ML ERASMUS, OCT 31, 2022

Resampling methods

With a limited data set, we want to give a reliable estimate of various expectation values.

Sample mean

$$\mathbb{E}[y] = \frac{1}{n}\sum_{i=0}^{n-1} y_i \quad (\neq \mu)$$

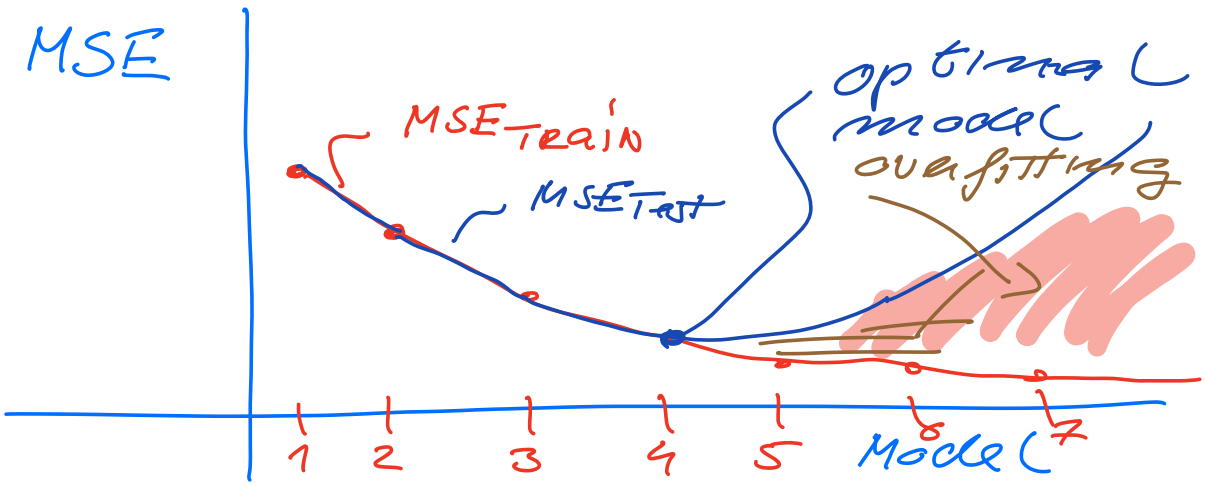
(True mean value

$$\mu = \int_D p(y)\, y\, dy \;)$$

$$\mathrm{Var}[y] = \frac{1}{n}\sum_{i=0}^{n-1} (y_i - \mathbb{E}[y])^2$$

$$MSE = C(\beta) = \frac{1}{n}(y - X\beta)^T (y - X\beta)$$

$$= \frac{1}{n}\sum_{i=0}^{n-1} \left(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j\right)^2$$

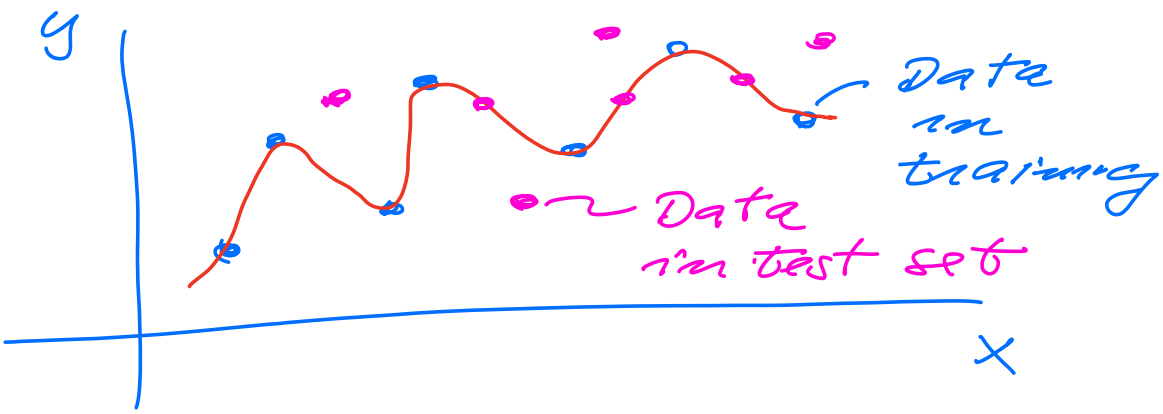


OLS

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

$$\tilde{y} = X\hat{\beta} = ?$$

$$\tilde{y}_i = \sum_{j=0}^{p-1} x_{ij} \beta_j$$



Aim with resampling methods is to provide an "accurate" estimate of

for example the MSE, variance and mean values',

- Bootstrap method

Basics steps :

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

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

1) compute for example μ (sample) and var

$\mu = \frac{1}{n} \sum_{i=0}^{n-1} x_i$ / var $= \sigma^2 = \frac{1}{n} \sum_{i=0}^{n-1} (x_i - \mu)^2$

2) reshuffle D randomly with replacement

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

$D = \{1, 2, 3, 4, 5\}$

$D^* = \{2, 2, 5, 4, 1\}$

compute new μ and var

$$\mu^* = \frac{1}{n} \sum_{i=0}^{n-1} x_i^*$$

$$var^* = \frac{1}{n} \sum_{i=0}^{n-1} (x_i^* - \mu^*)$$

3) Repeat 2) M times and store μ^*, var^*

4) Compute final $\bar{\mu}$

$$\bar{\mu} = \frac{1}{M} \sum_{i=0}^{M-1} \mu_i^*$$

$$var = \frac{1}{M} \sum_{i=0}^{M-1} (\mu_i^* - \bar{\mu})^2$$

For iid events this will approach true var and $\bar{\mu}$

MSE after M-bootstraps

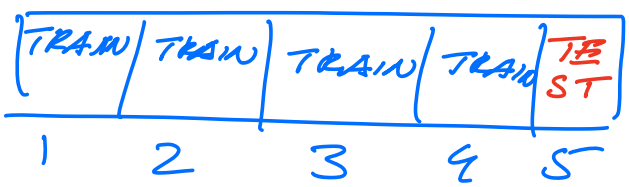
$$MSE = \frac{1}{M} \sum_{i=0}^{M-1} MSE_i^*$$

MSE_{Train} and MSE_{Test}.

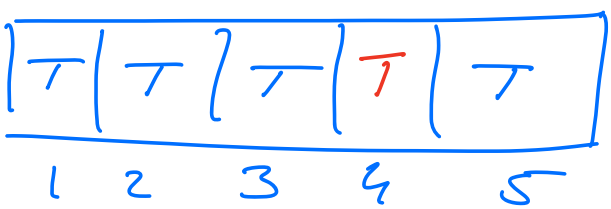
Cross-validation

Split data in k-folds (subdivision)

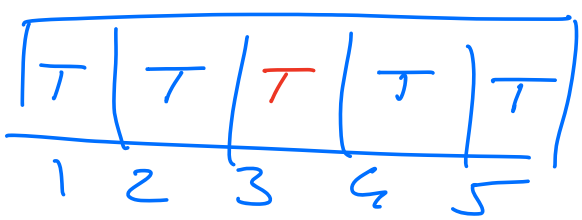
$K = 5$ new parameter

1) 

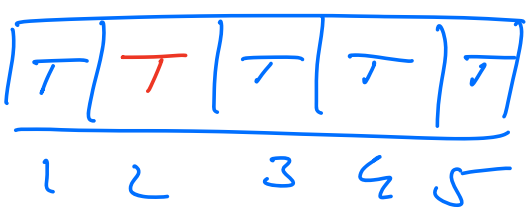
Train = 1, 2, 3, 4

2) 

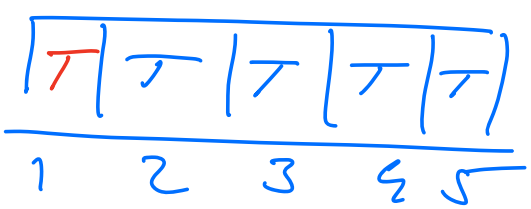
Train = 1, 2, 3 & 5

3) 

Train = 1, 2, 4, 5

4) 

Train 1, 3, 4 & 5

5) 

Train 2, 3, 4, 5

$$MSE_{Train} = \frac{1}{5} \sum_{i=0}^{4} MSE(i)_{TRAIN}$$

$$MSE_{Test} = \frac{1}{5} \sum_{i=0}^{4} MSE(i)_{Test}$$