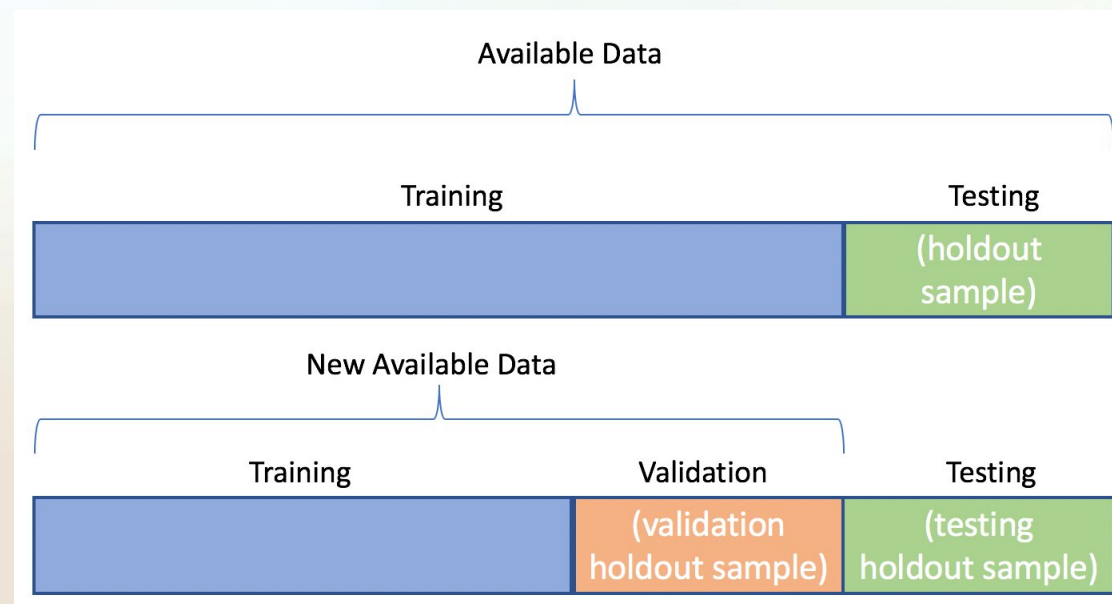
- Этот параметр называется learning rate. Мы еще много будем про него говорить на курсе по нейронкам

Как проверить качество?

- Допустим, мы обучили наш алгоритм. Как удостовериться, что он хорошо работает?
- Очевидно, считаем все те же метрики
- Метрика качества может быть такая же, как функция ошибки, а может быть другой
- Но нельзя ее считать на той выборке, на которой мы учились: это будет нечестно

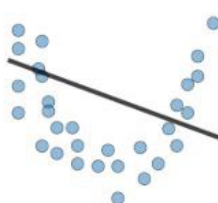


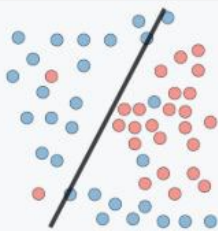
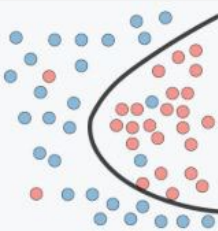
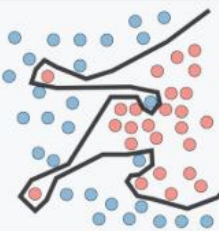

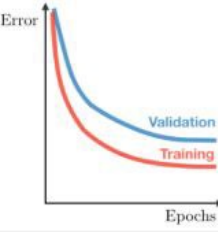
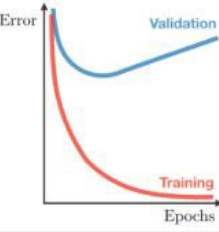
Как проверить качество?

- Следовательно, перед обучением нужно отложить какое-то количество данных, чтобы модель их не видела



Переобучение и недообучение

- В алгоритмах классического МО очень важна работа с признаками
- Если признаки линейно зависимы между собой, то высок риск переобучения
- Работа с фичами – это искусство

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none">• High training error• Training error close to test error• High bias	<ul style="list-style-type: none">• Training error slightly lower than test error	<ul style="list-style-type: none">• Very low training error• Training error much lower than test error• High variance
Regression illustration			
Classification illustration			
Deep learning illustration			
Possible remedies	<ul style="list-style-type: none">• Complexify model• Add more features• Train longer		<ul style="list-style-type: none">• Perform regularization• Get more data

Наконец - практическая часть!

устанавливаем scikit learn, если еще не!

```
pip install scikit-learn
```

```
(conda install -c anaconda scikit-learn)
```