Lecture September 17

Resampling techniques

- Bootstrap (jackknife)
- cross-validation

$D = [x_1, x_2, \ldots x_n]$

(i) Draw a sample with replacement (randomly)

$x_1^*, x_2^*, \ldots x_n^*$

Compute $\beta_{n,1}^* = g(x_1^* x_2^* \ldots x_n^*)$

$\left[\mu_{n,1}^* = \frac{1}{n} \sum_{i=1}^{n} x_i^* \right]$

(ii) Repeat previous steps B times

$\beta_{n,1}^*, \beta_{n,2}^* \ldots \beta_{n,B}^*$

$(\mu_{n,1}^* \ldots)$

(iii) STD

$$S = \sqrt{\frac{1}{B}\sum_{j=1}^{B}(\bar{\beta}^*_{m,j} - \bar{\beta})^2}$$

$$\bar{\beta} = \frac{1}{B}\sum_{j=1}^{B}\beta^*_{m,j}$$

First example

$$X \sim N(100, 15^2)$$

$$z = \frac{\bar{x}_1 + \bar{x}_2 + \cdots \bar{x}_m}{m} = \frac{m\bar{\mu}}{m}$$

$$= \mu$$

$X \sim N(\ldots)$

central limit theorem

$$z = \mu$$

$$\sigma_z^2 = \frac{\sigma^2}{m} \Rightarrow STD = \sigma/\sqrt{m}$$

in the code

$$STD = \sigma/\sqrt{B}$$

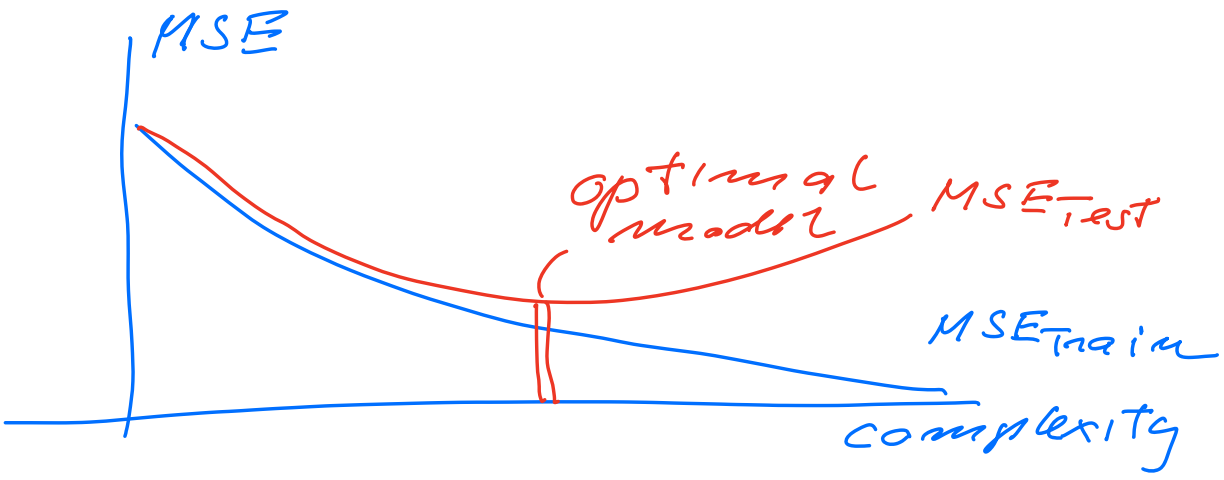
Bias - variance Tradeoff

$$y_i = f(x_i) + \varepsilon_i$$

$$f(x_i) \approx \underset{\text{Model}}{\tilde{y}_i} \quad (x_i * \beta)$$

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

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



Bootstrap to improve (or make reliable prediction of MSE_{Test})

Bias-variance tradeoff:

$$\sum_{i=0}^{n-1} (\quad \quad)^2$$

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

$$\left(\mathbb{E}[\tilde{y}_i] = X_{i*}\beta = \tilde{y}_i \right.$$

$$\operatorname{var}[\tilde{y}_i] = \operatorname{var}[y_i] =$$

$$\operatorname{var}[\varepsilon_i] = \sigma^2$$

$$\mathbb{E}[\hat{\beta}] = \hat{\beta} \qquad \operatorname{var}[\hat{\beta}]_{OLS} = \sigma^2 (X^T X)^{-1} \left. \right)$$

$$= \text{Bias} + \text{variance} + \sigma^2$$

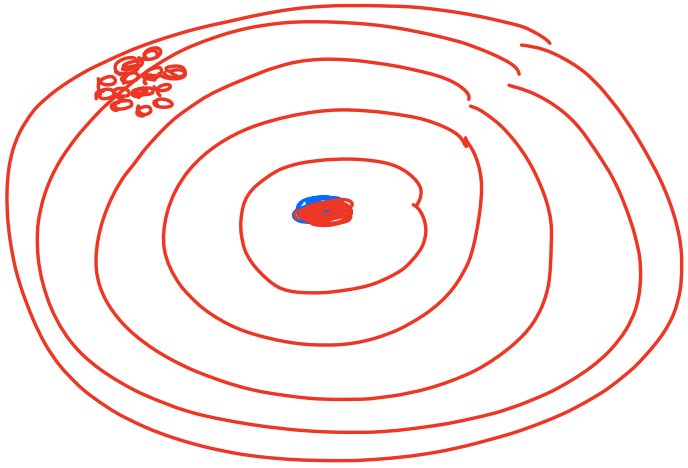
$$\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \mathbb{E}[\tilde{y}])^2$$

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

$$\mathbb{E}[\tilde{y}] = \frac{1}{n}\sum_{i=0}^{n-1}\tilde{y}_i$$

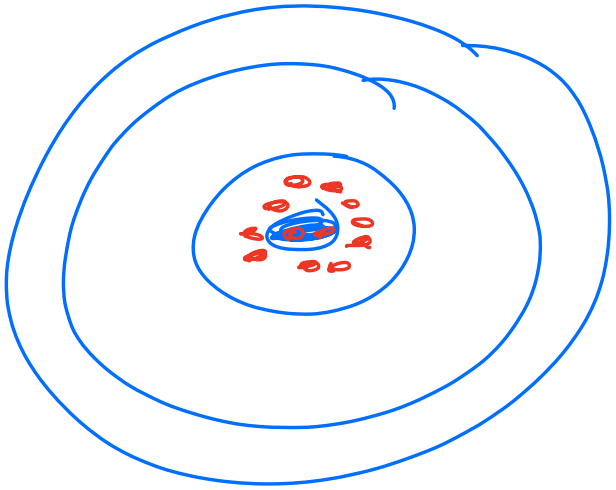
Low variance

High Bias



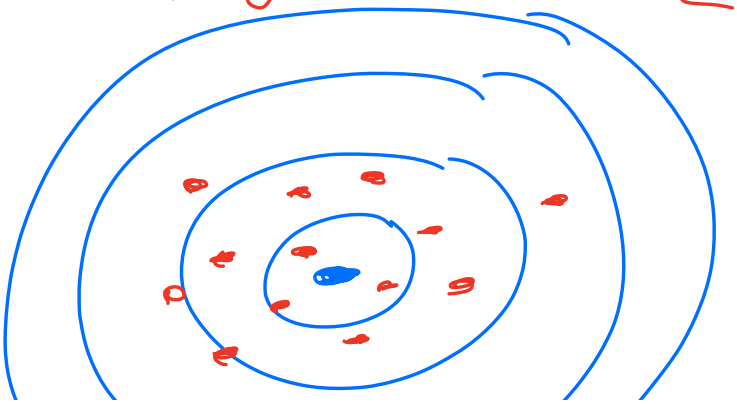
Low variance

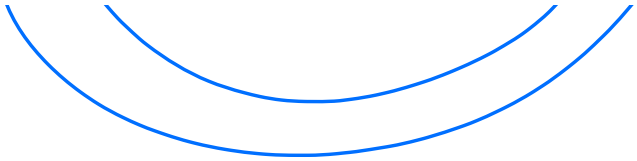
Low Bias



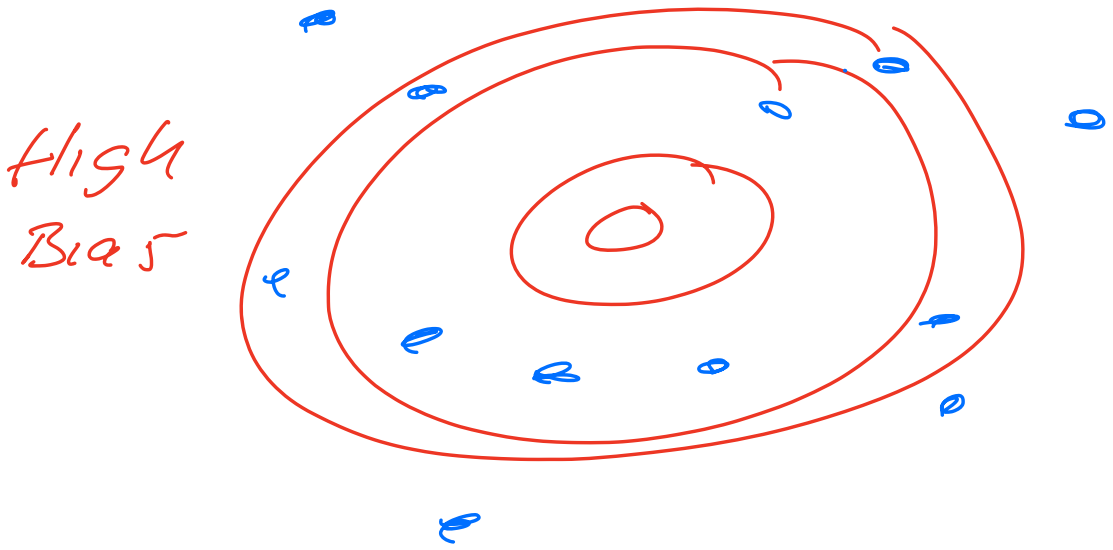
High variance

Low Bias





High variance



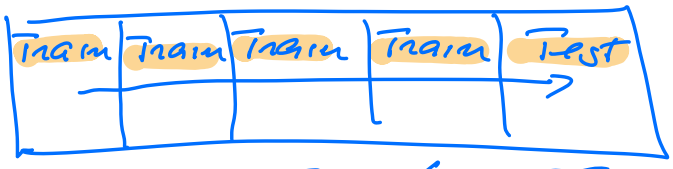
High Bias

Cross-validation (CV)

k-fold CV

$k = 5$

(1)

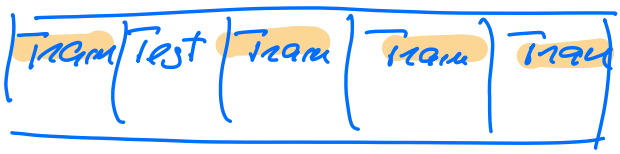


Data set

1 2 3 4 5

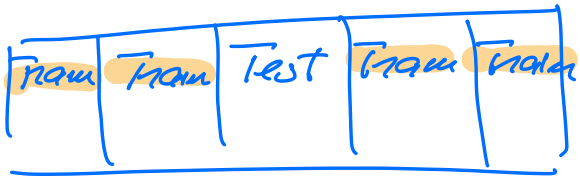
Testen 1

(2)



$Test_{err_2}$

(3)



$Test_{err_3}$

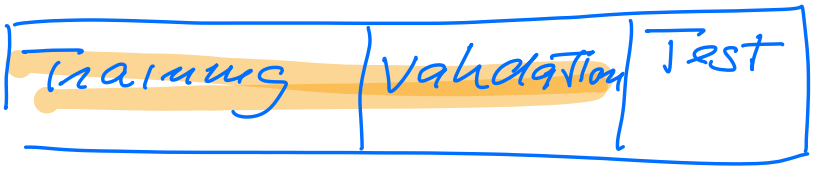
⋮

(5)

$$Err = \frac{1}{k} \sum_{i=1}^{k} Test_{err_i}$$

$k = 5 - 10$

Splitting of data



$k = n$ (number of data)

LOCV

$$x = [x_1 x_2 \ldots x_m]$$

$$y = [y_1 y_2 \ldots y_m]$$

for $i = 1, m$

$$x_{cv} = [x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m]$$

$$y_{cv} = [y_1 y_2 \ldots y_{i-1} y_{i+1} \ldots y_m]$$

$$x_{out} = x_i$$

$$\tilde{y}_{out} = f_{it}(x_{cv}, y_{cv}, x_{out})$$