Lecture September 3

$$y_i = f(x_i) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

Model

$$y_i = X_{i*}\beta + \varepsilon_i = \tilde{y}_i + \varepsilon_i$$

$$\rightarrow \mathbb{E}[y_i] = X_{i*}\beta \qquad \beta^T = [\beta_0, \beta_1 \ldots \beta_{p-1}]$$

$$\beta \in \mathbb{R}^p$$

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

$$y \in \mathbb{R}^n$$

$$\rightarrow \mathrm{var}(y_i) = \sigma^2$$

$$\boxed{y_i \sim N(X_{i*}\beta, \sigma^2)}$$

$$\mathrm{var}(\hat{\beta}), \quad \mathbb{E}[\hat{\beta}]$$

$$\rightarrow \mathbb{E}[\hat{\beta}] = \beta \qquad \left(\hat{\beta} = (X^T X)^{-1} X^T y\right)$$

$$\mathrm{var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

$$\rightarrow \mathrm{var}(\hat{\beta}_j) = \sigma^2 (X^T X)^{-1}_{jj}$$

Baye's theorem

$$P(B|A) = \frac{P(A|B)P(B)}{\rightarrow \sum P(B=k)P(A|B=k)}$$

$p(A) = \frac{\;}{b}$

Assumption:

$$\boxed{P(y_i \mid x_{i*}\beta)} = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y_i - x_{i*}\beta)^2}{2\sigma^2}}$$

$$P(D \mid X, \beta) = \prod_{i=0}^{n-1} P(y_i \mid x_{i*}\beta)$$

MLE

$$= \underline{\sum_i \log P(y_i \mid x_{i*}\beta)}$$

$$\hat{\beta} = \arg\max_{\beta \in \mathbb{R}^P} \qquad = \;\_\_\_$$

or

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^P} \; -\sum_i \log P(y_i \mid x_{i*}\beta)$$

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

Bias-variance trade-off

$\mathbb{E}[\tilde{y}]$ $\qquad + \mathbb{E}[\tilde{y}] - \mathbb{E}[\tilde{y}]$

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

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

Bias

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

variance of $\tilde{y}$

noise variance.

$$\text{OLS} = \frac{p}{n}\sigma^2$$

Model Assessment and Selection (chapter 7 of Hastie et al)

- Bias-variance trade-off
- Discussion of errors
- Resampling techniques
  - Bootstrap
  - cross-validation

We have a cost/loss function

$$C(\hat{\beta}) = \frac{1}{n}\sum_i (y_i - \tilde{y}_i)^2$$

Test Error for the test data $\boxed{T}$ with a given training set

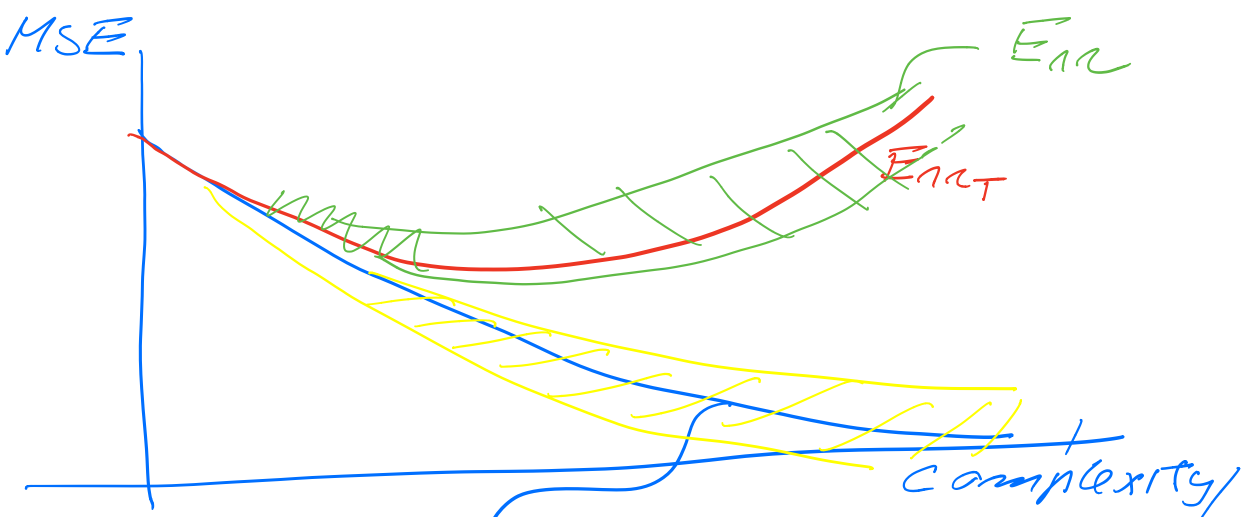
-

$$Err_T = E[L(\hat{\beta}), T]$$

Define a better estimate of $Err_T \Rightarrow \boxed{Err} = \frac{1}{B} \sum_{i=1}^{B} Err_{T_i}$

Making B - samples for training (Err_{T_i})

prediction/test Error. The test Error is more to a statistical interpretation.



$$Err_{Train} = \frac{1}{n_{Train}} \sum_i (y_i - \hat{y}_i)^2$$

$T \leftarrow$

MODEL SELECTION

estimate the performance

of different models in order to pick the best one

MODEL ASSESSMENT with a final model, estimate the prediction error on new data

TRAIN 50-60%	validate ~20%	TEST 20%

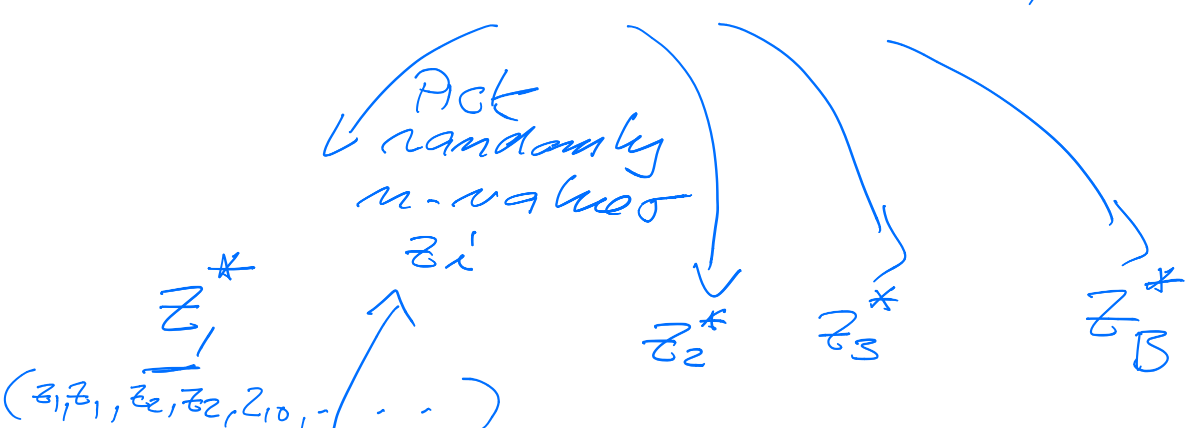
divided Randomly,

Resampling methods

Bootstrap

Training sample

$$Z = (z_1, z_2, \dots z_n)$$



can have same
z_i more than
once

For every sample (training)
we can compute some expected
value, $S(z_1^*)$, $S(z_2^*)$ ...,
$S(z_B^*)$

$$var(S(z)) = \frac{1}{B-1} \sum_{i=1}^{B} \left(S(z_i^*) - \bar{S}^* \right)^2$$

B = # of resamples