Section 3C. Linear Regression

Statistics for Data Science

Victor M. Preciado, PhD MIT EECS
Dept of Electrical & Systems Engineering
University of Pennsylvania
preciado@seas.upenn.edu

Linear Regression

- **Additive model**: We assume that the output (random) variable Y follows a linear model:

$$Y = f_L(\mathbf{X}; \beta) + \varepsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

  where:
  - the input (random) variable is drawn from a known distribution $\mathbf{X} \sim f_X$
  - the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are deterministic (but unknown) coefficients
  - the measurement noise follows a distribution $\varepsilon \sim f_\varepsilon$

- **Joint PDF** f_{XY}: Notice that the additive model induces a joint PDF

$$f_{XY}(\mathbf{x}, y) = f_{Y|X}(y|\mathbf{x})\, f_X(\mathbf{x})$$

Linear Regression (cont.)

- **Linear Regression Problem**:
  - Given a training dataset $\mathcal{D}_{\text{Tr}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ consisting of N independent samples $(\mathbf{x}_i, y_i) \sim f_{XY}$
  - Estimate values for the unknown coefficients $\beta_0, \beta_1, \ldots, \beta_p$ denoted by $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$
  - Once we have these estimates, we can make a prediction about the output variable corresponding to a new input $\mathbf{x} = [x_1, \ldots, x_p]^\mathsf{T}$, as follows

$$\widehat{y} = f_L\left(\mathbf{x}; \widehat{\beta}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_p x_p$$

where $\widehat{\beta} = \left[\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p\right]^\mathsf{T}$

Linear Regression (cont.)

- To find the vector of estimated coefficients $\widehat{\beta} = \left[\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p\right]^{\mathsf{T}}$ from $\mathcal{D}_{\text{Tr}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, we solve the following optimization problem:

$$\widehat{\beta} = \arg\min_{\theta} \sum_{i=1}^{N} \left(y_i - f_L\left(\mathbf{x}_i; \theta\right)\right)^2$$

where $\theta = [\theta_0, \theta_1, \ldots, \theta_p]^{\mathsf{T}}$ and the term $e_i = y_i - f_L\left(\mathbf{x}_i; \theta\right)$ is called the i-th *residual*. The summation above is called the *Residual Sum of Squares* (RSS).

Linear Regression (cont.)

- We can write the above equation in matrix form as

$$\widehat{\beta} = \arg\min_{\theta} \|\mathbf{y} - M_X \theta\|^2$$

where

$$M_X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

- The solution to the above optimization problem is given by (without proof)

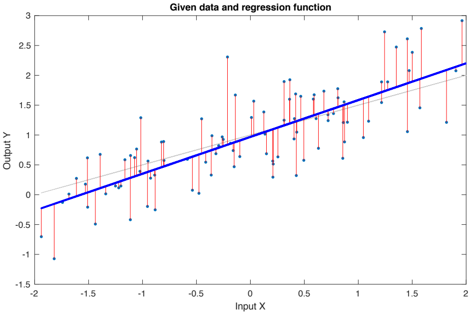
$$\widehat{\beta} = (M_X^\intercal M_X)^{-1} M_X^\intercal \mathbf{y}$$

Linear Regression: Univariate Case

- For the particular case $p = 1$, our training dataset is given by $\mathcal{D}_{\text{Tr}} = \{(x_i, y_i)\}_{i=1}^{N}$ and the linear regression takes the form:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1$$

where $\widehat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$ and $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$ with $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$



Copyright 2020 University of Pennsylvania
No reproduction or distribution without permission.