Лекция 8

Введение в машинное обучение

Курс: Введение в DS на УБ и МиРА (весна, 2022)
Преподаватель: Владимир Омелюсик

16 мая 2022 г.

Что такое машинное обучение?

Неформальное определение
Машинное обучение – дисциплина, изучающая построение моделей, позволяющих компьютерам воспроизводить зависимости между разными объектами без их непосредственного программирования.

- Есть конечная выборка, на которой обучаем модель.
- В ходе обучения происходит «запоминание» зависимостей.
- После обучения модель способна давать «хорошие» предсказания на новых данных.

Зависимости

- Иногда можно получить явный математический вид.
- Иногда нет.
  - Какая завтра погода?
  - Какая тональность у текста?
  - На фотографии кошка или собака?

тонн -> кг

Зав: 1 т = 1000 кг

$f(x) = \frac{x}{1000}$, x - масса в т.

- Найти точные математические функции для ответа на эти вопросы сложно или невозможно. Но если у нас есть некоторый набор данных, то можно попытаться приблизить истинные зависимости некоторыми математическими моделями.
- Статистика – про объяснения, машинное обучение – про предсказания.

Основные понятия

	x_i^1	.....	x_i^d	
x_1	1.01	"red"	1	...
x_2	2.01	"blue"	1	...
x_3	7.15	"green"	0	...

- Наблюдение: $x_i = (x_i^1 \dots x_i^d)$ (1 объект) $X_{N \times d}$
- Признак: $x_i^1, \dots, x_i^d$
- Целевая переменная: $y_{N \times 1}$ target
- Модель: $a(x_i)$ – выдает прогноз на набл. x_i
- Параметры: Хар-и, к. подбир. при обучении
- Гиперпараметры: Хар-и, к. устанавл. извне и не уч. при обучении
- Обучающая выборка: (X_{train}, y_{train})
- Тестовая выборка: (X_{test}, y_{test})
- Функция потерь: ф-я, к. минимиз. при обучении
- Метрика качества: ф-я, по кот. смотрим кач-о

Пример: предсказание стоимости квартиры

X_{train} :

# комн.	кол-во ком	S кв.	расстоян. до бл. метро	y_{train} (у.е.)
1	3	100	100	10 000
2	2	50	550	6000
⋮	⋮	⋮	⋮	
10 000	1	35	80	3000

y_i	x_1	x_2
10000	8000	0.03

β_1 – мал. ≈ 1

β_2 – больш. ≈ 1000

$$a(x_i) = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{кол-в ком.} + \hat{\beta}_2 \, S_{кв} + \hat{\beta}_3 \, \text{расст-е}$$

1) Обуч. $a(x_i)$ — находим $\hat{\beta}_0 \dots \hat{\beta}_3$ на X_{train}, y_{train}

$y_{train} = a(x_i)$

$$\hat{\beta} = (X_{train}^T X_{train})^{-1} X_{Train}^T y_{train}$$

MSE → min (ф. потерь)

2) X_{test}, y_{test}

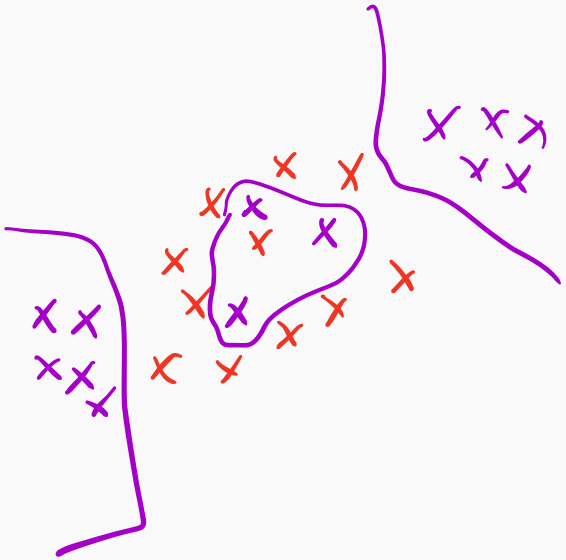
└ кол-во к., $S_{кв}$, расст.

$a(X_{test}) \to \hat{y}_{test}$

Метрики кач-а: MSE, MAE, R^2 (y_{test}, $\hat{y}_{test}$)

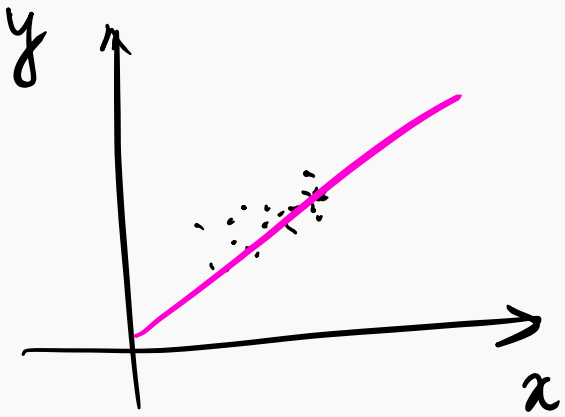
Виды задач в машинном обучении

1. Обучение с учителем. (есть y) X_{train}, y_{train}
   - Регрессия.
   - Классификация. — y прин. конеч. число зн.
     - Бинарная.
     - Многоклассовая.
     - С пересекающимися классами.
   - Ранжирование.
2. Обучение без учителя. (нет y)
   - Кластеризация.
   - Понижение размерности.
   - Визуализация.



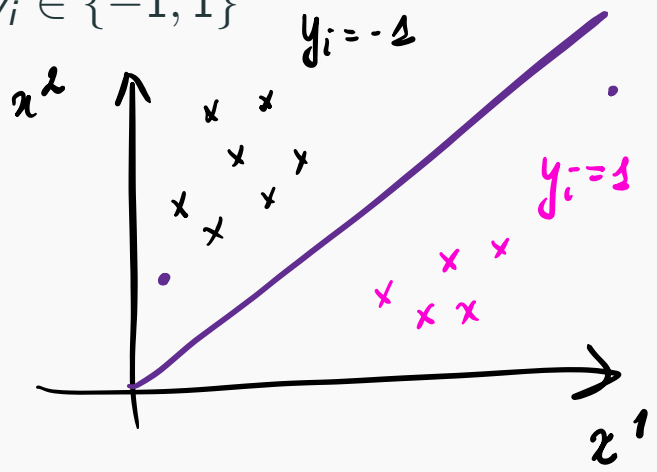
Задача регрессии

- $y_i \in \mathbb{R}$ (прогноз. стоимость квартиры)



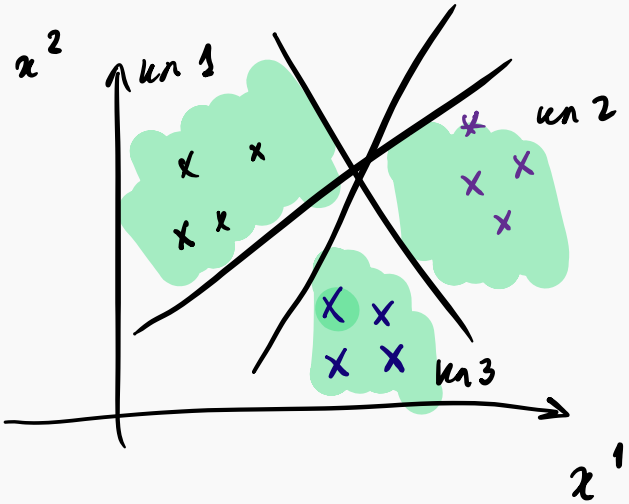
Задача классификации (бинарная)

- $y_i \in \{-1, 1\}$



Задача классификации (многоклассовая)

- $y_i \in \{1, \ldots, K\}$



Задача классификации (с пересекающимися классами)

- $y_i \in 0, 1^K$
- Ответ – вектор из нулей и единиц длины K.
- Единица на позиции i означает, что объект принадлежит к классу i.

Задача ранжирования

- Есть набор документов $d_1, \ldots, d_N$ и некоторый запрос q.
- Хотим сортировать документы в соответствии с релевантностью запросу.
- Алгоритм должен выдавать оценку релевантности.

Задача кластеризации

- Есть только X, а y отсутствует.
- Хотим найти группы «похожих» объектов в X, используя только характеристики X.
- Как определить «похожесть»? Как оценить качество? Как выбрать число групп?

Задача понижения размерности

- X имеет размеры $N \times d$, где d – очень большое.
- Пример: медицинские измерения.
- Проблемы:
  - Модели долго обучаются.
  - Некоторые модели могут неправильно обучиться.
- Раешение – построить алгоритм, который на основании выборки X построит новую выборку с меньшим числом признаков.

Задача визуализации

- Частный случай задачи понижения размерности, где новая матрица состоит из 2 или 3 признаков.