Машинное обучение

День 3. Регрессия.

ПЛАН

День 3. Регрессия.

01 ОСНОВЫ 10 МИН

Дополнительные понятия и проблемы. Масштабирование признаков, метрики регрессии.

02 МЕТОДЫ 15 МИН

Методы регрессии: линейные методы, регуляризация.

03 ПЛАН РЕШЕНИЯ ML-ЗАДАЧ 5 МИН

Процесс решения ML-задачи: второе приближение.

04 ПРАКТИКА 10 МИН

Решим задачу регрессии.

01 ОСНОВЫ 10 МИН Постановка задачи и scaling

X – множество объектов / features
Y – целевое значение / target
f – решающая функция или алгоритм ML

$$f(X) = Y$$

X =

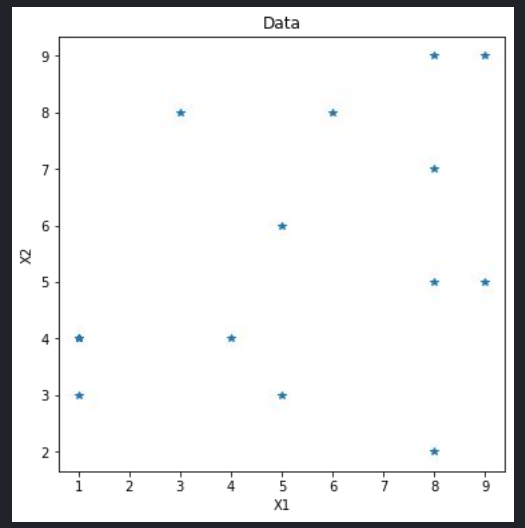
F1	F2	F3	F4	F5
14.3	4	type1	127	2.8
13.35	5	type2	100	2.65
12	1	type1	101	2.8
17.1	4	type1	113	3.85

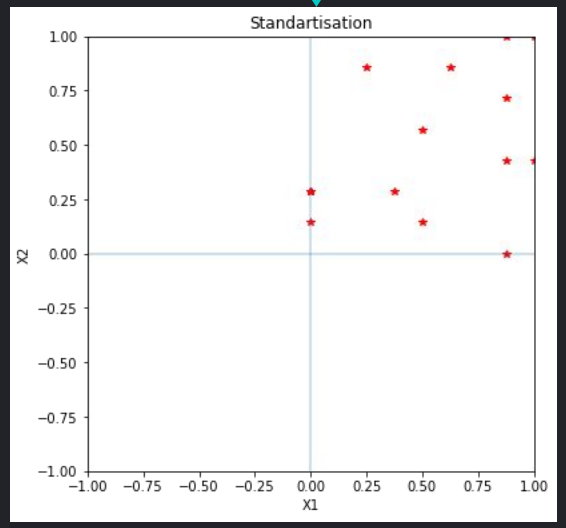
Y =

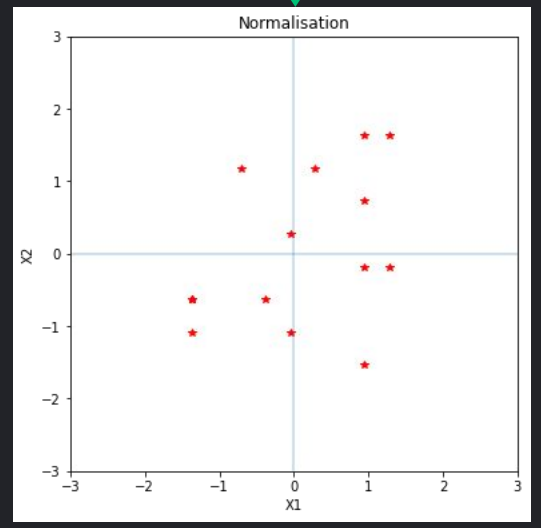
class
12.5
-7.3
-14
26

Масштабирование признаков:

- стандартизация $X \in [0; 1]$ $\frac{X - X_{min}}{X_{max} - X_{min}}$
- нормализация $X \sim D(0,1)$ $\frac{X - \mu}{\sigma}$







01 ОСНОВЫ 10 МИН Метрики

Метрика

- Метрика – это число.
- Это показатель того, насколько хорошо работает наш алгоритм и какая у него обобщающая способность.
- Метрика считается только на тестовой выборке.

Максимальная ошибка | MaxE

Max Error

$$MaxError = \max(|Y_{true} - Y_{pred}|)$$

Максимальная ошибка в % | MaxPE

Max Percentage Error

$$MaxError = \max\left(\left|\frac{Y_{true} - Y_{pred}}{\max(\epsilon, Y_{true})}\right|\right)$$

Средняя абсолютная ошибка | MAE

Mean Absolut Error

$$MAE = \frac{1}{N} \cdot \sum |Y_{true} - Y_{pred}|$$

Средняя абсолютная ошибка в % | MAPE

Mean Absolut Percentage Error

$$MAPE = \frac{1}{N} \cdot \sum \left|\frac{Y_{true} - Y_{pred}}{\max(\epsilon, Y_{true})}\right|$$

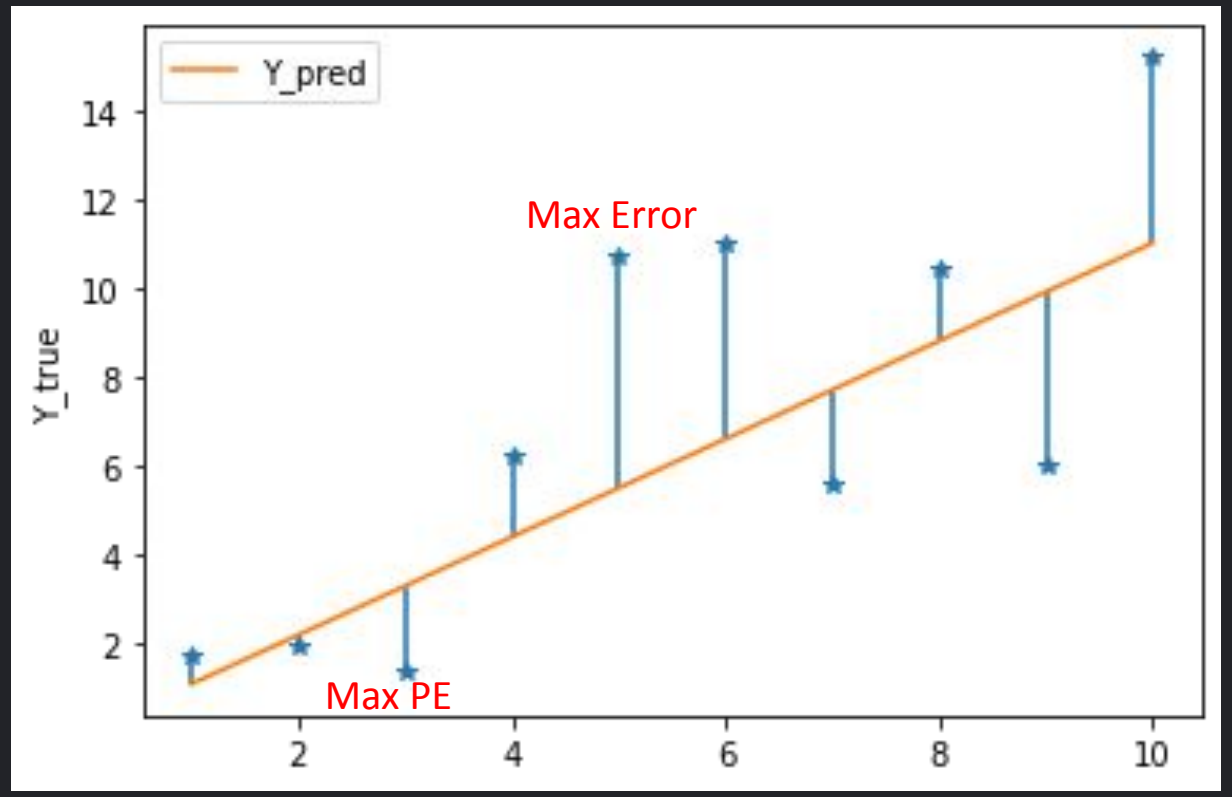
Среднеквадратичная ошибка | MSE

Mean Squared Error

$$MSE = \frac{1}{N} \cdot \sum (Y_{true} - Y_{pred})^2$$

01 ОСНОВЫ 10 МИН

Метрики



Max E = ***5.22***

Max PE = **1.41**

MAE = ***2.59***

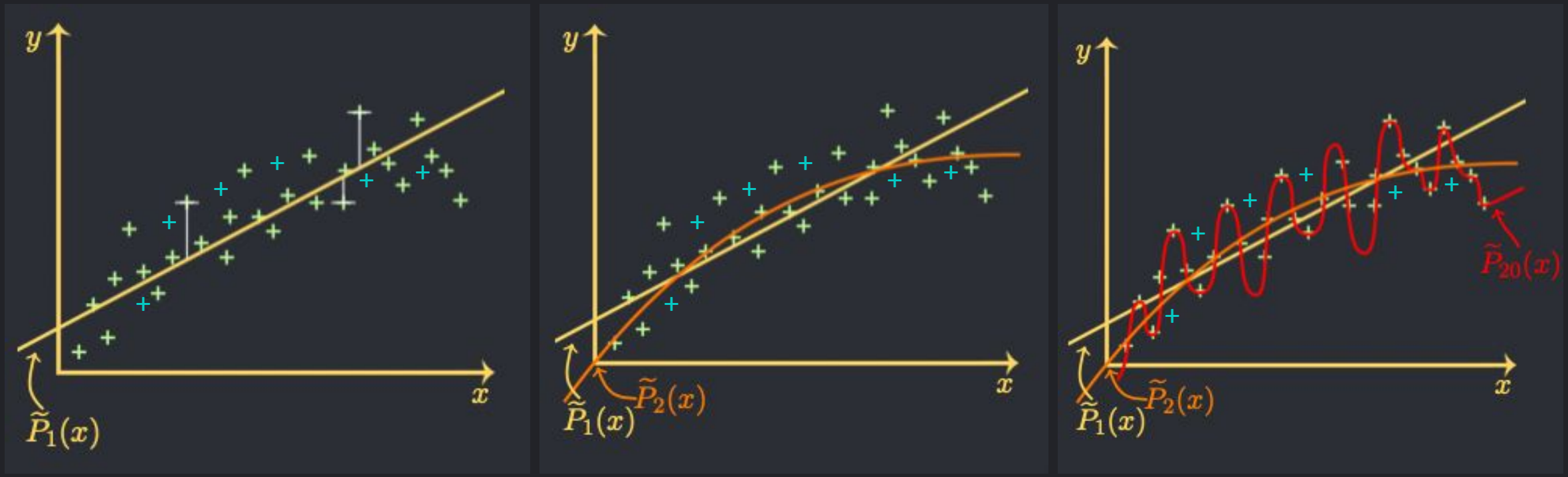
MAPE = ***0.45***

MSE = ***9.3***

01 ОСНОВЫ 10 МИН

Переобучение

Это ситуация, когда модель теряет обобщающую способность и решает конкретную задачу, «подстраиваясь» под обучающую выборку.

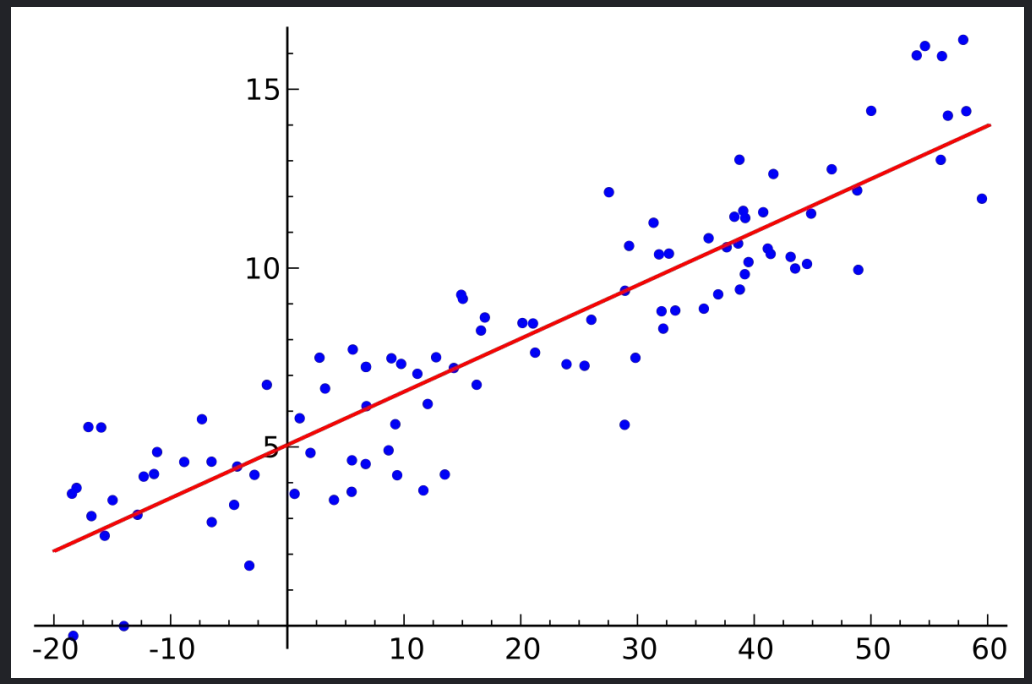


02 МЕТОДЫ
15 МИН

Типы алгоритмов регрессии*

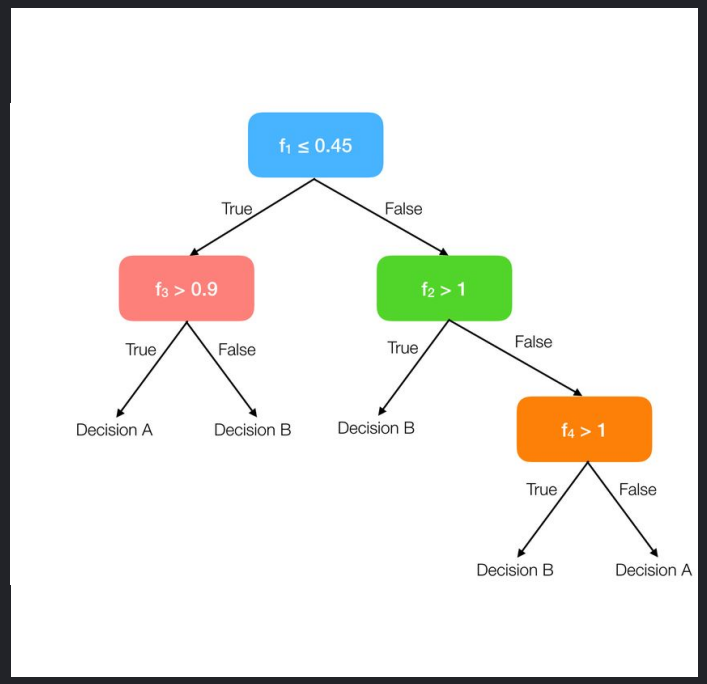
Линейные методы

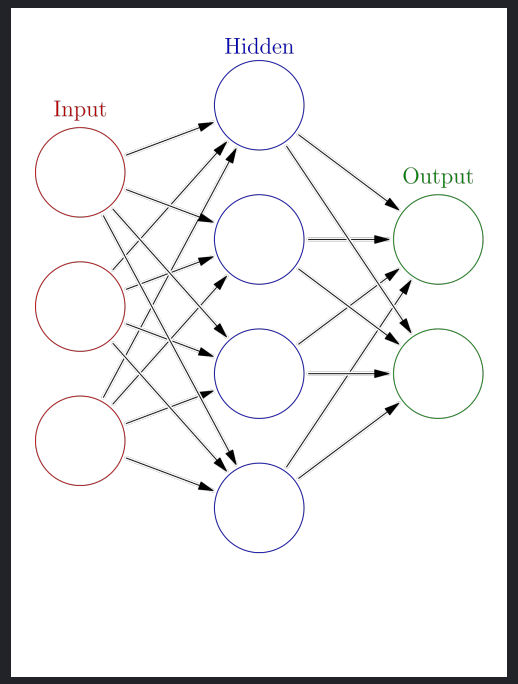
$$Y = \langle X \cdot W \rangle$$



Least Squares, Ridge regression, LASSO, etc.

Нелинейные





Decision tree regressor, neural networks, etc.

*существуют и другие типы, но мы остановимся только на этих двух

Линейные методы

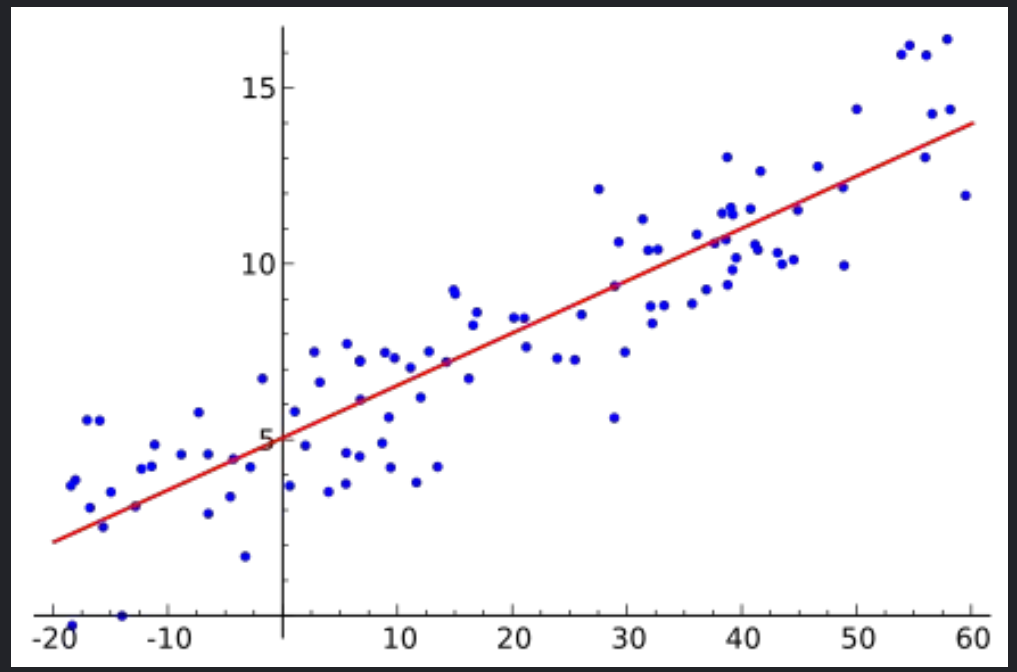
Дано:
X – матрица объектов-признаков, $x_i \in R^n$
Y – вектор ответов, $y_i \in (-\infty;\ +\infty)$

Постановка задачи линейной регрессии:
$Y = f(X \cdot w) = \langle X \ \cdot w \rangle$

Найти:
w – ?

w – вектор весов, $w_i \in R^n$
$\langle X \ \cdot w \rangle$ – скалярное произведение

$$\underbrace{\begin{pmatrix} x_{1,1} & x_{1,2} & \cdots \\ x_{2,1} & x_{2,2} & \cdots \\ \vdots & \vdots & \ddots \\ x_{n,1} & x_{n,2} & \cdots \end{pmatrix}}_{X} \quad \underbrace{\qquad\qquad}_{w} \quad \underbrace{\qquad}_{Y}$$



Решение задачи

Решение задачи – поиск вектора весов.
Алгоритмы решения: SVD, итеративные, точные.

Точное решение

$$Q = \sum_{i=1}^{n} (f(x_i, w) - y_i)^2 = \|X \cdot W - Y\|^2 \to \min_w.$$

$$\frac{dQ}{dw} = 2X^T(X \cdot W - Y) = 0$$

$$X^T X \cdot W = X^T Y$$

$$W = (X^T X)^{-1} X^T Y$$

Зачастую, невозможно, из-за линейно-зависимости, почти линейно-зависимости.
Матрица $X^T X$ вырождена.

SVD

Singular Value Decomposition (сингулярное разложение).
$X = VDU^T$

V: $V^T V = I$ – матрица собственных векторов XX^T;
U: $U^T U = I$ – матрица собственных векторов $X^T X$;
D: $D = diag(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots \sqrt{\lambda_n})$ – диагональная матрица сингулярных чисел чисел X.

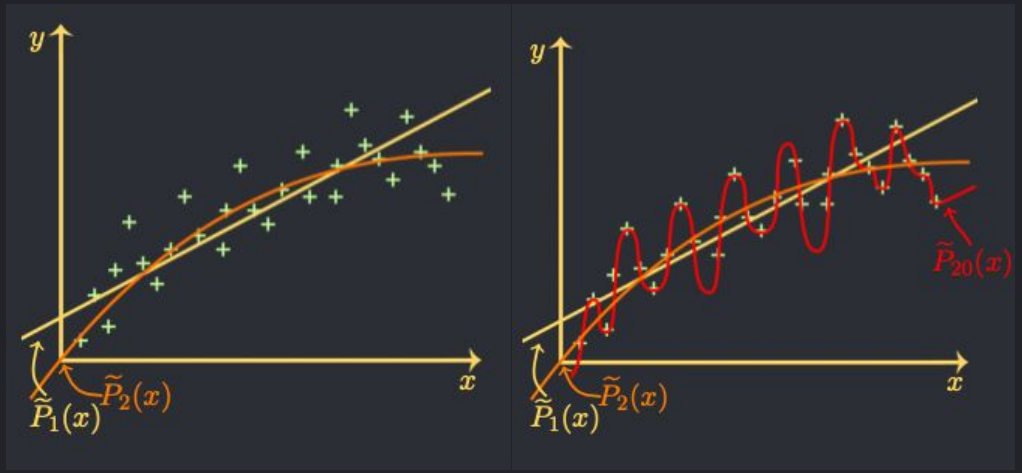
Где ещё нужно SVD?

В областях от медицины до энергетики.
В задачах шумоподавления, сжатия, упрощения вычислений и др.

02 МЕТОДЫ
15 МИН

Проблемы

Переобучение



Мультиколлинеарность

$$X = \begin{vmatrix} 1 & 4 & 2 & 2 \\ 0 & 1 & 0 & 9 \\ 1 & 4 & 2 & 2 \\ 3 & 0 & 6 & 3 \end{vmatrix}$$

Линейно-зависимые строки

Линейно-зависимые столбцы

В чем это выражается?

Падение качества

MAPE ⬇

MSE ⬇

Дисбаланс весов w

$$W = (-1234 \quad 1802 \quad 2 \quad 150 \quad -180 \quad 0.5)$$

Регуляризация

Регуляризация – метод упрощения модели, борьба с переобучением, мультиколлинеаностью и другими проблемами.

L1

LASSO regression,
обнуление
весов.

$$W = (0.7 \quad 0 \quad 0 \quad -0.3 \quad 0 \quad 0.2)$$

нулевые веса

$$Q = \|X \cdot W - Y\|^2 + \alpha\|W\| \to \min_w,$$

α – параметр регуляризации.

L2

Ridge (гребневая) regression,
weight decay (сокращение
весов),
регуляризация Тихонова.

$$W = (0.7 \quad -0.2 \quad 0.9 \quad -0.3 \quad -0.15 \quad 0.2)$$

отсутствуют большие по модулю веса

$$Q = \|X \cdot W - Y\|^2 + \alpha\|W\|^2 \to \min_w,$$

α – параметр регуляризации.

03 ПЛАН РЕШЕНИЯ ML-ЗАДАЧ 5 МИН

Второе приближение

Меняем ?

step 1

Загрузка данных

Предобработка данных: заполняем пропуски, удаляем дубликаты, кодируем категориальные переменные.

step 2

Разделение

Делим на обучающую и тестовую выборки.

new step

Scaling

Масштабируем данные

step 4

Выбор методов

Выбираем алгоритм, метрику качества.

step 5

Обучаем

Обучаем модель.

step 6

Контроль

Анализируем результаты.

04 ПРАКТИКА
10 МИН

Практика 3

Вторая ML-задача.

Откройте учебный Notebook из архива с материалами к занятию.