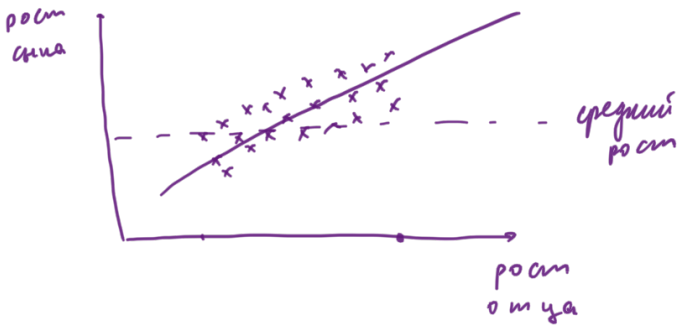
Линейная регрессия

$Y = \mathbb{R}$



① Модель линейной регрессии

$$a(x) = w_0 + w_1 x_1 + \dots + w_d x_d = w_0 + \sum_{j=1}^{d} w_j x_j \quad (=)$$

w_0 — свободный коэффициент (bias)

$w_1, \dots, w_d$ — веса коэффициенты параметры

$(d+1)$ параметр

(=) $w_0 + \langle w, x \rangle$

предположение: 1-й признак всегда = 1

$w_1 \cdot x_1$

$w_1 \cdot 1$

$a(x) = \langle w, x \rangle$

② Области применимости

x - квартира в Москве

y - рыночная стоимость

$a(x) = w_0 + w_1 \cdot$ (площадь) $+ w_2 \cdot$ (кол-во комнат) $+ \dots$

$+ w_3 \cdot (\text{расст до метро}) + \sout{w_4 (\text{район})}$

1) категориальные признаки

x_1 - район

$C = \{c_1, \dots c_n\}$ - мн-во значений признака

$b_1(x), \dots b_n(x)$ - новые признаки

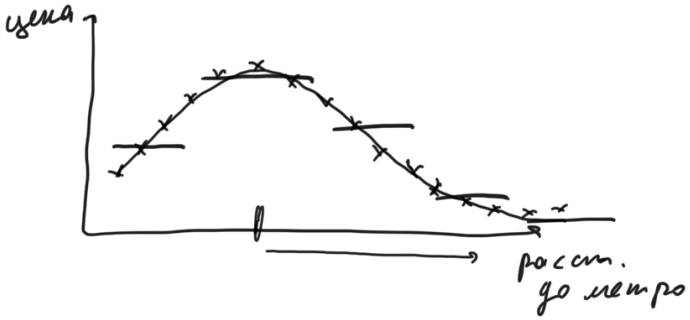
$b_j(x) = [x_1 = c_j]$

$b_1(x) + \dots + b_n(x) = 1$

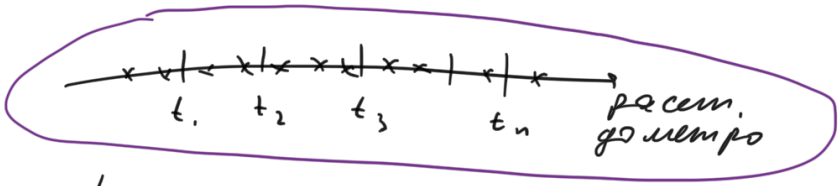
} one-hot кодирование

$$a(x) = w_0 + w_1 \cdot [x_1 = c_1] + w_2 \cdot [x_1 = c_2] + \dots + w_n \cdot [x_1 = c_n] + \dots$$

2) бинаризация числовых признаков
бинаризация



$w_3 \cdot (\text{расст} \cdot \text{до метро})$



$t_0 = -\infty$

$t_{n+1} = +\infty$

$b_i(x) = [t_{i-1} \le x_1 < t_i]$

$$a(x) = w_1 [t_0 \le x_1 < t_1] + \dots + w_{n+1} [t_n \le x_1 < t_{n+1}]$$

Для линейных моделей признаки нужно готовить так, чтобы модель была осмысленной

③ Измерение ошибки в задачах регрессии

1) $L(y, z) = (y - z)^2$ — квадратичная ф-ция потерь

$$MSE(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \underbrace{(a(x_i) - y_i)^2}_{\text{квадратные рубли}}$$

($a(x_i)$ — рубли, y_i — рубли)

$$RMSE(a, X) = \underbrace{\sqrt{MSE(a, X)}}_{\text{рубли}}$$

коэф. детерминации:

$$R^2(a, X) = 1 - \frac{\sum_{i=1}^{\ell} (a(x_i) - y_i)^2}{\sum_{i=1}^{\ell} (y_i - \bar{y})^2}$$

$a(x)$ идеальная $\Rightarrow R^2 = 1$

$a(x)$ константа $\Rightarrow R^2 = 0$
$(a(x) = \bar{y})$

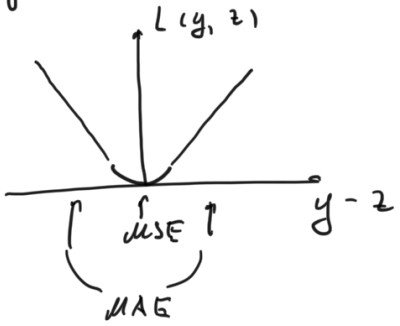
2) $L(y, z) = |y - z|$

$$MAE(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|$$

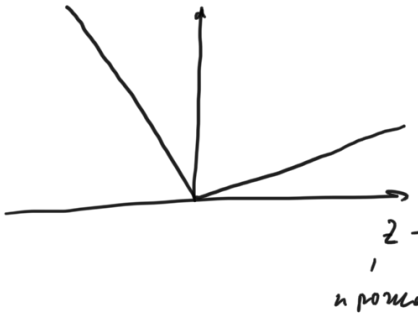
	y	$a(x)$	$(y - a(x))^2$	$\lvert y - a(x) \rvert$
выброс	1	2	1	1
	1000	2	996004	998
	1	1	0	0
	1000	3	994009	997

=> MAE устойчива к выбросам

3) Функция Хубера (Huber loss)



4) несимметричные ф-ции потерь



$$L(y,z) = \begin{cases} \\ \end{cases}$$

5) $L(y,z) = (\log(z+1) - \log(y+1))^2$ - MSLE

6) $L(y,z) = \left|\frac{y-z}{y}\right|$ $\quad$ 2 (1000) $\quad$ 1 (999)

$$MAPE(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}\frac{|y_i - a(x_i)|}{|y_i|}$$

④ Обучение

$$MSE: \frac{1}{\ell}\sum_{i=1}^{\ell}(\langle w, x_i\rangle - y_i)^2 \to \min_w$$

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ \vdots & & & \\ x_{\ell 1} & x_{\ell 2} & \dots & x_{\ell d} \end{pmatrix}$$ - матрица объекты - признаки

$$y = \begin{pmatrix} y_1 \\ \vdots \end{pmatrix} \qquad w = \begin{pmatrix} w_1 \\ \vdots \end{pmatrix}$$ — вектор весов

$\begin{pmatrix} y_1 \\ \vdots \\ y_\ell \end{pmatrix}$ - вектор ответов $\quad \begin{pmatrix} w_1 \\ \vdots \\ w_d \end{pmatrix}$

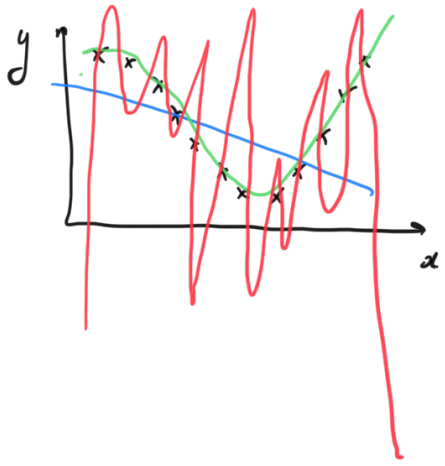
$$MSE = \frac{1}{\ell} \| \underbrace{Xw}_{\begin{pmatrix} \langle w, x_1 \rangle \\ \langle w, x_2 \rangle \\ \vdots \\ \langle w, x_\ell \rangle \end{pmatrix}} - y \|_2^2 \to \min_w$$

$$\nabla_w MSE = 0$$

$$w = \underbrace{(X^T X)^{-1}}_{d \times d} X^T y \quad (\text{только если } X \text{ полного ранга})$$

$\underline{O}(d^3)$ - сложность обращения

⑤ Обобщающая способность модели



$a(x) = w_0 + w_1 x$

$a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$

$a(x) = w_0 + w_1 x + w_2 x^2 + \dots + w_{13} x^{13}$

Переобучение (overfitting) - ошибка на новых данных больше, чем на обуч. выборке

Обнаружение: 1) Отлож. выборка

X - обуч.

X_t - тестовая выборка

$$\mathcal{L}(\theta, X) \leq \mathcal{L}(\theta, X_t)$$

2) процесс - вакцинация