$$Q(\beta) = \underbrace{(y - X\beta)^T}_{1\times m}\underbrace{(y - X\beta)}_{m\times 1} + \lambda\beta^T\beta$$

① взять производную (по вектору)

② взять дифференциал

$$d(AB) = (dA)B + A\,d(B)$$

$$d(Q(\beta)) = -\underbrace{d\beta^T}_{1\times m} X^T \underbrace{(y - X\beta)}_{m\times 1} - (y - X\beta)^T X\, d\beta + 2\beta^T d\beta = 0 \Big|_{\beta=\hat{\beta}}$$

$$+ 2(y - X\hat{\beta})^T X\, d\hat{\beta} = 2\lambda\hat{\beta}^T d\hat{\beta}$$

$$(y - X\hat{\beta})^T X = \lambda\beta^T$$

$$X^T(y - X\hat{\beta}) = \lambda\hat{\beta}$$

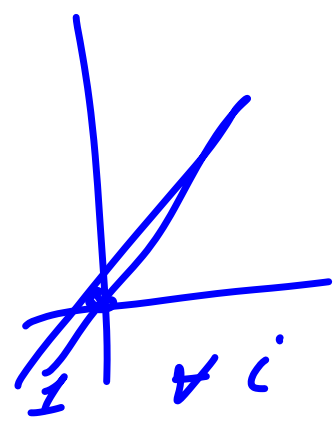
$$\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^Ty$$

$$\underbrace{(x_1\ x_2\ x_3)}_{x^T}\underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}}_{x} = \boxed{x^Tx}$$

$$\{(x_i, y_i)\}_{i=1}^{N}$$

$y_i \in R$ - наблюдения

$x_i \in R^{d+1}$ - признаки



(модель для y_i)

$$a(x_i) = \underline{\langle w, x_i \rangle} = w_0 + w_1 x_{1i} + w_2 x_{2i} + \dots w_d x_{di}$$

Обучение ☺

① аналитически

$$(X^T X)^{-1}$$

② градиентный спуск

$$\nabla f(x_1, \dots x_n) = (f'_{x_1}, \dots f'_{x_n})$$

- иниц. веса $w^{(0)} = (w_0^{(0)}, w_1^{(0)} \dots w_d^{(0)})$
- обновим

$$w^{(k)} = w^{(k-1)} - \eta_k \nabla Q(w^{(k-1)})$$

↑ learning rate

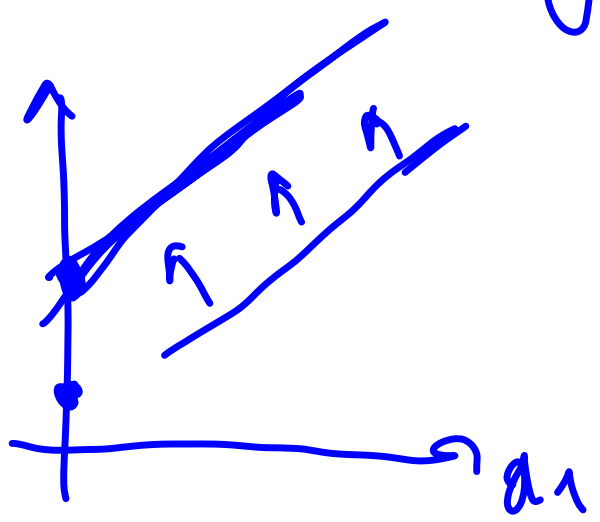
$$Q(w) = \frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \underbrace{\langle w, x_i\rangle}_{\hat{y}_i}\Big)^2 + \underbrace{\lambda \sum_{j=1}^{d} w_j^2}_{\lambda\langle w, w\rangle}$$

$\hat{y}_i = \sum_j x_{ij} w_j$

---

① $\nabla_w Q(w) = -\frac{1}{n}\sum 2x_i(y - \langle w, x_i\rangle) + 2\lambda w$

$y_1 \dots y_n$



$\eta > 0$

$$w_0^{(1)} = w_0^{(0)} - \eta\left[-\frac{2}{n}\sum_{i=1}^{n} x_{i0}(y_i - \hat{y}_i)\right]$$

$$w_d^{(1)} = w_d^{(0)} - \eta\left[\underbrace{-\frac{2}{n}\sum_{i=1}^{n} x_{id}(y_i - \hat{y}_i)}_{\frac{\partial}{\partial w_d} Q(w^{(0)})} + 2\lambda w_d^{(0)}\right]$$

$$w_d^{(k)} = w_d^{(k-1)} - \eta[\ \dots\ ]$$

$$w_d^{(1)} = \left(w_d^{(0)} - 2\eta\lambda w_d^{(0)}\right) - \eta\, grad_{OLS}$$

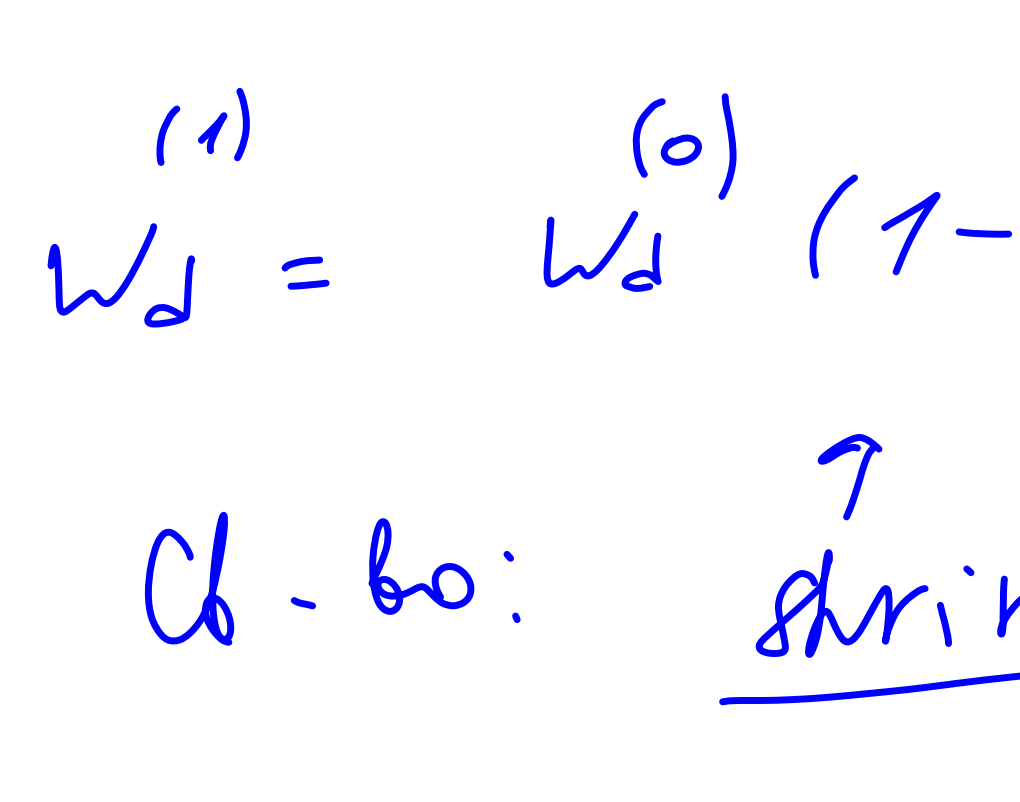
где MNK (OLS)

$$w_d^{(1)} = w_d^{(0)} - \eta\, grad_{OLS}$$

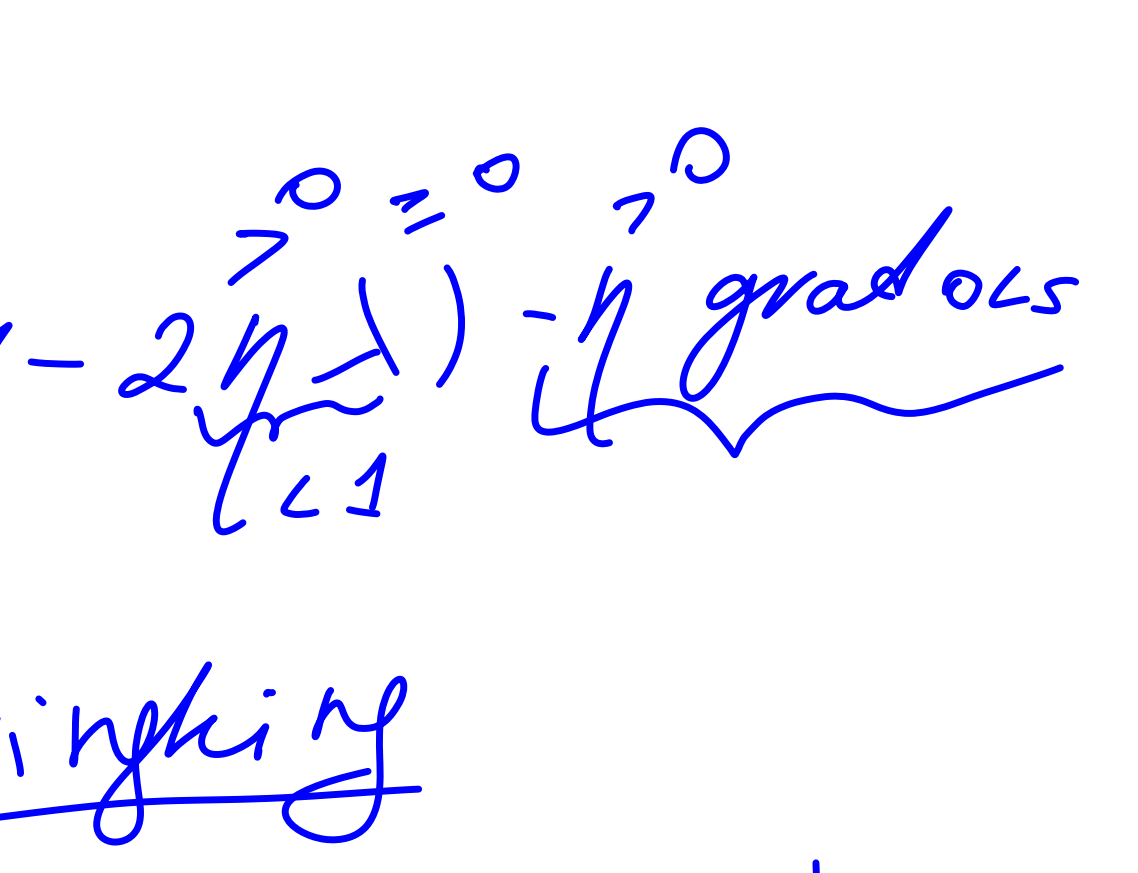
$$w_d^{(1)} = w_d^{(0)}\underbrace{(1 - 2\eta\lambda)}_{\geq 0,\ <1} - \eta\, grad_{OLS}$$

Сл-во: shrinking

Lasso



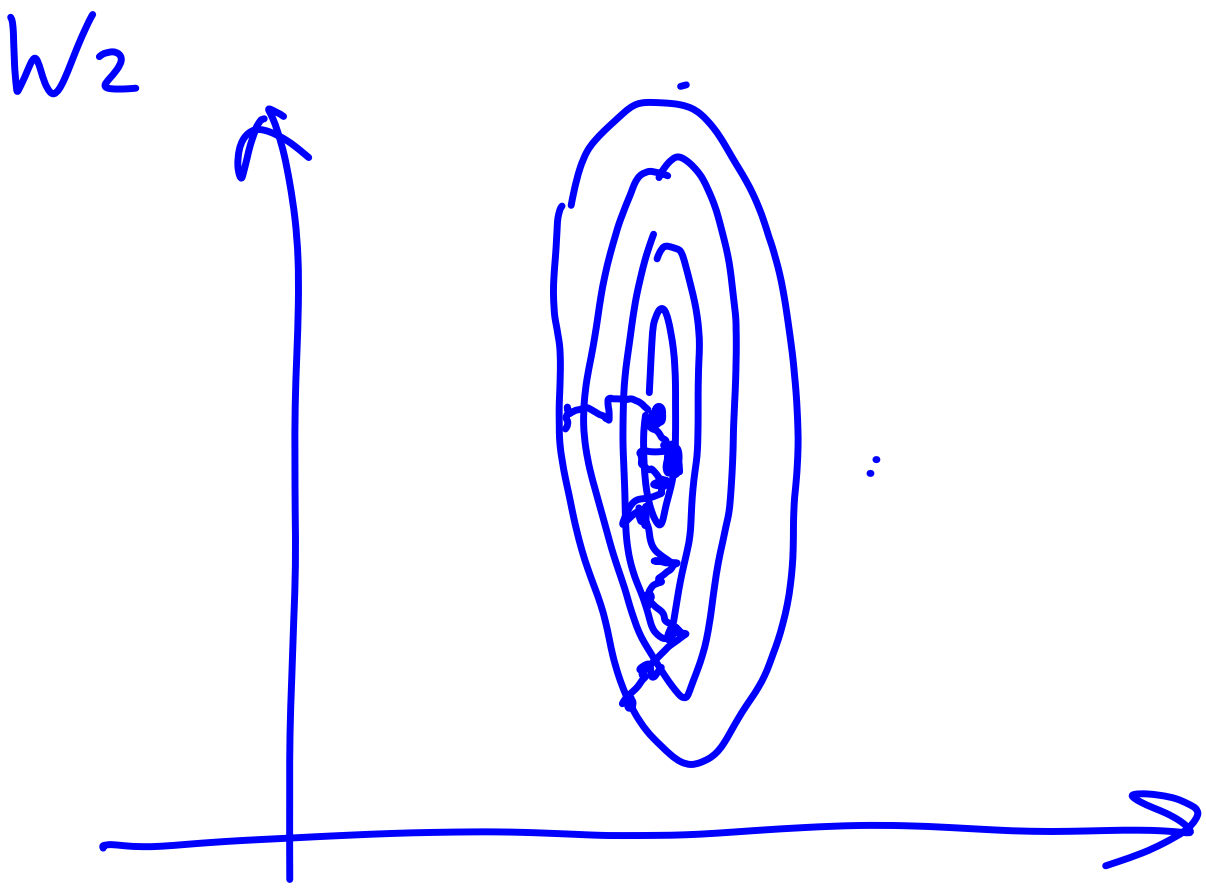
Ridge

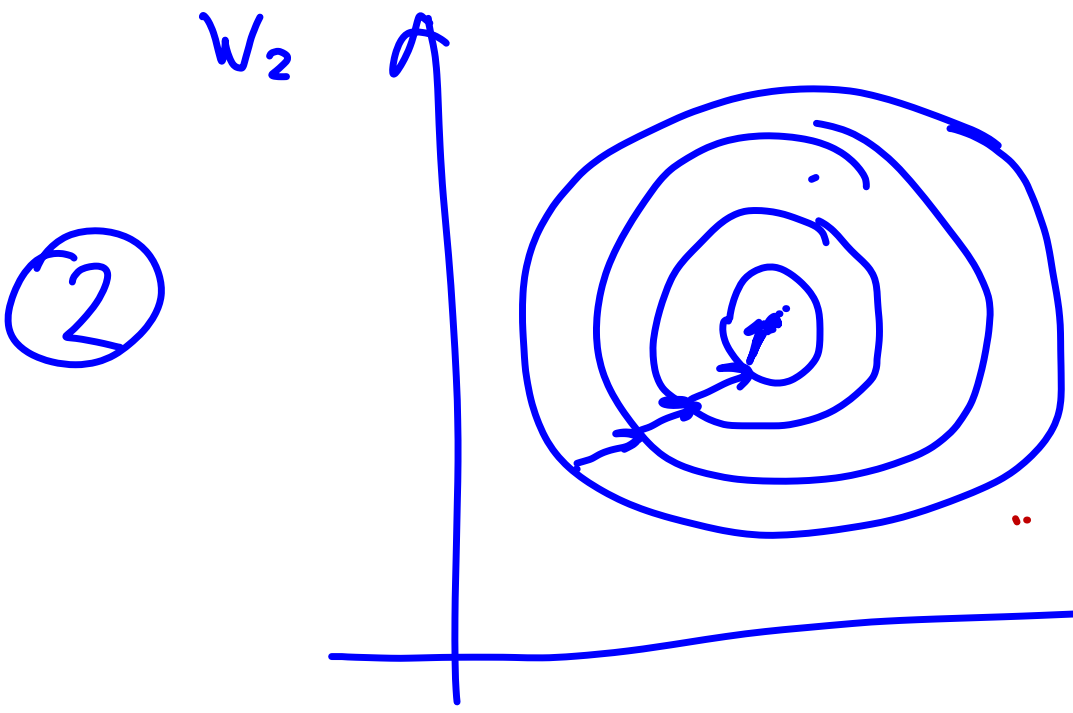


Нормализация

① $x_1 \in [20, 200] \rightarrow w_1$

$x_2 \in [1, 5] \rightarrow w_2$



② 

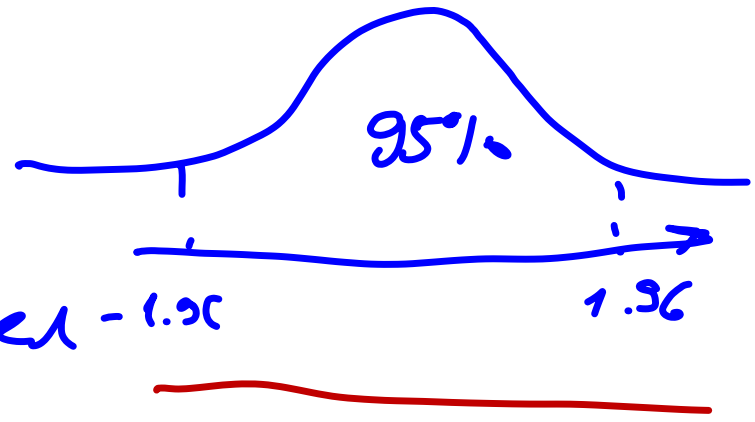
нормируем x_1, x_2

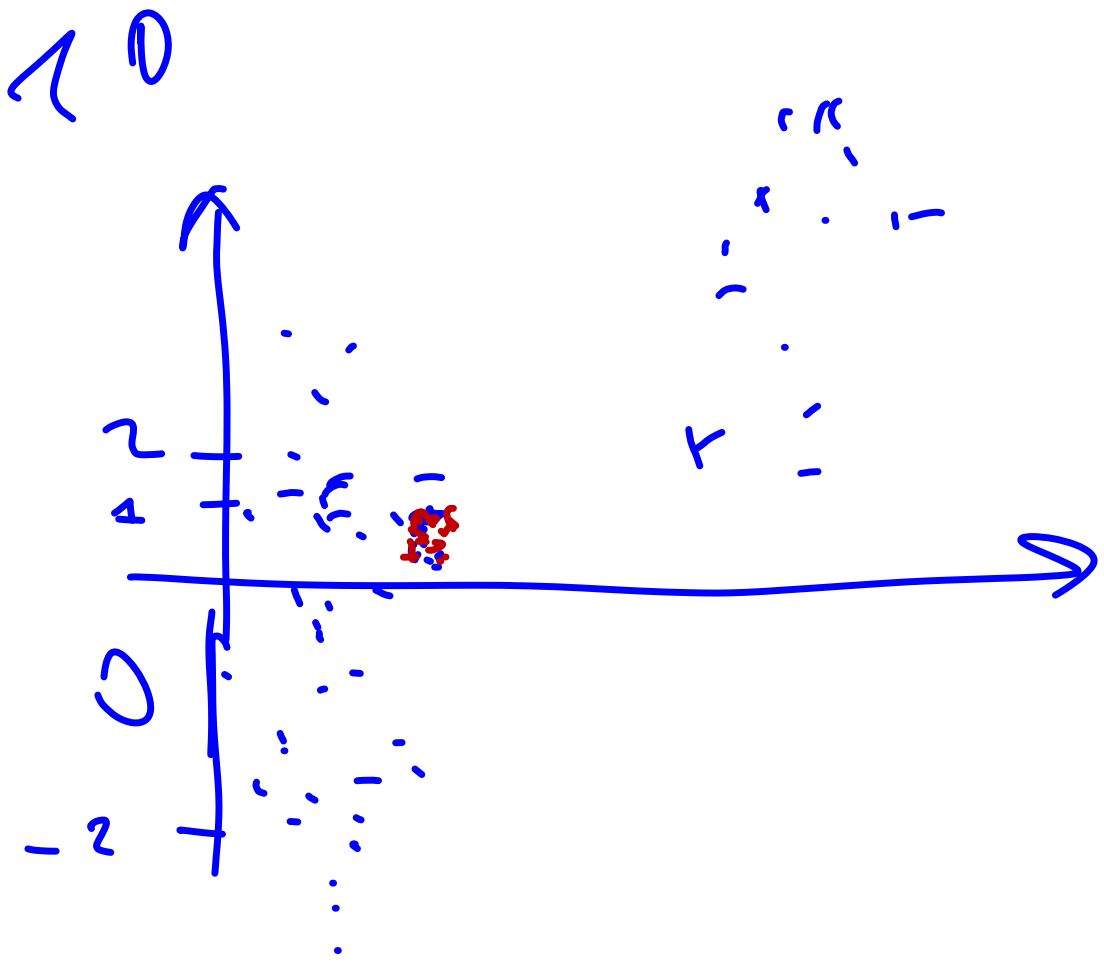
$0 \leq x_1 \leq 1$

$0 \leq x_2 \leq 1$

---

① $x' = \frac{x - \bar{x}}{\hat{\sigma}}$ - ст. откл





② min - max

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \longrightarrow [0; 1]$$

$$x' (b - a) + a \longrightarrow [a, b]$$

Кросс - валидация.

① подбор обучающей и тестовой выборки

② настройка гипер параметров

а) k - fold CV

б) LOO CV
Leave - one - out

Набор данных $D = \{D_1, D_2 \dots D_k\}$
k частей равного размера

$$\bigcup_i D_i = D$$

$$D_i \cap D_j = \varnothing \quad \forall i,j = \{1 \dots k\} \; i \neq j$$

Алгоритм k-Fold

① Случайно разбить выборку на k частей. — "—

② для k от 1 до K повторить

Test set ← D_k

Train ← $D \setminus D_k$

Обучить модель на Train

Посч. ошибку для D_k и запомнить Error$_k$

$$Error = \frac{1}{k} \sum_{i=1}^{k} Error_i$$

↑ итоговая ошибка (кросс-вал.) данной модели

LOOCV (когда мало данных)

① для k от 1 до N ← |D| = N (размер выборки)

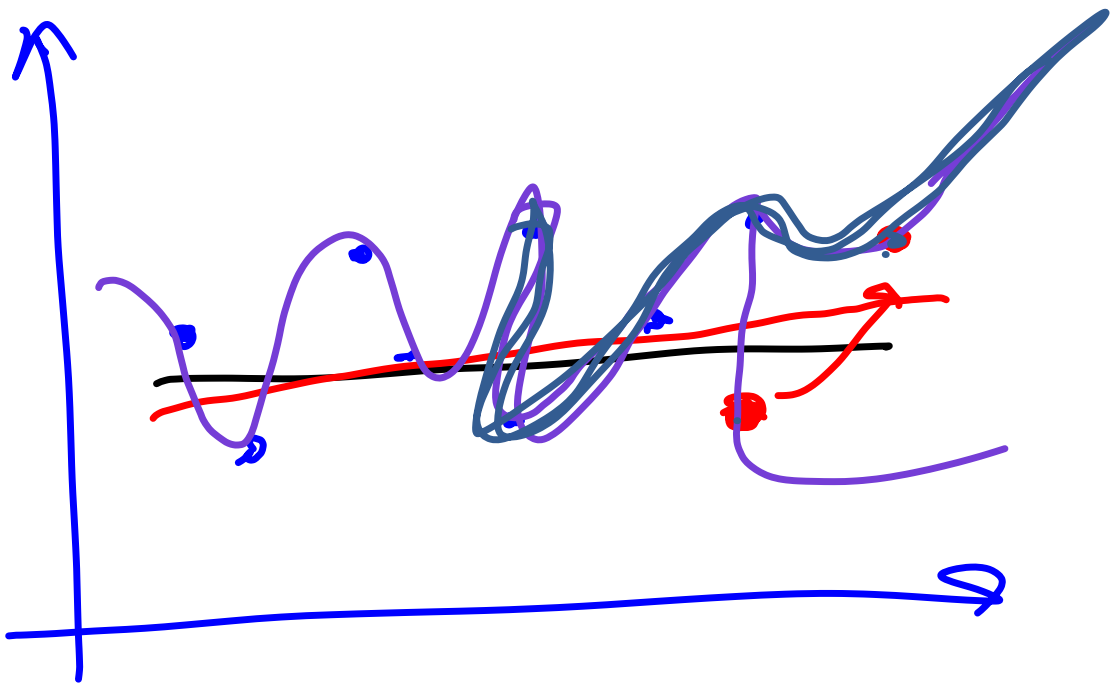
Test ← k-й элемент выборки

Train ← D \ k-й эл.

Об. модель на Train.

— " — — " — — "

$$Error = \frac{1}{N} \sum_{k=1}^{N} Error_k$$



Настройка гиперпар.

настраиваем гиперпар.

X_{train} (60), X_{test} (20), $X_{validation}$ (20)