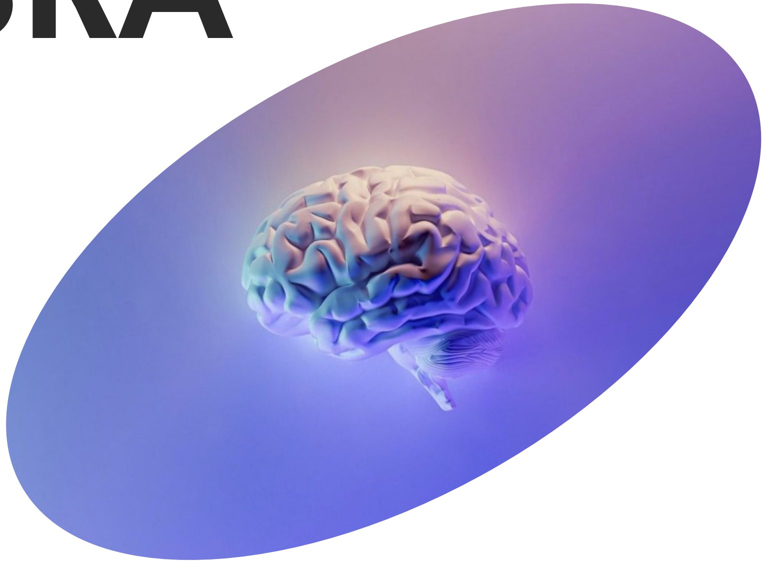
ПОСТАНОВКА ЗАДАЧИ



ДАННЫЕ ДЛЯ ОБУЧЕНИЯ

X – множество объектов (входные данные)

Y – множество ответов (выходные данные)

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

n – количество признаков

m – количество примеров

ЦЕЛЕВАЯ ФУНКЦИЯ

$f : X \to Y$ - истинная зависимость (закон природы)

$\hat{f}$ – приближающая зависимость (хотим ее получить)

Введем ограничения на функции:

1. Функция должна быть вычислима.

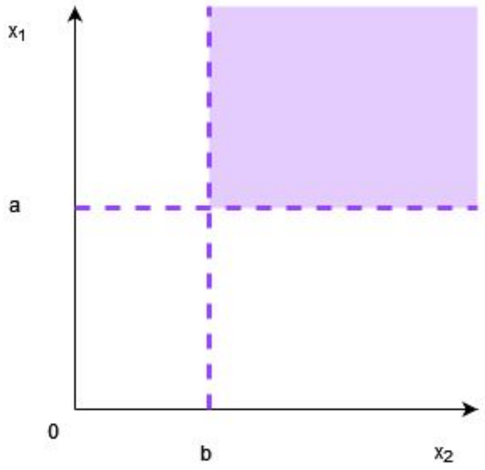
2. Выбираем функцию из некоторого параметризованного семейства.

ПРИМЕР

Задача определить можно ли пройти ребенку на аттракцион

в зависимости от его роста и возраста.

$$\hat{f}_{(a,b)}(x_1, x_2) = \begin{cases} 1 & x_1 > a \quad \& \quad x_2 > b \\ 0 & \text{иначе} \end{cases}$$



Вектор (a,b) - параметр

ФУНКЦИЯ ПОТЕРЬ

$L(y, \hat{y})$ – функция потерь (loss). Показывает как сильно отличается предсказанные значения от реальных.

Примеры:

1. $L(y, \hat{y}) = (y - \hat{y})^2$ - квадратичная
2. $L(y, \hat{y}) = |y - \hat{y}|$ - абсолютная

ЭМПИРИЧЕСКИЙ РИСК

Эмпирический риск – среднее значение функции потерь на обучающем датасете.

$$\frac{1}{m}\sum_{i=1}^{m} L(y^i, \hat{y}_w^i)$$

$$w_{best} = \arg\min_{w \in \Theta} \frac{1}{m}\sum_{i=1}^{m} L(y^i, \hat{y}_w^i)$$

ЭМПИРИЧЕСКИЙ РИСК

Минимизируем

1

MSE (Mean Squared Error)

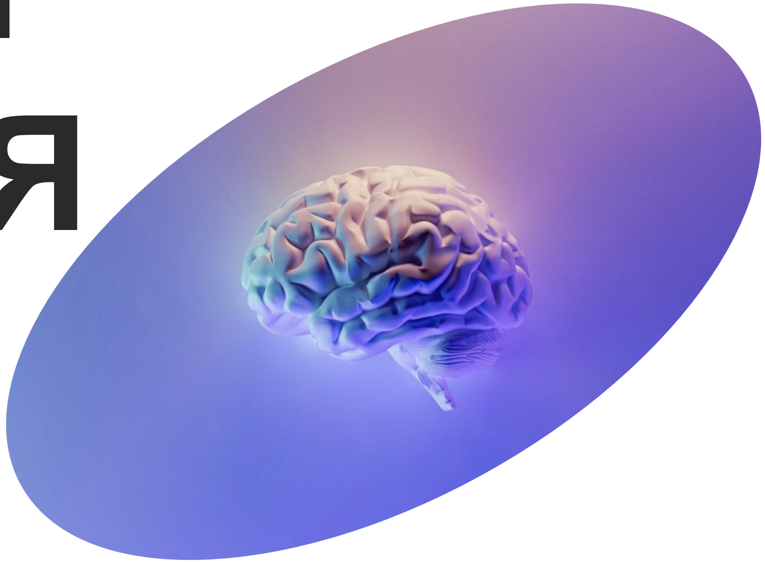
$$w = \arg\min_{w \in \Theta} \frac{1}{m} \sum_i (y_i - \hat{y}_i)^2$$

2

MAE (Mean Absolute Error)

$$w = \arg\min_{w \in \Theta} \frac{1}{m} \sum_i |y_i - \hat{y}_i|$$

ЛИНЕЙНАЯ РЕГРЕССИЯ



ДАННЫЕ ДЛЯ ОБУЧЕНИЯ

X – множество объектов (входные данные)

Y – множество ответов (выходные данные)

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

$$L(f) = \sum_{i=1}^{m} (y^i - f(x_1^i, \dots, x_n^i))^2$$

МОДЕЛЬ

Будем искать неизвестную функцию f в виде:

$$f(x) = w_0 + w_1 x_1 + \cdots + w_n x_n$$

Переопределим вектор $x = (1, x_1, \ldots, x_n)$

и запишем регрессию в векторном виде

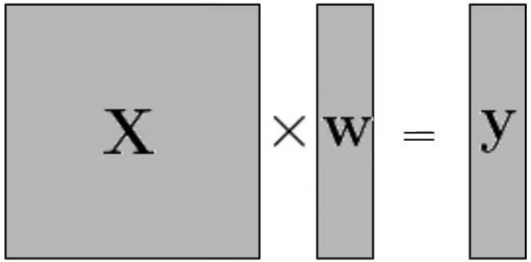
$$f(x) = \sum_{i=1}^{n} w_i x_i = \mathbf{xw}$$

ТОЧНОЕ РЕШЕНИЕ

1

$\mathbf{X}$ - квадратная матрица $(m = n)$

Тогда решение $w = X^{-1}y$

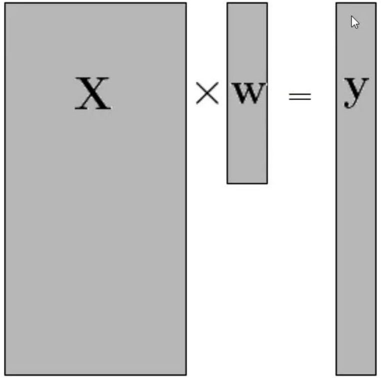


ТОЧНОЕ РЕШЕНИЕ

2

X - прямоугольная матрица $(m >> n)$

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



ТОЧНОЕ РЕШЕНИЕ ЧЕРЕЗ ПРОИЗВОДНУЮ

2

Найдем решение через производную.

Подставим выражение $f(x)$ в функцию потерь.

$$L(f) = \sum_{i=1}^{m} (y_{true}^{i} - x_i^T w)^2 = (Xw - y)^T (Xw - y)$$

$$\frac{\partial L(f)}{\partial w} = 2X^T (Xw - y)$$

ТОЧНОЕ РЕШЕНИЕ ЧЕРЕЗ ПРОИЗВОДНУЮ

2

$$2X^T(Xw - y) = 0$$

Решение

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

ПОЧЕМУ ЕЩЕ ИСПОЛЬЗУЮТСЯ ЛИНЕЙНЫЕ МОДЕЛИ?

1. **Понятно какие признаки вносят больший вклад в результат**
2. **Легко бороться с переобучением**
3. **Легко применять**

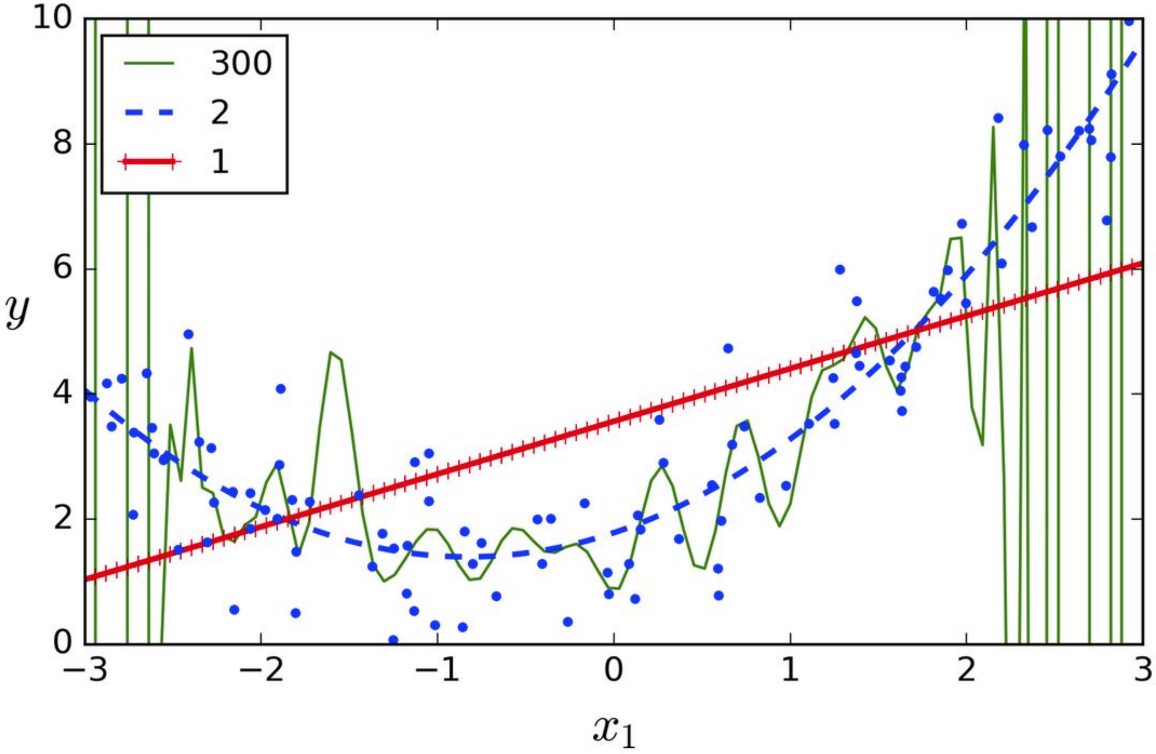
ПОЛИНОМИАЛЬНАЯ РЕГРЕССИЯ

Пусть у нас изначально есть только один признак x. Создадим новые:
$x_1 = x, x_2 = x^2, \ldots, x_n = x^n$

Линейная регрессия от таких признаков будет полиномиальной.

$f(x) = w_0 + w_1 x + w_2 x^2 + \cdots + w_n x^n$

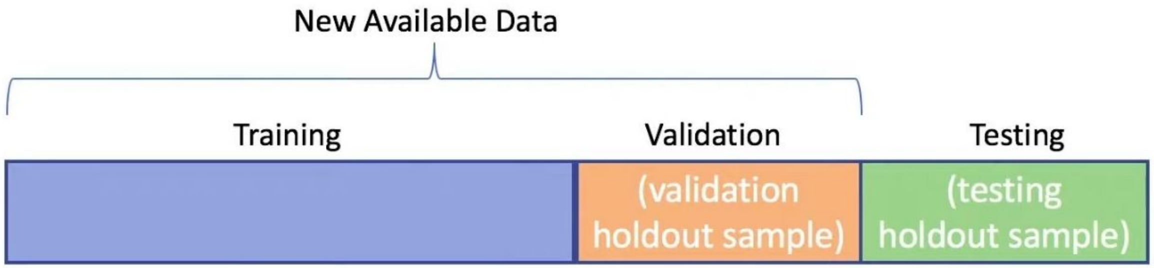
ПЕРЕОБУЧЕНИЕ



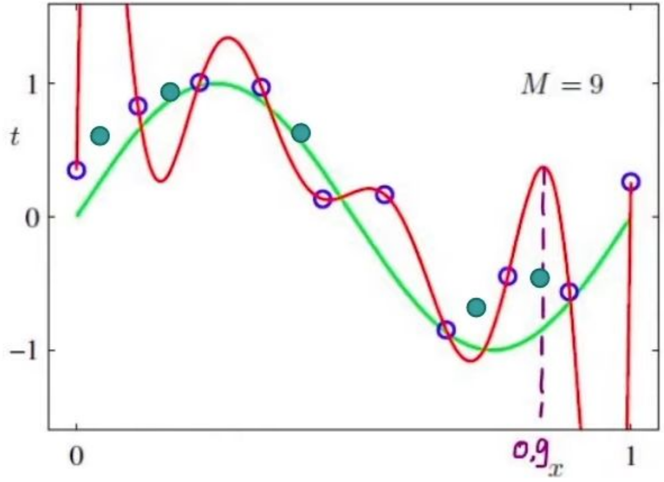
ПЕРЕОБУЧЕНИЕ

1) **Train** - данные для обучения
2) **Validation** - данные для оценки качества модели
3) **Test** - для финальной оценки качества

Переобучение - ситуация, когда качество на train значительно лучше, чем на validation.



ПЕРЕОБУЧЕНИЕ



● -точка из test датасета

○ -точка из train датасета

КОДИРОВАНИЕ ПРИЗНАКОВ

LABEL ENCODER

Однозначное соответствие числа и уникального значения

1. BMW
2. Mercedes
3. Nissan
4. Infinity
5. Audi
6. Volvo
7. Skoda

КОДИРОВАНИЕ ПРИЗНАКОВ

ONE-HOT ENCODER

	BMW	Mercedes	Nissan	Infinity	Audi	Volvo	Skoda
BMW	1	0	0	0	0	0	0
Mercedes	0	1	0	0	0	0	0
Skoda	0	0	1	0	0	0	1
Volvo	0	0	0	0	0	1	0