楊宜樹 D-C1-2828-0 MATH3017-001 Lecture 6

Suppose here is K dimensional

Suppose $g(x, \theta) = \theta_K x^K + ... + \theta_1 x + \theta_0$ with $\theta_K \in \mathbb{R}, ..., \theta_1 \in \mathbb{R}$ and $\theta_0 \in \mathbb{R}$ being unknown parameters to be estimated, therefore,

$$\hat{\theta} = \arg \min_{\substack{\theta=(\theta_0,...,\theta_K)\in\mathbb{R}^{K+1} \\ \theta \in \mathbb{R}^{K+1}}} \left[\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_K x_i^K - ... - \theta_1 x_i - \theta_0)^2\right]. \quad (2)$$

Denote $f_n(\theta_0, ..., \theta_K) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_K x_i^K - ... - \theta_1 x_i - \theta_0)^2$, to find the minimizer of f_n, we differentiate it with respect to $\theta_0, ..., \theta_K$ and let these partial derivatives be 0, i.e.

$$\frac{\partial f_n(\theta)}{\partial \theta_k} = 0, \quad k = 0, ..., K,$$

which read as

$$\sum_{i=1}^{n}(y_i - \theta_K x_i^K - ... - \theta_1 x_i - \theta_0)x_i^k = 0, \quad k = 0, ..., K. \quad (3)$$

Write

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & ... & x_1^K \\ 1 & x_2 & x_2^2 & ... & x_2^K \\ ... & ... & ... & ... & \\ 1 & x_n & x_n^2 & ... & x_n^K \end{bmatrix}$$

Assignment 1: Verify that (3) can be rewritten as a vector equation as

$$\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\theta.$$

where $\mathbf{y} = [y_1, ..., y_n]^\top$ and $\theta = [\theta_0, ..., \theta_K]^\top$.

Assignment I Solution:

$$(3) \Rightarrow \sum_{i=1}^{n}(y_i - \theta_K x_i^K - ... - \theta_1 x_i - \theta_0)x_i^k = 0, k = 0, ..., K$$

$$\Rightarrow \sum_{i=1}^{n} y_i x_i^k = \sum_{i=1}^{n}(\theta_K x_i^K + \theta_{K-1} x_i^{K-1} + \cdots + \theta_1 x_i + \theta_0)x_i^k, k = 0, ..., K$$

Now from (3) we have

$$\begin{bmatrix} \sum_{i=1}^{n} y_i = \sum_{i=1}^{n}(\theta_K x_i^K + \theta_{K-1} x_i^{K-1} + \cdots + \theta_1 x_i + \theta_0)\cdot 1 \\ \sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n}(\theta_K x_i^K + \theta_{K-1} x_i^{K-1} + \cdots + \theta_1 x_i + \theta_0)\cdot x_i \\ \vdots \\ \sum_{i=1}^{n} y_i x_i^K = \sum_{i=1}^{n}(\theta_K \cdot x_i^K + \theta_{K-1} x_i^{K-1} + \cdots + \theta_1 x_i + \theta_0) x_i^K \end{bmatrix} \Rightarrow \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_i \\ \vdots \\ \sum_{i=1}^{n} y_i x_i^K \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n}(\theta_K x_i^K + \theta_{K-1} x_i^{K-1} + \cdots + \theta_1 x_i + \theta_0)\cdot 1 \\ \sum_{i=1}^{n}(\theta_K x_i^K + \theta_{K-1} x_i^{K-1} + \cdots + \theta_1 x_i + \theta_0)\cdot x_i \\ \vdots \\ \sum_{i=1}^{n}(\theta_K \cdot x_i^K + \theta_{K-1} x_i^{K-1} + \cdots + \theta_1 x_i + \theta_0) x_i^K \end{bmatrix}$$

$\Rightarrow X^T \vec{y} = X^T X \vec{\theta}$ and hence shown.

here

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & ... & x_1^K \\ 1 & x_2 & x_2^2 & ... & x_2^K \\ ... & ... & ... & ... & \\ 1 & x_n & x_n^2 & ... & x_n^K \end{bmatrix}$$

and then $X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^K & x_2^K & \cdots & x_n^K \end{bmatrix}$

also $\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $\vec{\theta} = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_K \end{bmatrix}$

L.H.S. is obvious;

R.H.S. is because $X\vec{\theta} = \begin{bmatrix} \sum_{i=1}^{n}(\theta_K x_i^K + \cdots + \theta_1 x_i + \theta_0)\cdot 1 \\ \vdots \\ \sum_{i=1}^{n}(\theta_K x_i^K + \cdots + \theta_1 x_i + \theta_0)\cdot x_i^K \end{bmatrix} \in \mathbb{R}^{K+1}$

thus $X^T(X\vec{\theta}) = \begin{bmatrix} \sum_{i=1}^{n}(\theta_K x_i^K + \theta_{K-1} x_i^{K-1} + \cdots + \theta_1 x_i + \theta_0)\cdot 1 \\ \sum_{i=1}^{n}(\theta_K x_i^K + \theta_{K-1} x_i^{K-1} + \cdots + \theta_1 x_i + \theta_0)\cdot x_i \\ \vdots \\ \sum_{i=1}^{n}(\theta_K \cdot x_i^K + \theta_{K-1} x_i^{K-1} + \cdots + \theta_1 x_i + \theta_0) x_i^K \end{bmatrix} \in \mathbb{R}^{K+1}$

D.

- We consider the linear regression:

$$\mathbf{y} = \mathbf{X}\theta + \epsilon, \quad (6)$$

with the restriction: for some $K > 0$, $\|\theta\| \leq K$.

- By Lagrangian duality, to estimate the parameter θ, we solve the following optimization problem:

$$\min_{\theta}\left[\|\mathbf{y} - \mathbf{X}\theta\|^2 + \lambda\|\theta\|^2\right], \quad (7)$$

where $\lambda > 0$ is some tuning parameter to be chosen.

- Assignment 4: For (7), we have

$$\hat{\theta} = (\lambda I_d + \mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}, \quad (8)$$

where I_d is the $d \times d$ identity matrix.

Assignment 4: Solution:

Suppose $J(\theta) = \|y - X\theta\|^2 + \lambda\|\theta\|^2$

$\therefore J(\theta) = (y - X\theta)^T(y - X\theta) + \lambda\theta^T\theta = y^Ty - y^TX\theta - \theta^TX^Ty + \theta^TX^TX\theta + \lambda\theta^T\theta = y^Ty - 2\theta^TX^Ty + \theta^TX^TX\theta + \lambda\theta^T\theta \;\therefore \frac{dJ(\theta)}{d\theta} = -2X^Ty + 2X^TX\theta + 2\lambda\theta = 0 \Rightarrow (X^TX + \lambda I_d)\theta = X^Ty \Rightarrow \theta = (X^TX + \lambda I_d)^{-1}X^Ty$

shown.