Линейная регрессия

Семинар 2

БАЗА

Прогнозирование цен на недвижимость

Подбор персональных рекомендаций музыки

Анализ эмоциональной окраски текста

Поиск интересующего объекта на фото

Перевод текста с русского языка на английский

Определение сегментов клиентов для создания рекламных кампаний

Если уже есть какое-то правило/функция, которая описывает объект и решает данную проблему, то машинное обучение здесь не нужно

Восстановить сложные зависимости по конечному числу примеров!

БАЗА

- **Регрессия** – множество ответов вещественное число

$$Y = R$$

Определить какой процент рекомендованной книги прочитает пользователь.

- **Классификация** – множество ответов конечное, порядка на данном множестве нет

$$Y = \{1, \ldots, k\}$$

Определить жанр книги (единственным образом)

- **Бинарная классификация** – есть два ответа

$$Y = \{-1, +1\}$$

Определить прочитает ли пользователь данную книгу или нет.

- **Multi-label classification** – многоклассовая классификация с пересекающимися классами

$$Y = \{0, 1\}^k$$

Определить жанр книги.

БАЗА

Задача: Хотим предсказать какую оценку за МО-1 получит каждый студент.

Объект (то для чего мы делаем прогноз): Студент

Целевая переменная, таргет (то что предсказываем): Оценка за МО-1

БАЗА

Сущность	**Обозначение**
Объект	x
Целевая переменная, таргет	y
Множество объектов	X
Множество ответов, таргетов	Y

БАЗА

Как мы сделаем прогноз по объекту?
Что такое признаки, фичи объекта?

Признаки – характеристики, описание объекта
Пример:
Объект - Студент
Признаки, фичи:
- Пол
- Средняя оценка за прошлый семестр
- Оценки в школе
- Опыт в МЛ
- И т.д.

БАЗА

Типы признаков:

- Бинарные ϵ {0, 1}
- Числовые ϵ R
- Категориальные – значения из множества классов
- Порядковые

БАЗА

Признаки объекта x можно записать в виде вектора

$$(x_1, \dots, x_d)$$

Матрица “объекты-признаки”:

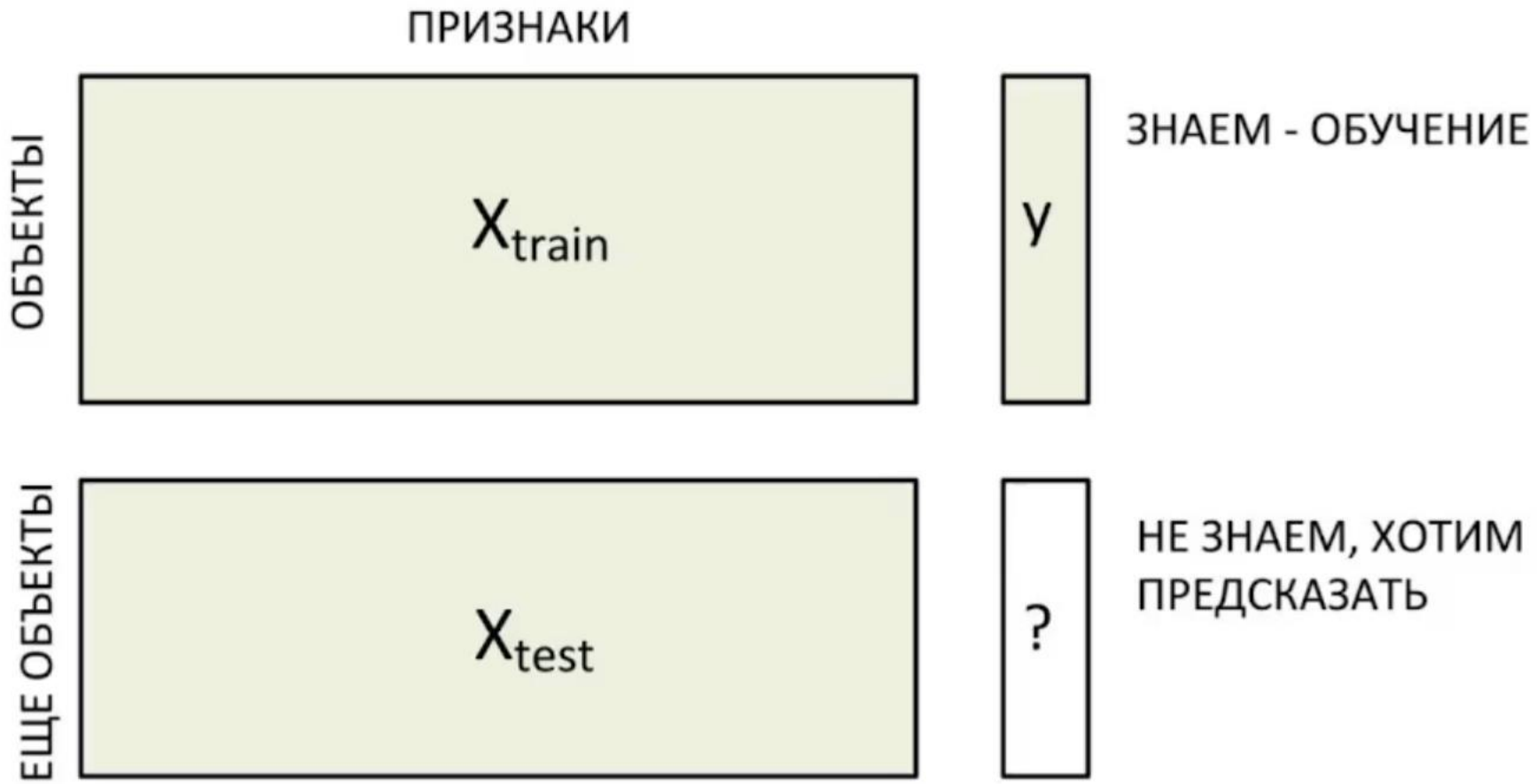
$$\begin{pmatrix} x_1^1 & \cdots & x_d^1 \\ \vdots & \ddots & \vdots \\ x_1^l & \cdots & x_d^l \end{pmatrix}$$

Признаки первого объекта

Значение признака d

Движемся к модели

Движемся к модели



Линейная модель

Пример: линейная модель

$$a(x) = w_0 + w_1 x_1 + \ldots + w_d x_d$$

a – модель

$\boldsymbol{w_i}$ - веса модели

$\boldsymbol{x_i}$ - признаки, фичи

d – кол-во признаков, фичей

Обучение модели

Подбор оптимальных весов модели, при которых качество будет наилучшим.

Как определить качество?

Обучение модели

Функция потерь

Модель сделала предсказание.

Чем больше значение функции потерь, тем сильнее модель ошиблась в своем предсказании.

Функция потерь оценивает качество предсказания на одном объекте.

Какие вам известны функции потерь для задачи регрессии?

* Мб напишу на доске

Обучение модели

Функционал ошибки – для выборки.
При обучении модели мы минимизируем функционал ошибки.

$$Q(a, X) = \frac{1}{l}\sum_{i=1}^{l}(a(x_i) - y_i)^2 \to \boldsymbol{min}$$

Функция потерь оценивает качество предсказания на одном объекте.

* Мб напишу на доске

Переобучение

Переобучение - это явление, при котором модель слишком точно адаптируется к обучающим данным, теряя при этом способность к обобщению на новых данных.

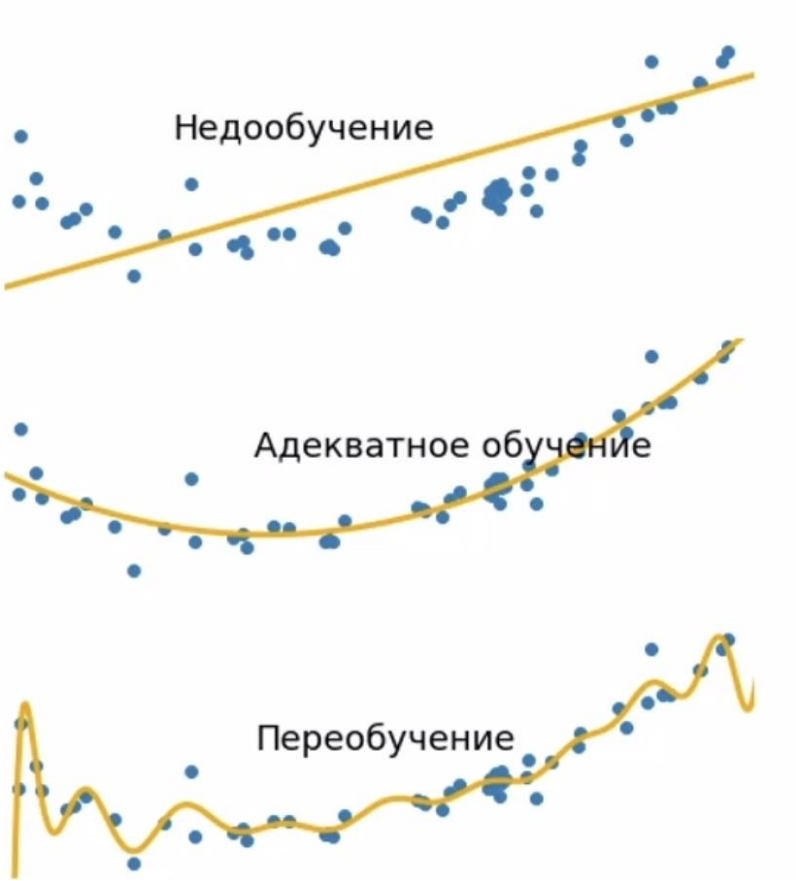
Отследить переобучение можно, сравнивая качество модели на обучающем наборе данных и на валидационном/тестовом наборе.

Разрыв в производительности указывает на переобучение.

Переобучение

Признаки переобучения:

- Ошибка на тесте сильно выше, чем на трейне
- Большие веса при признаках



Переобучение

Виды регуляризации

$$Q(w,X)+\lambda\sum_{i=1}^{d}|w_i|\to\min_w$$

L1 – регуляризация, Lasso

$$Q(w,X)+\lambda\sum_{i=1}^{d}w_i^2\to\min_w$$

L2 – регуляризация, Ridge

$$Q(w,X)+\lambda_1\sum_{i=1}^{d}|w_i|+\lambda_2\sum_{i=1}^{d}w_i^2\to\min_w$$

Elastic Net