5 Cross-validation

Data set

train 50%   validate 25%   test 25%

25% → Model selection

25% → Model assessment
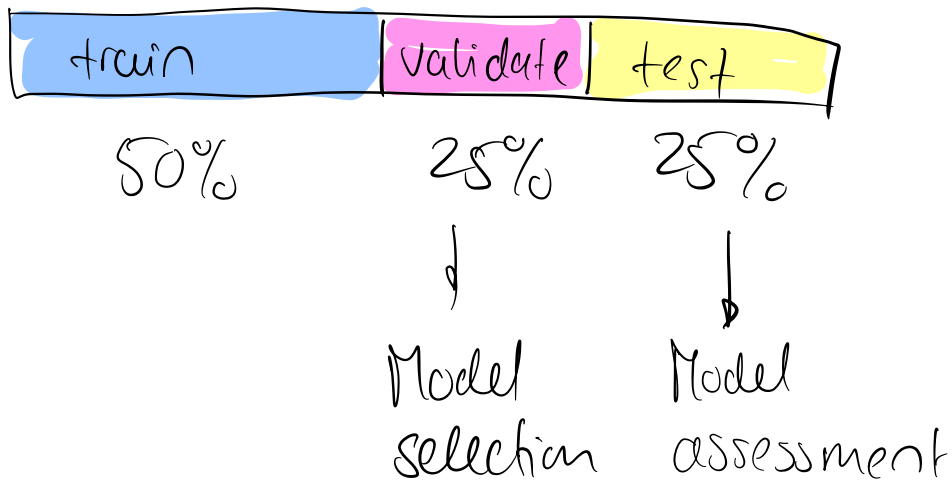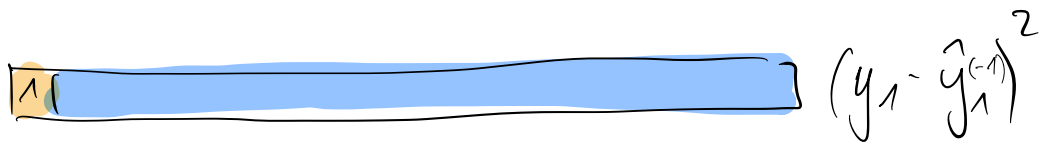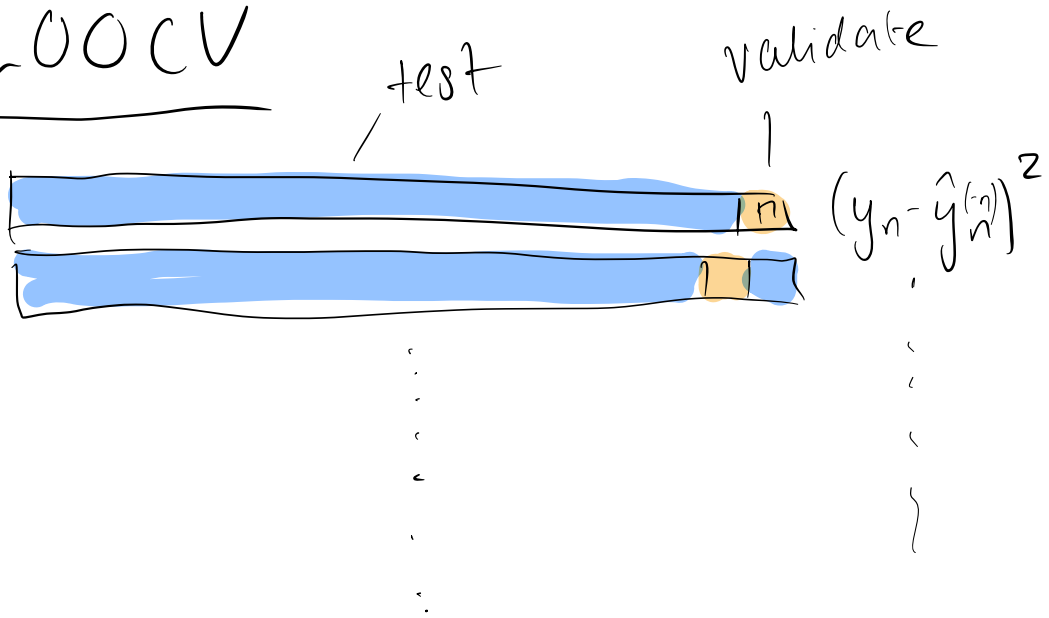
Data-rich situation (often not the case!)

test 25%

75%
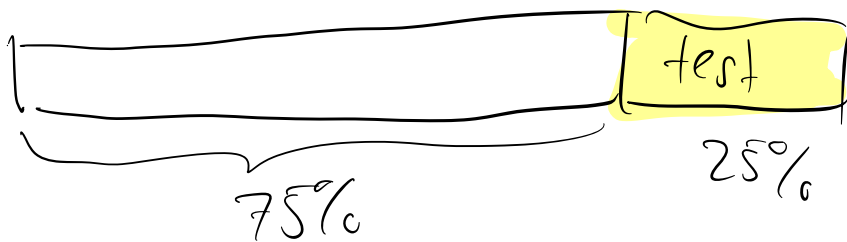
can be used smarter

LOOCV

test　　　　　　validate

$$\left(y_n - \hat{y}_n^{(-n)}\right)^2$$

$$\left(y_1 - \hat{y}_1^{(-1)}\right)^2$$

$$MSE_{LOOCV} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

k-fold CV (k=5 in this case)

$$MSE_1 = \frac{1}{n_1}\sum_{i \in C_1}(y_i - \hat{y}_i)^2$$

$$MSE_5 = \frac{1}{n_5}\sum_{i \in C_5}(y_i - \hat{y}_i)^2$$
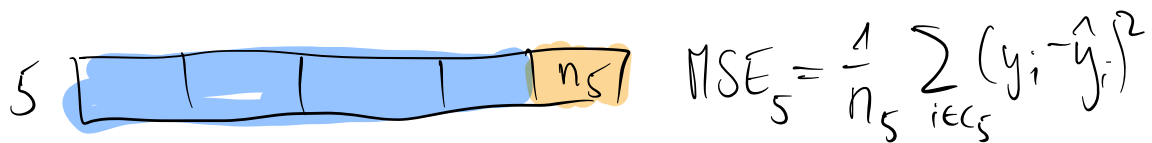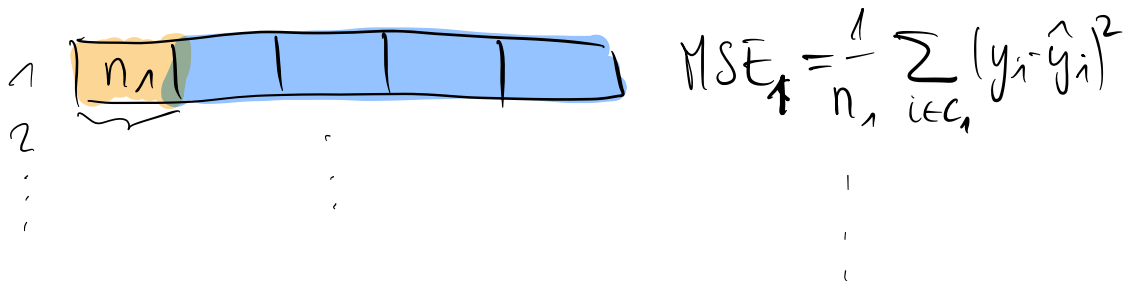
C_i = indices in fold i

$$CV_5 = \frac{1}{n} \sum_{j=1}^{5} n_j \, MSE_j$$

$$\left(\text{if } n_1 = \dots = n_5 : \right.$$

$$CV_5 = \frac{1}{5} \sum_{j=1}^{k} MSE_j \left. \right)$$

CV

$\bar{q} \pm 1 \, SE$
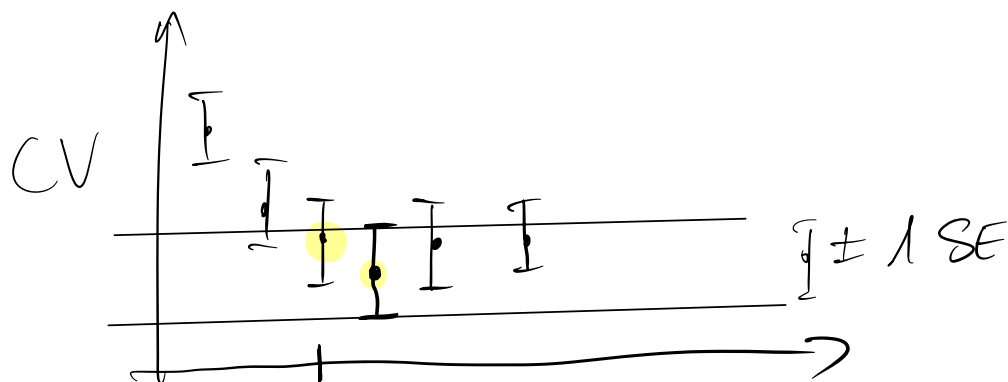
$\ominus$ (Complexity

Nested CV:

CV

test

test
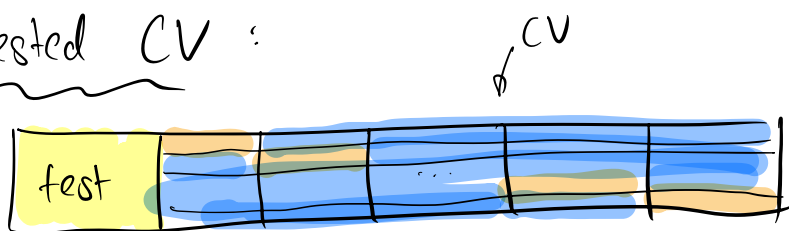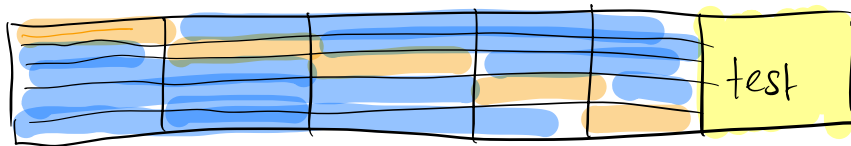
The Bootstrap

Idea: $\hat{P}(x) = \frac{1}{n}$

(empirical distribution)