Erasmus+ course on machine learning, September 25, 2023

1) input & output data

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

$$X = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_{n-1} \end{bmatrix} \qquad y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \end{bmatrix}$$

input

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

output

2) Regression, basic assumption that there exists a continuous function $f(x)$, which is non-stochastic

$$y = f(x) + \varepsilon$$

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

$$\varepsilon_i \sim N(0, \sigma^2)$$ ε_i are i.i.d.

Model : $f(x) \simeq \tilde{y}$

Example $\tilde{y}(x_i) = \tilde{y}_i = \sum_{j=0}^{p-1} \beta_j x_i^j$

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

3) Objective function / cost/loss function to find

optimal β — $\hat{\beta}$ — which minimize the cost function

$C(\hat{\beta})$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{p-1}}{\arg\min}\ C(\beta)$$

Example: Mean square Error

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

$$\tilde{y}_i = \tilde{g}(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{p-1} x_i^{p-1}$$

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

$$\tilde{y}_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \cdots + \beta_{p-1} x_1^{p-1}$$

$$\vdots$$

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

$$\beta^T = [\beta_0 \ \beta_1 \ \beta_2 \ \cdots \ \beta_{p-1}] \in \mathbb{R}^p \qquad p \leq n$$

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

Design/feature matrix

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

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

$$\frac{\partial C(\beta)}{\partial \beta} = 0$$

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

$$\frac{\partial C}{\partial \beta} = -\frac{2}{n} X^T (y - X\beta) = 0$$

(Derivation in exercise 1)

$$\Rightarrow \quad X^T y = X^T X \beta \quad \Rightarrow$$

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

ordinary least square

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

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

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

$\rightarrow$ $\hat{\beta}$ estimator.

Learning through $(X^T X)^{-1} X^T y$