Linear Regression Analysis
- Maximum Likelihood Approach -

Tseng-Ching James Shen, PhD

Linear Regression Model

- Assumptions
  - ➢w_0 is a random variable from a Gaussian distribution

  - ➢w_1, w_2, … w_d are scalars
- For each dataset point ($\mathbf{x}_i$, y_i) i=1, 2, …, N, y_i can be estimated by the following linear equation:

$$y_i = w^T \mathbf{x}_i + w_0 \quad w_0 \sim N(0,)$$

Training Loss

- Suppose there exists a distribution density function from $f(\mathbf{x}_i)$ which y_i is drawn
- The estimated density function $g(\mathbf{x}_i \,|\, \mathbf{w}, \,)$ is

$$\mathbf{w}^T \mathbf{x}_i + w_0 \quad w_0 \sim N(0, \,)$$

- Use *Kullback-Leibler (KL) divergence* to measure the distance between two density functions, which is also called Training Loss

- The goal is to find the optimal which will minimize the training loss

Likelihood

- In fact, minimizing the training loss is equivalent to *maximizing the likelihood p*(y_1, y_2, … y_N|) which can be expressed as

- Since the Gaussian distribution assumption of $w_{0,}$ will be also a Gaussian distribution

$$N(w^T \mathbf{x_i},)$$

- Therefore, the likelihood equation becomes

- We will find optimal which maximizes the likelihood

Natural Logarithm of Likelihood

- Take the natural logarithm of likelihood

log L = log

()

Finding Optimal **w**

- Take the partial derivative on **w**

- The optimal **w**, is arrived when = 0, so we will get

$= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

where $\mathbf{X} = (\mathbf{x}_{\mathbf{1}}{}^T, \mathbf{x}_{\mathbf{2}}{}^T, \ldots, \mathbf{x}_{\mathbf{N}}{}^T)^T$

Finding Optimal

- Take the partial derivative on **w** =
- The optimal , is arrived when = 0, so we will get

=

Variability in Model Parameters

- For a given set of observed dataset points, we can get one pair of optimal and
- If we use different datasets to train the model, we will get multiple pairs of and
- The potential variability of estimated is encapsulated in the covariance matrix of

$$(\mathbf{X}^{T}\mathbf{X})^{-1}$$

- The diagonal elements tell us how much variability to be expected in the individual parameters
- The off-diagonal elements tell us how parameters co-vary
- Section 2.10.3 uses Olympic data example to illustrate the above points