Notes Tuesday July 20

output $y^T = [y_0, y_1 \dots y_{n-1}]$

input $x^T = [x_0, x_1 \dots x_{n-1}]$

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

Model: Example polynomial of degree $p-1$

$$y(x) \longrightarrow y(x_i) = y_i = f(x_i) + \varepsilon_i$$

assumption

ε_i error $\sim N(0, \sigma_i^2)$

$\mu = 0$

$\sigma_i^2 = \sigma^2$

$$f(x_i) \simeq \tilde{y}(x_i) = \tilde{y}_i$$

$$= \sum_{j=0}^{p-1} \beta_j x_i^j$$

$$= \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \dots + \beta_{p-1} x_i^{p-1}$$

$$\tilde{y}_0 = \beta_0 + \beta_1 x_0^1 + \beta_2 x_0^2 + \dots \beta_{p-1} x_0^{p-1}$$

$$\tilde{y}_1 = \beta_0 + \beta_1 x_1^1 + \beta_2 x_1^2 + \dots$$

$$\vdots$$

$$\tilde{y}_{n-1} = \beta_0 + \beta_1 x_{n-1}^1 + \beta_2 x_{n-1}^2 + \dots$$

$$\tilde{y}^T = [\tilde{y}_0, \tilde{y}_1, \dots \tilde{y}_{n-1}]$$

$$X = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^{p-1} \\ 1 & x_1 & x_1^2 & \dots & \\ 1 & x_2 & x_2^2 & \dots & \\ \vdots & \vdots & \vdots & \dots & \\ 1 & x_{n-1} & x_{n-1}^2 & \dots & x_{n-1}^{p-1} \end{bmatrix}$$

x_{00} x_{10}

$$\beta^T = [\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_{p-1}]$$

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

$$p = n-1 \qquad X \in {}^{n \times p}$$

Linear Regression $n \gg p$

How to assess the quality of the model?

Measure: Mean square error

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

$$= \frac{1}{n}\left\{(y-\hat{y})^T(y-\hat{y})\right\}$$

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

(cost/loss/error/risk ... function)

$$\min_{\beta \in R^p} \frac{1}{n}\left((y-X\beta)^T(y-X\beta)\right)$$

$$\frac{\partial MSE(\beta)}{\partial \beta_j} = 0$$

$$\frac{\partial}{\partial \beta_j}\left[\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \beta_0 x_{i0} - \beta_1 x_{i1} - \cdots \underset{\beta_j x_{ij}}{\cdots} - \beta_{p-1} x_{ip-1})^2\right]$$

$$= 0$$

$$= -\frac{2}{n}\sum_{i=0}^{n-1} x_{ij}(y_i - \beta_0 x_{i0} - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{ip-1})$$

Or in matrix-vector form ($x - \frac{n}{2}$)

$$0 = X^T(y - X\beta)$$

$$X^T X \beta^{opt} = X^T y \implies$$

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

if we can invert problem solved.

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

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

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

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

$$X^T y \in \mathbb{R}^p$$

statistical quantities

mean (continuous PDF $p(x)$)

$$\mu = \int p(x)\, x\, dx$$

$$\sigma^2 = \int p(x)(x-\mu)^2 dx$$

Discrete $p(x) \to p(x_i) = p_i$

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

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

sample mean/variance

$$\underline{\bar{\mu}} = \frac{1}{n} \sum x_i \neq \mu$$

$$\underline{\bar{\sigma}}^2 = \frac{1}{n} \sum (x_i - \bar{\mu})^2 \neq \sigma^2$$

Linear regression

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

$\sim N(0, \sigma^2)$

our model
(deterministic)

$N(x_i * \beta, \sigma^2)$

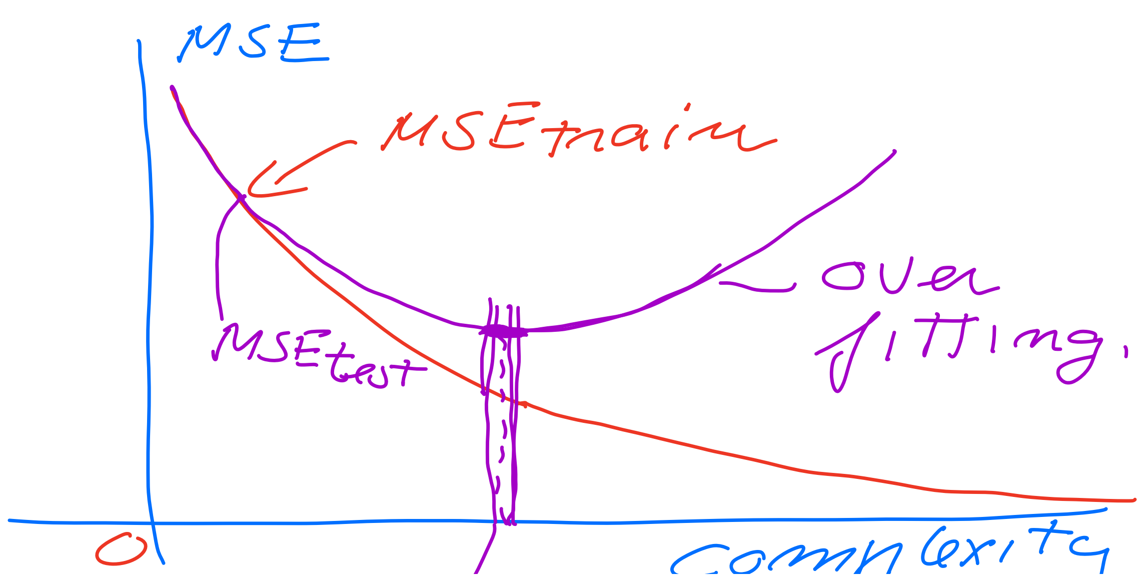
$$\boxed{MSE(\beta)} = \frac{1}{n}\sum_{i=0}^{n-1}(y_i - \tilde{y}_i)^2$$

$$= \frac{1}{n}\sum_{i=0}^{n-1}\underline{(y_i - \underline{x_i} * \beta)^2}$$

Bias-variance trade off.

- split data in train and test.

MSE train MSE Test



optimal model   of model

How do we get the best estimate of MSE_{train} and MSE_{test}?

$\Rightarrow$ Resampling

- Bootstrap
- cross validation

1) Train and test. compute MSE_1.
2) Reshuffle train data

1	5	5	51	..	70	70

$n=100$

MSE_2

3) Repeat MSE_3

⋮ repeat b-times

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