



Машинное

День 3. Регрессия.

обучение



ПЛАН

День 3. Регрессия.

01

ОСНОВЫ
10 МИН



Дополнительные понятия и проблемы.
Масштабирование признаков, метрики регрессии.

Методы регрессии: линейные методы,
регуляризация.



МЕТОДЫ
15 МИН

02

03

ПЛАН РЕШЕНИЯ
ML-ЗАДАЧ 5 МИН



Процесс решения ML-задачи: второе
приближение.

Решим задачу регрессии.



ПРАКТИКА
10 МИН

04

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

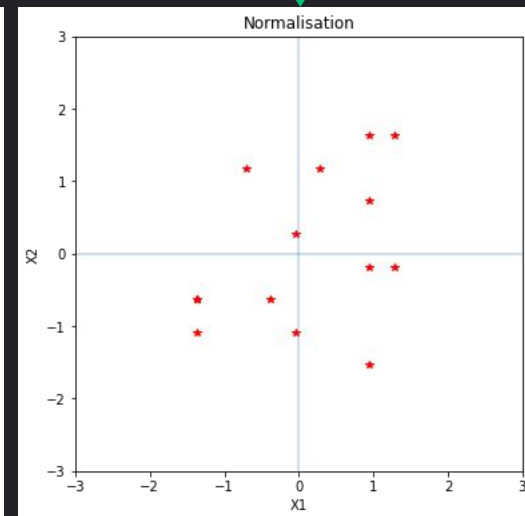
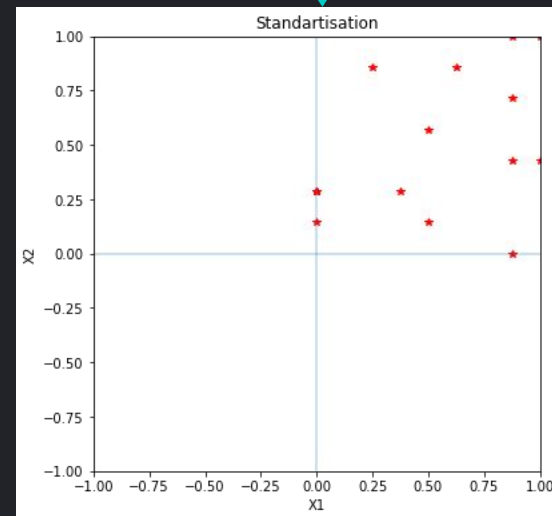
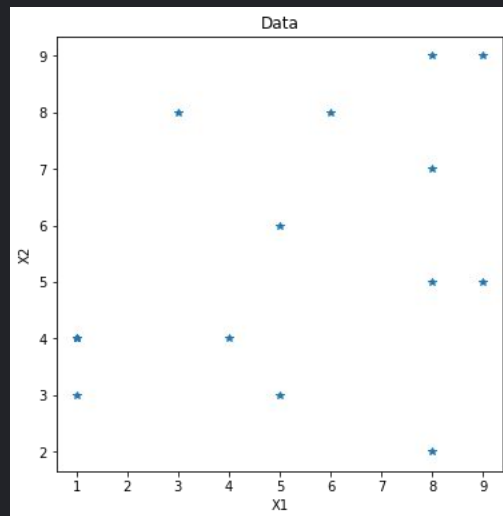
Масштабирование признаков:

▶ стандартизац
 $X \in [0; 1]$

▶ нормализаци
 $X \sim D(0,1)$

$$\frac{X - X_{min}}{X_{max} - X_{min}}$$

$$\frac{X - \mu}{\sigma}$$



Метрик

а

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**

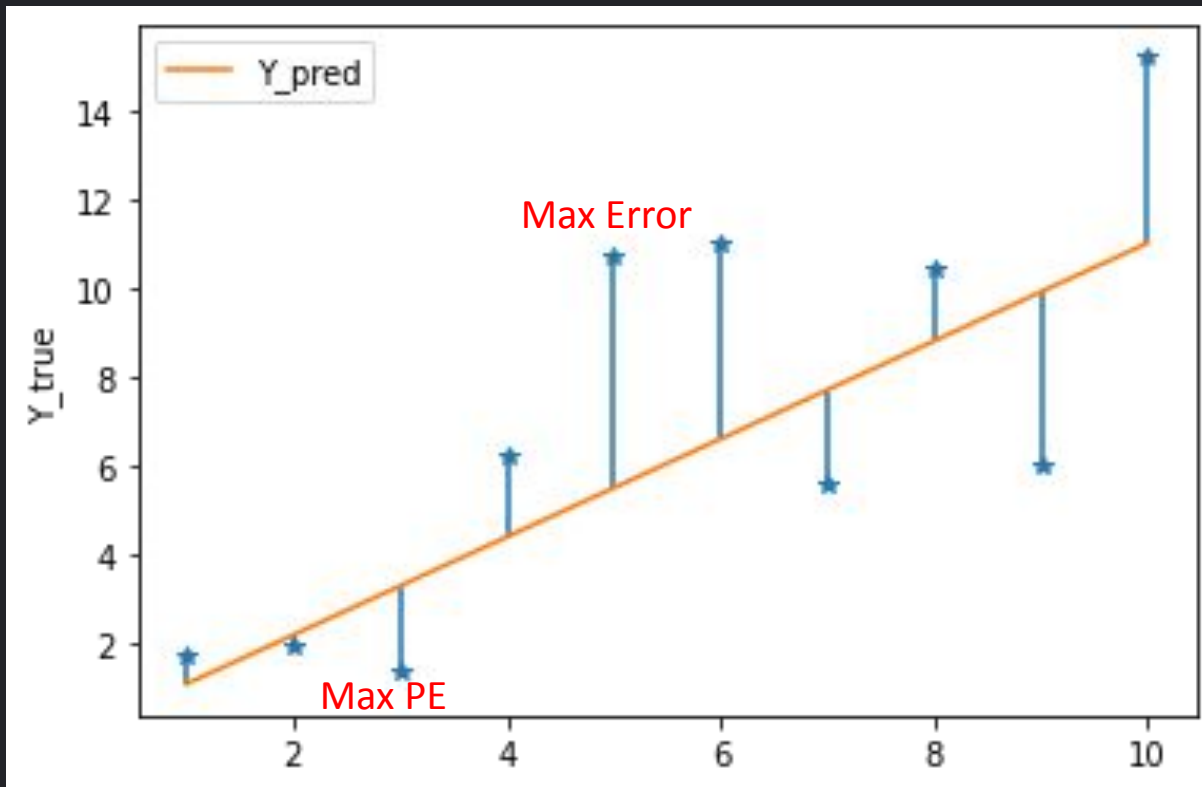
Mean Absolut Percentage Error

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

Среднеквадратичная ошибка | **MSE**

Mean Squared Error

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



$Max E = 5.22$

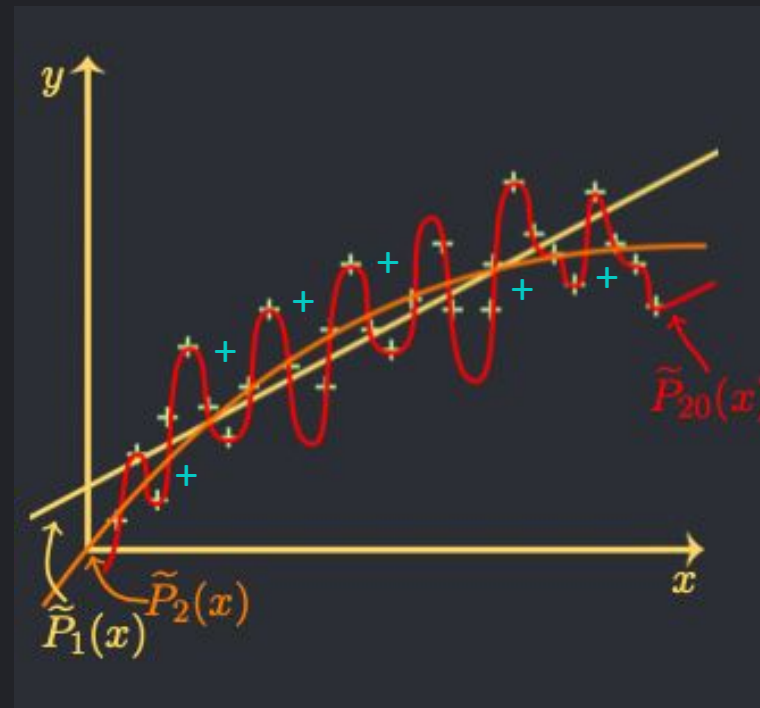
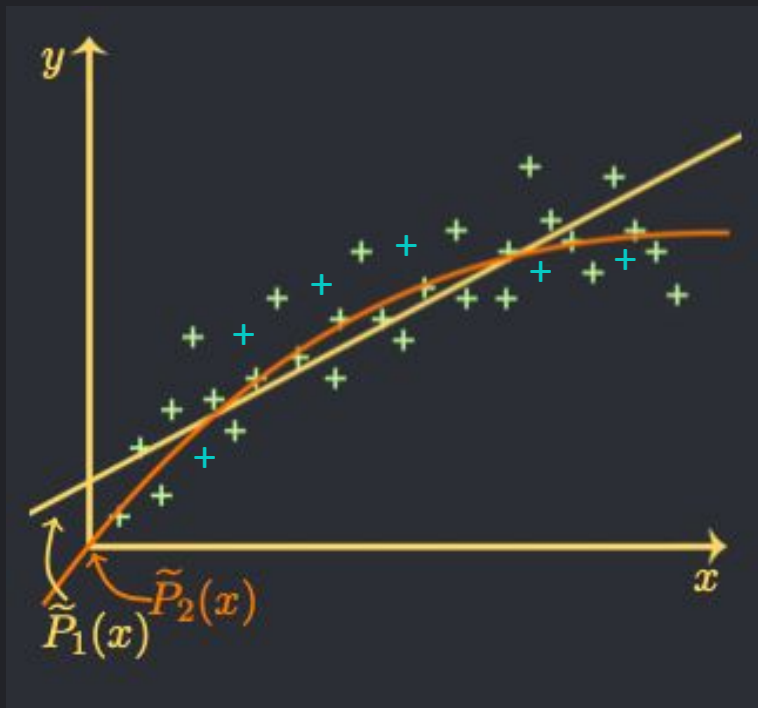
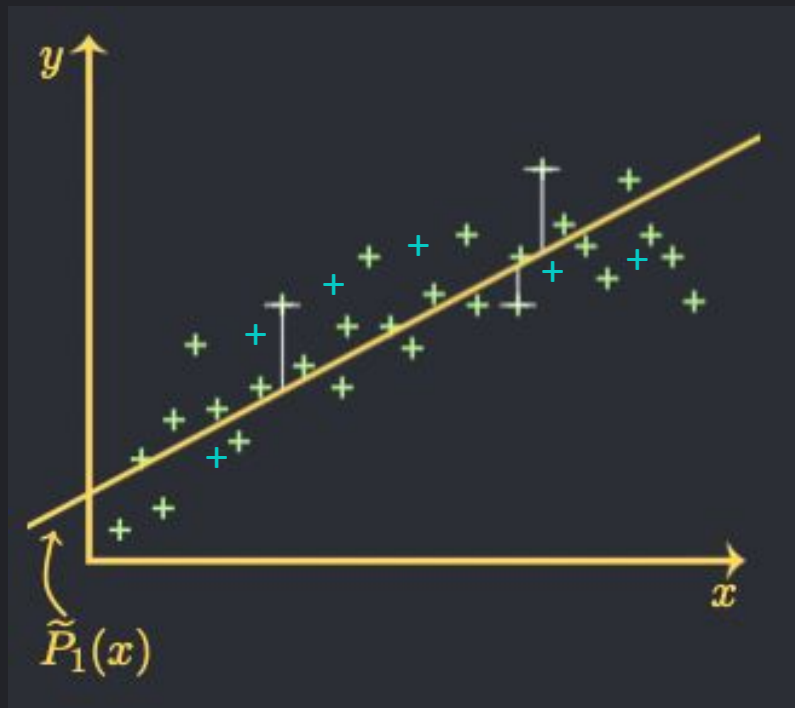
$Max PE = 1.41$

$MAE = 2.59$

$MAPE = 0.45$

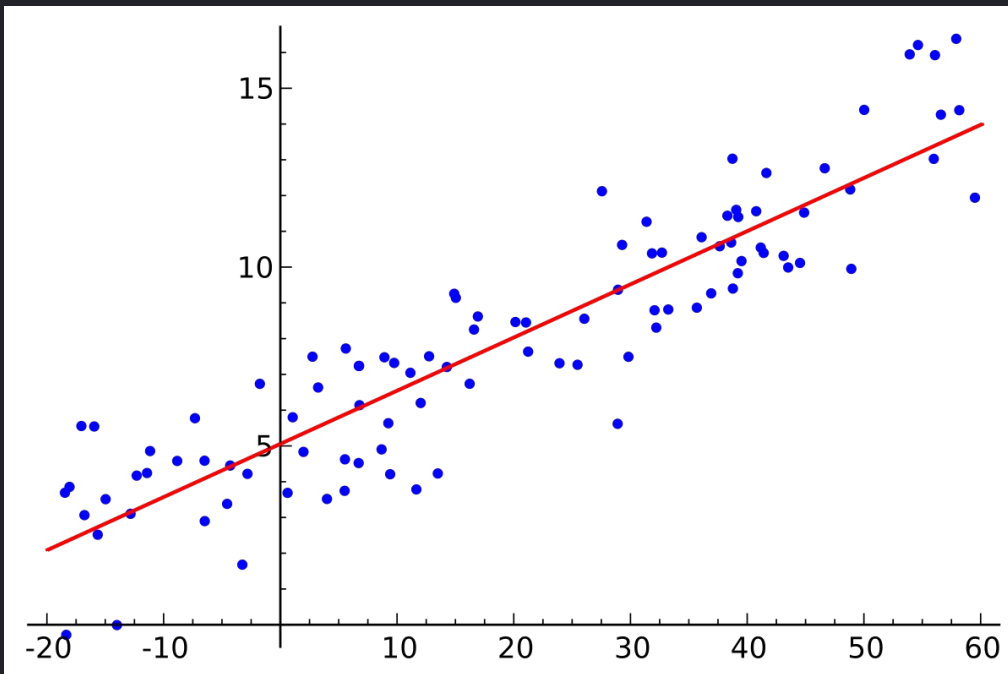
$MSE = 9.3$

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



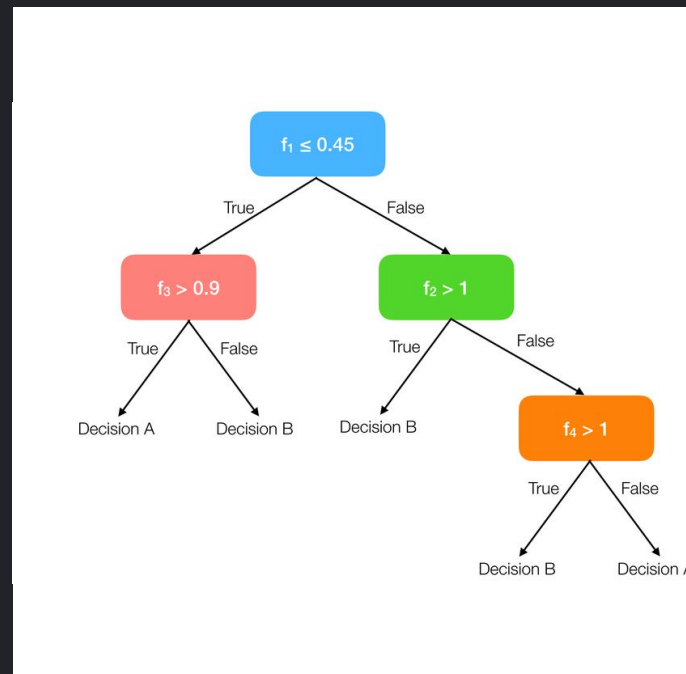
Линейные методы

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

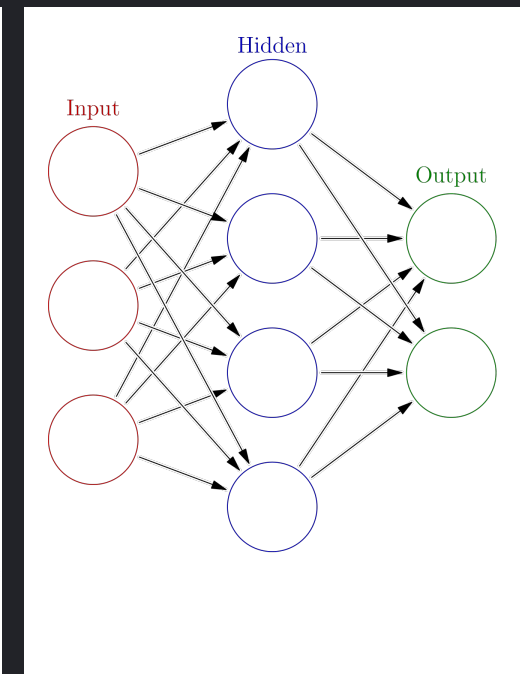


Least Squares, Ridge regression, LASSO, etc.

Нелинейные



Decision tree regressor, neural networks, etc.



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

Дано:

X – матрица объектов-признаков, $x_i \in R^n$

Y – вектор ответов, $y_i \in (-\infty; +\infty)$

$$\begin{pmatrix} x_{1,1} & x_{1,2} & \dots \\ x_{2,1} & x_{2,2} & \dots \\ \vdots & \vdots & \ddots \\ x_{n,1} & x_{n,2} & \dots \end{pmatrix}$$

X

w

Y

Постановка задачи линейной регрессии:

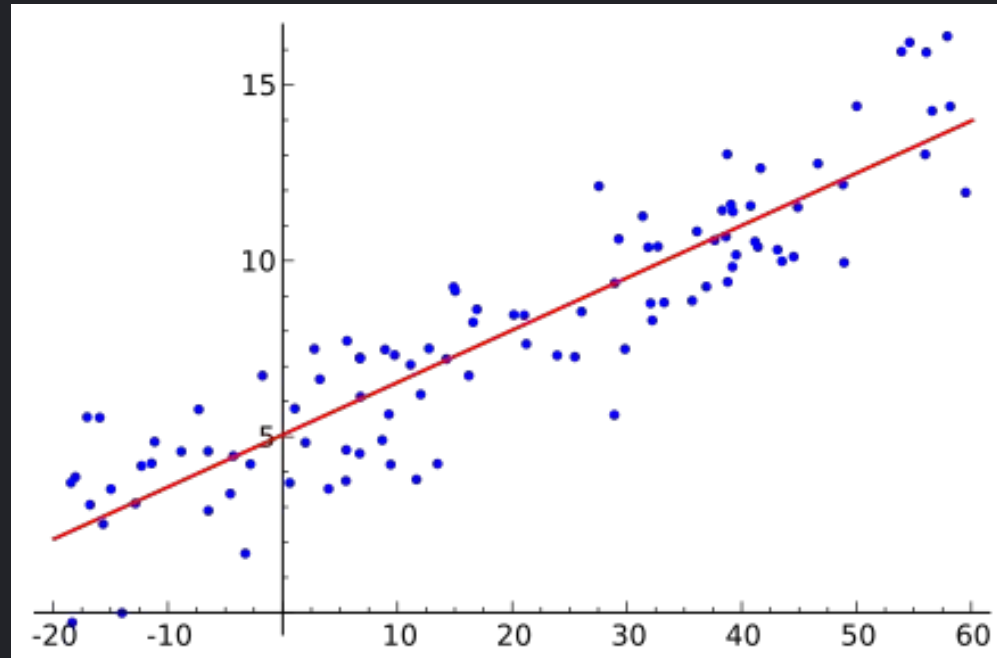
$$Y = f(X \cdot w) = \langle X \cdot w \rangle$$

Найти:

w – ?

w – вектор весов, $w_i \in R^n$

$\langle X \cdot w \rangle$ – скалярное произведение



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

Точное
решение

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

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

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

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

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 $X^T X$;

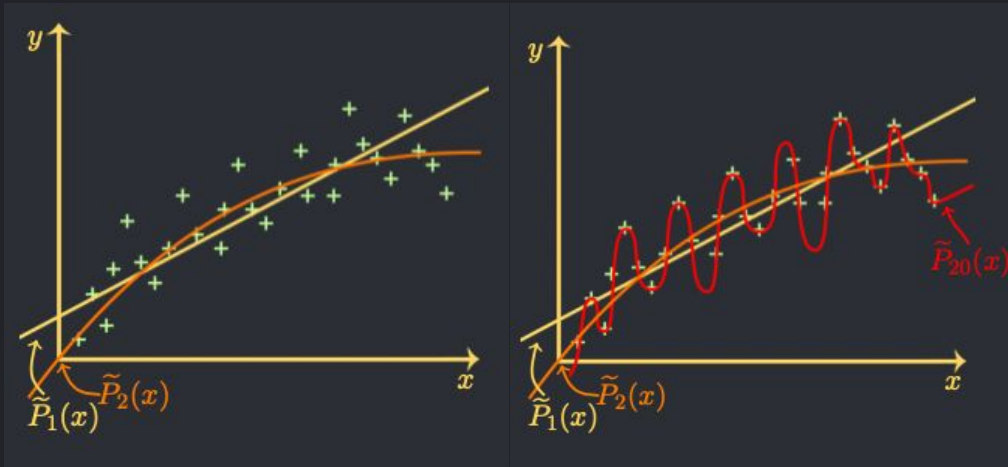
D : $D = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_n})$ – диагональная матрица сингулярных чисел X .

Где ещё нужно SVD?

В областях от медицины до энергетики.

В задачах шумоподавления, сжатия, упрощения вычислений и др.

Переобучение



В чем это
выражается?

Падение
качества

MAPE ↓

MSE ↓

Мультиколлинеарность

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

Линейно-зависимые строки

Линейно-зависимые столбцы

Дисбаланс весов

W

$W =$

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

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

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

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

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

α – параметр регуляризации.

03

ПЛАН РЕШЕНИЯ
ML-ЗАДАЧ 5 МИН

Второе приближение



Практика 3

Вторая ML-задача.

Откройте учебный Notebook из архива с материалами к занятию.