# ML ERASMUS, NOV 7, 2022

$$\mathbb{E}[x] = \int_D dx\, x\, p(x) = \mu_x$$

$$\left( \sum_{i \in D} x_i\, p(x_i) \right)$$

mean of the sample

$$\overline{\mu_x} = \frac{1}{n} \sum_{i=1}^{n} x_i \neq \mu_x$$

$$MSE = \mathbb{E}\left[(y-\tilde{y})^2\right]$$

$$= \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \tilde{y}_i)^2$$

Resampling methods aim at arriving at a "reliable" estimate of various expectation values.

Bootstrap also

$$D = \{x_0, x_1, \dots, x_{n-1}\}$$

(i) calculate $\mu_x$

(ii) Pick $x$ randomly with replacement

$D^* = \{ x_0^* \, x_1^* , \dots x_{n-1}^* \}$

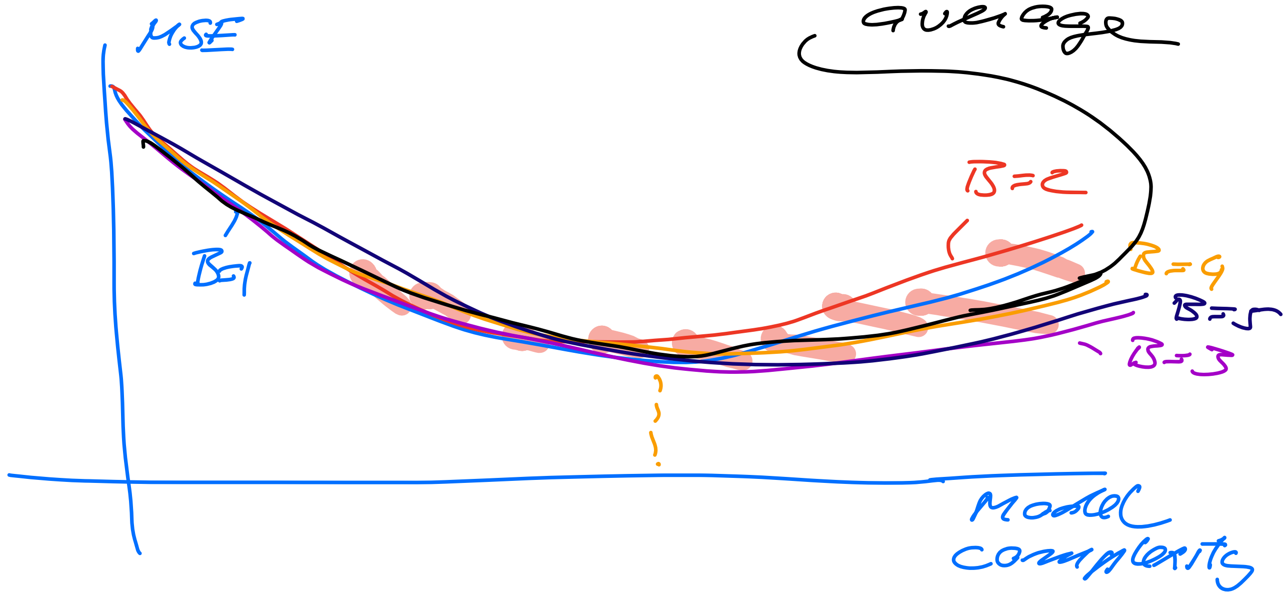
calculate $\mu_x^*$

(iii) repeat (ii) B-times

(iv) calculate final

$$\mu = \frac{1}{B} \sum_{j=0}^{B-1} \mu_x^*(j)$$

suppose $B = 5$ and we calculate MSE (Test data)




Cross-validation (CV)

- Folds K

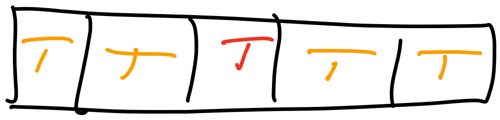
$K = 5$

| | | | | |
|---|---|---|---|---|
| T | T | T | T | T |

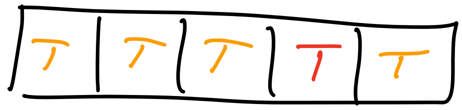
TEST

TRAIN

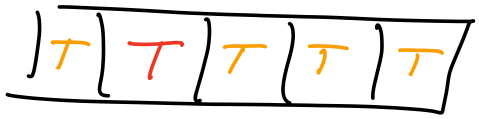
$MSE_1$



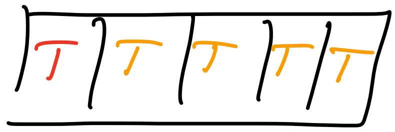

$MSE_2$ on test




$MSE_3$ on test

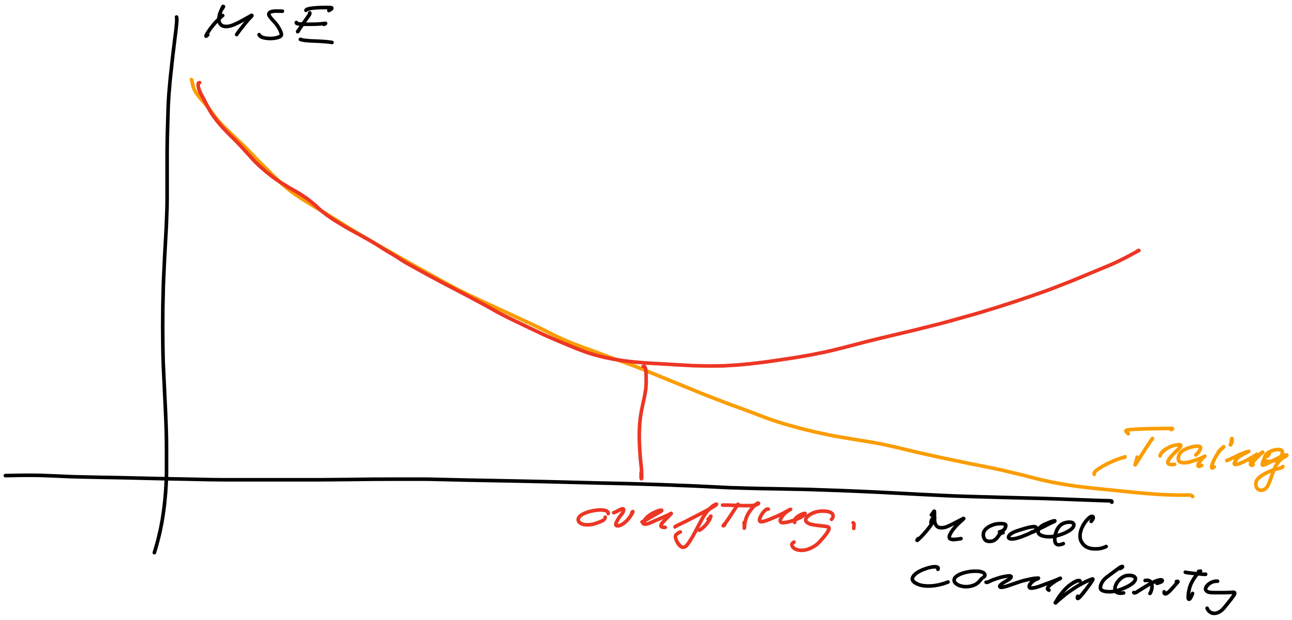


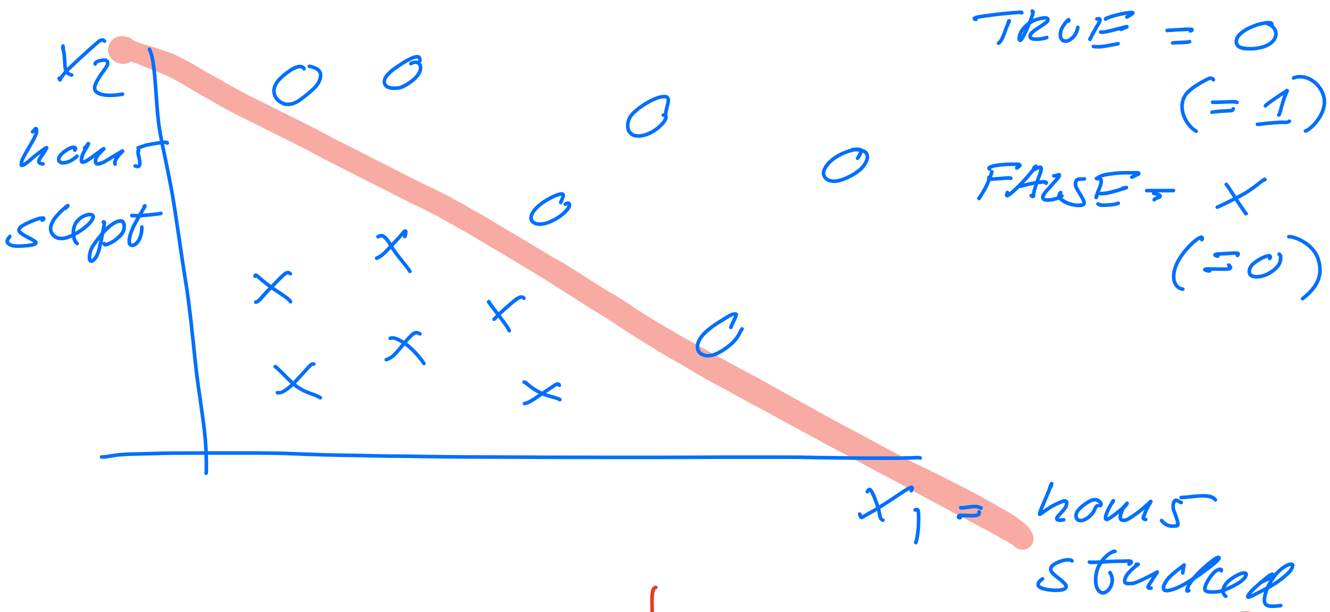

$MSE_4$

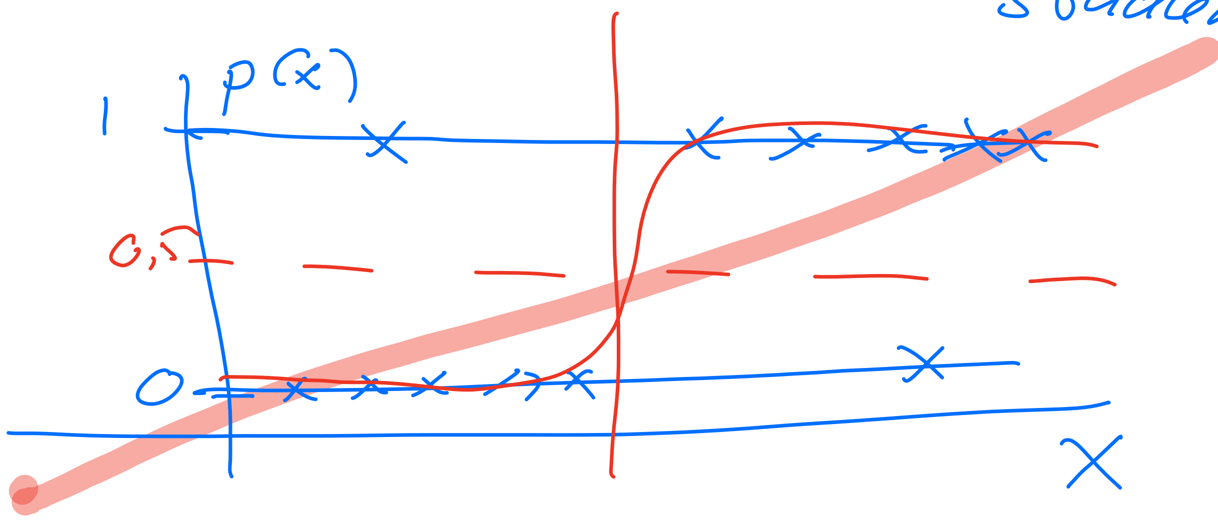



$MSE_5$ $K=5$

$$MSE = \frac{1}{5} \sum_{i=1}^{5} MSE_i$$






Regression

$$y = f(x) + \varepsilon \qquad y \in (-\infty, \infty)$$

Classification

Binary

$$y = p(x) + \varepsilon \qquad y \in \{0, 1\}$$

$$p(x) \in [0, 1]$$

$$\int_D p(x)\,dx = \int_0^1 p(x)\,dx = \underline{1}$$

$$x \in [0, 1]$$

$$p(x) = \frac{1}{1+e^{+x}}$$

$$p(x) \longrightarrow p(y_i, x_i | \beta)$$

$$= \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

parameters

$$p(y | x; \beta) \longrightarrow P_i$$

$$= p(y_i = \underline{1} | x_i; \beta)$$

$$p(y_i = 0 | x_i; \beta) = 1 - P_i$$

$$p(y_i = 1) + p(y_i = 0) = \underline{1}$$

$$= P_i + 1 - P_i$$

$$D = \{(x_0, y_0), (x_1, y_1) \dots (x_{n-1}, y_{n-1})\}$$

$P(D|\beta)$

Assumption $y_i$ are i.i.d.

$$P(D/\beta) = \prod_{i=0}^{n-1} P_i^{y_i} [1-P_i]^{1-y_i}$$

To find $\beta$, what should we aim at in the optimization of $P(D/\beta)$?

— max $P(D/\beta)$

min $P(D|\beta)$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\max}\, P(D/\beta)$$

$$\Rightarrow \quad \frac{\partial P(D/\beta)}{\partial \beta} = 0$$

$$\frac{\partial \log(P(D/\beta))}{\partial \beta} = 0$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \underbrace{\left[-\log(P(D|\beta))\right]}_{C(\beta)}$$

$$C(\beta) = -\sum_{i=0}^{n-1} \left[ y_i \, e^{\beta_0 + \beta_1 x_i} - \log(1 + e^{\beta_0 + \beta_1 x_i}) \right]$$

$$\frac{\partial C}{\partial \beta_0} = 0 = -\sum_{i=0}^{n-1} (y_i - p(x_i))$$

$$p(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$\frac{\partial C}{\partial \beta_1} = 0 = -\sum_{i=0}^{n-1} x_i (y_i - p(x_i))$$

$$\frac{\partial C}{\partial \beta} = 0 = -X^T (y - p)$$

$$y^T = [y_0, y_1, y_2 \ldots y_{n-1}]$$
$$\in \mathbb{R}^n$$

$$p^T = [p_0, p_1, \ldots, p_{n-1}] \in \mathbb{R}^n$$

$$X \in \mathbb{R}^{n \times p}$$

$$\frac{\partial C}{\partial \beta} \in \mathbb{R}^p$$

OLS

$$\frac{\partial C}{\partial \beta} = 0 = -\frac{2}{n} X^T (y - X\beta)$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

linear dependence on $\beta$.

Here:

$$\frac{\partial C}{\partial \beta} = X^T (p - y) = 0$$

non-linear dependence on $\beta$.

$$\frac{\partial^2 C}{\partial \beta \partial \beta^T} = X^T W X = H$$

$$W_{ii} = p(x_i)(1 - p(x_i))$$

W is a diagonal matrix

Roots of $\frac{\partial C}{\partial \beta} = 0$

Newton-Raphson's method

$$f(s) = 0$$

Taylor expand around

$$f(s + \Delta x)$$

---

X

---

$$\beta^{new} = \beta^{old} - \underbrace{\left(H(\beta^{old})\right)^{-1}}_{} g(\beta^{old})$$

$\gamma$ = learning rate.

ADAgrad

RMSprop

Stochastic Gradient

descent

ADAM

⋮