Statistical Machine Learning

1주차

담당: 15기 염윤석

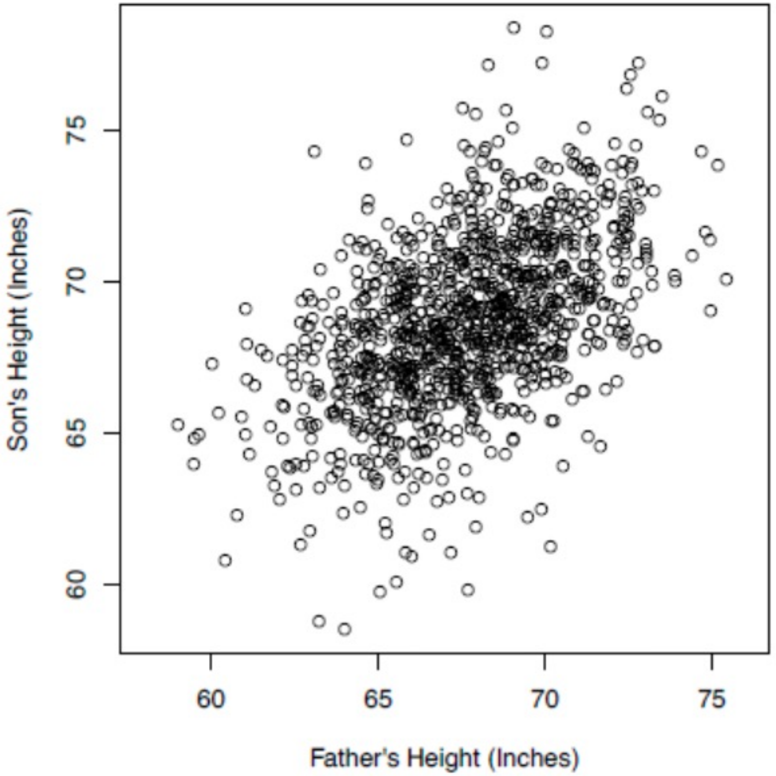
1.Regression

2. Linear Regression

3. Regularization

1. Regression

Regression



Son's Height (Inches)
60
65
70
75
Father's Height (Inches)
60
65
70
75

Linear Discriminant

Discriminant Method

- Assume model $g_i(x)$ directly, **no density estimation**
- **Estimate boundary $g_i(x)$ from data x**

Discriminant : $g_i(x) = w_i^T \mathrm{x} + w_{i0}$ = score = z

$$P(Y\,|X) = p^y(1-p)^{(1-y)}$$

$$P(Y = 1\,|X) = p = \frac{1}{1+e^{-z}} = \sigma(z) = \textit{sigmoid function}$$

$$\log L(p) = \sum_{i=1}^{n}(y_i \log p + (1-y_i)\log(1-p))$$

Maximize Log Likelihood

- Binary Cross Entropy

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot \log(\hat{y}_i) + (1-y_i) \cdot \log(1-\hat{y}_i)$$

Minimize Loss Function

Parametric method - Discriminant

$r = \underline{f(x)} + \varepsilon$

Estimator this one directly! = $g(x|w)$

[Assumptions of error] : Normality & Homoscedasticity & independent

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

$r \sim N(g(x|w), \sigma^2)$

We need g(x|w), we need “w”

MLE! : Maximize log p(r|x)

$$\boxed{log \prod_{t=1}^{N} p(r^t|x^t) = log \prod_{t}^{N} \frac{1}{\sqrt{2\pi\sigma}} exp\left[-\frac{r^t - g(x^t|w)}{2\sigma^2}\right] \rightarrow \text{Maximize!}}$$

From Log-likelihood to Error

$$\boxed{log\prod_{t=1}^{N} p(r^t|x^t) = log\prod_{t}^{N}\frac{1}{\sqrt{2\pi\sigma}}\,exp\left[-\frac{(r^t-g(x^t|w))^2}{2\sigma^2}\right] \rightarrow \text{Maximize!}}$$

Minimize : E(w|x) = $\frac{1}{2}\sum_{t=1}^{N}[(r^t-g(x^t|w))^2]$

2. Linear Regression

Linearity & Linear Model

- Linearity?

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_p X_i^p + \epsilon_i$$

Linear Regression

$r = f(x) + \varepsilon$

Estimator this one directly! = $g(x|w) = w_1x_1+\cdots+w_dx_d+w_0 = w^T x + w_0$

Assume as Linear model

[Assumptions of error] : Normality & Homoscedasticity & independent

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

$r \sim N(g(x|w), \sigma^2)$

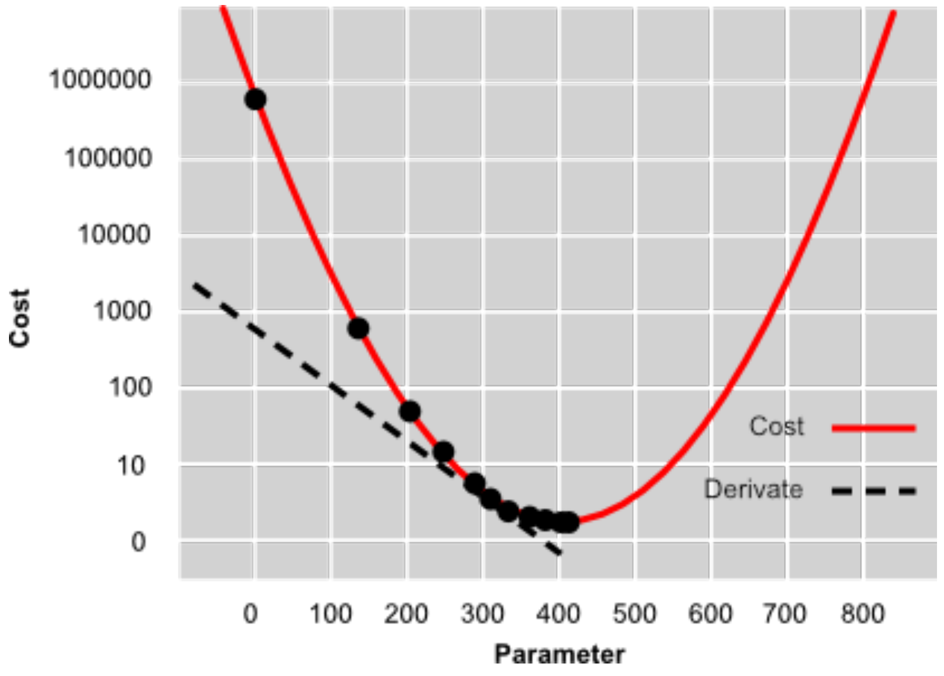
$$\log \prod_{t=1}^{N} p(r^t|x^t) = \log \prod_{t}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{(r^t - g(x^t|w))^2}{2\sigma^2}\right]$$ → Maximize!

Minimize : Loss function E(w|x) = $\frac{1}{2}\sum_{t=1}^{N}[(r^t - g(x^t|w))^2]$

Gradient Descent

Minimize : Loss function E(w|x) = $\frac{1}{N}\sum_{t=1}^{N}[(r^t - g(x^t|w))^2] = MSE(Mean\ Squared\ Error)$

$w^* = argmin_w E(w|x)$ $\qquad w_{j+1} \leftarrow w_j - \eta \frac{\partial E}{\partial w_j}$ iteratively



Least Square Estimation

Minimize : Loss function E(w|x) = $\frac{1}{2}\sum_{t=1}^{N}[(r^t - g(x^t|w))^2]$

** $g(x^t|w) = w_1 x^t + w_0$: 1st order

$$w^* = argmin_w E(w|x) \rightarrow \frac{\partial E}{\partial w_1} = 0 \;\&\; \frac{\partial E}{\partial w_0} = 0$$

$$A = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t (x^t)^2 \end{bmatrix} \; \text{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \; y = \begin{bmatrix} \sum_t y \\ \sum_t r^t x^t \end{bmatrix}$$

$$w^* = A^{-1} y$$

Polynomial Regression

$$g\left(x^t \mid w_k,\ldots,w_2,w_1,w_0\right) = w_k\left(x^t\right)^k + \cdots + w_2\left(x^t\right)^2 + w_1 x^t + w_0$$

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t & \sum_t (x^t)^2 & \cdots & \sum_t (x^t)^k \\ \sum_t x^t & \sum_t (x^t)^2 & \sum_t (x^t)^3 & \cdots & \sum_t (x^t)^{k+1} \\ \vdots & & & & \\ \sum_t (x^t)^k & \sum_t (x^t)^{k+1} & \sum_t (x^t)^{k+2} & \cdots & \sum_t (x^t)^{2k} \end{bmatrix}$$

$$\boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}, \; \boldsymbol{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \\ \sum_t r^t (x^t)^2 \\ \vdots \\ \sum_t r^t (x^t)^k \end{bmatrix}$$

$$\boxed{w^* = A^{-1} y}$$

Multivariate Regression

$$r^t = g(\mathbf{x}^t | w_0, w_1, \ldots, w_d) + \epsilon = w_0 + w_1 x_1^t + w_2 x_2^t + \cdots + w_d x_d^t + \epsilon$$

$$E(w_0, w_1, \ldots, w_d | \mathcal{X}) = \frac{1}{2} \sum_t (r^t - w_0 - w_1 x_1^t - w_2 x_2^t - \cdots - w_d x_d^t)^2$$

$$\begin{aligned}
\sum_t r^t &= N w_0 + w_1 \sum_t x_1^t + w_2 \sum_t x_2^t + \cdots + w_d \sum_t x_d^t \\
\sum_t x_1^t r^t &= w_0 \sum_t x_1^t + w_1 \sum_t (x_1^t)^2 + w_2 \sum_t x_1^t x_2^t + \cdots + w_d \sum_t x_1^t x_d^t \\
\sum_t x_2^t r^t &= w_0 \sum_t x_2^t + w_1 \sum_t x_1^t x_2^t + w_2 \sum_t (x_2^t)^2 + \cdots + w_d \sum_t x_2^t x_d^t \\
&\vdots \\
\sum_t x_d^t r^t &= w_0 \sum_t x_d^t + w_1 \sum_t x_d^t x_1^t + w_2 \sum_t x_d^t x_2^t + \cdots + w_d \sum_t (x_d^t)^2
\end{aligned}$$

22

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_d^1 \\ 1 & x_1^2 & x_2^2 & \cdots & x_d^2 \\ \vdots & & & & \\ 1 & x_1^N & x_2^N & \cdots & x_d^N \end{bmatrix}, \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d, \end{bmatrix}, \boldsymbol{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}$$

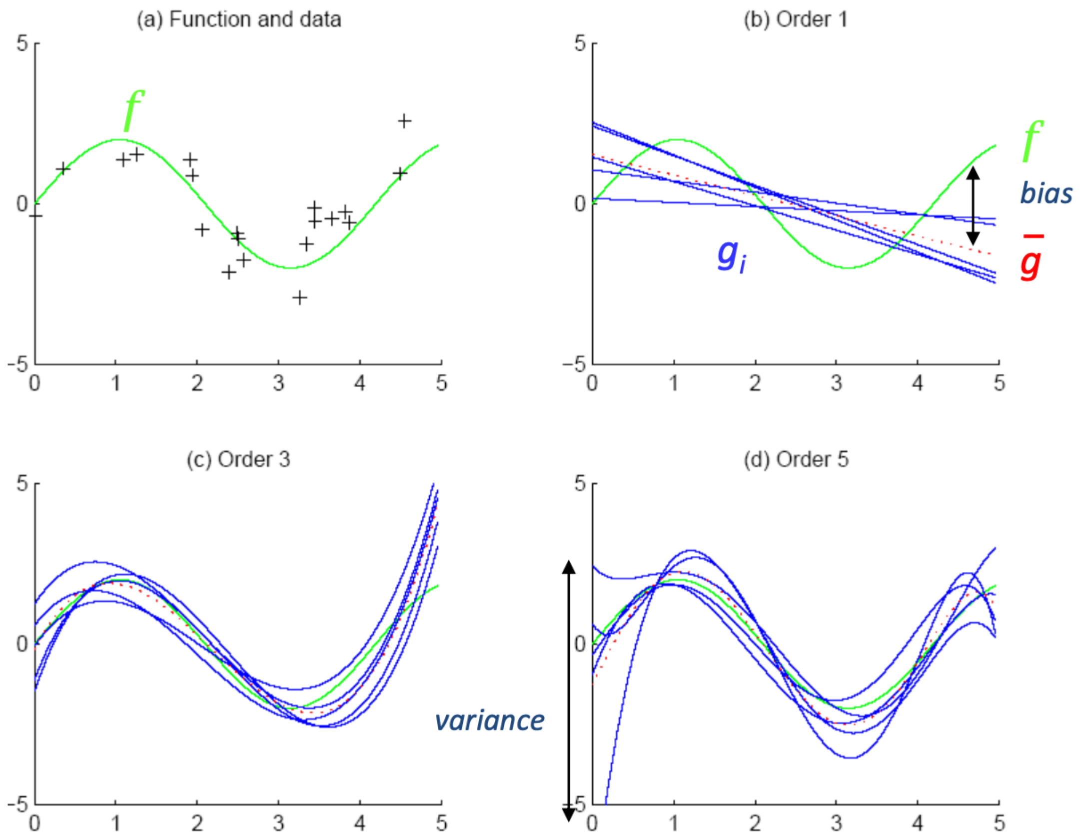
Then the normal equations can be written as

$$\mathbf{X}^T \mathbf{X} \boldsymbol{w} = \mathbf{X}^T \boldsymbol{r}$$

$$\boldsymbol{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{r}$$

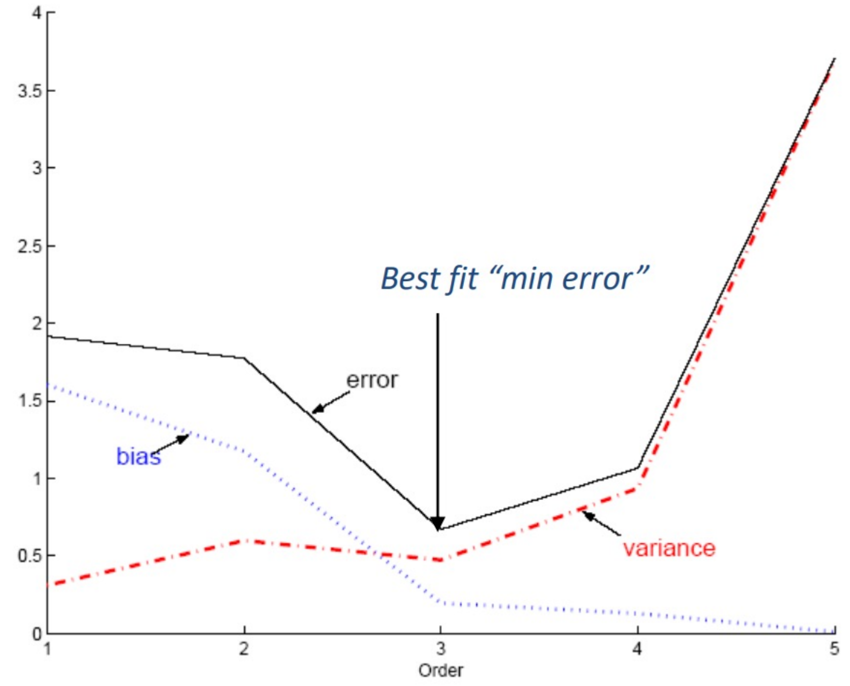
$$\boxed{w^* = A^{-1} y}$$

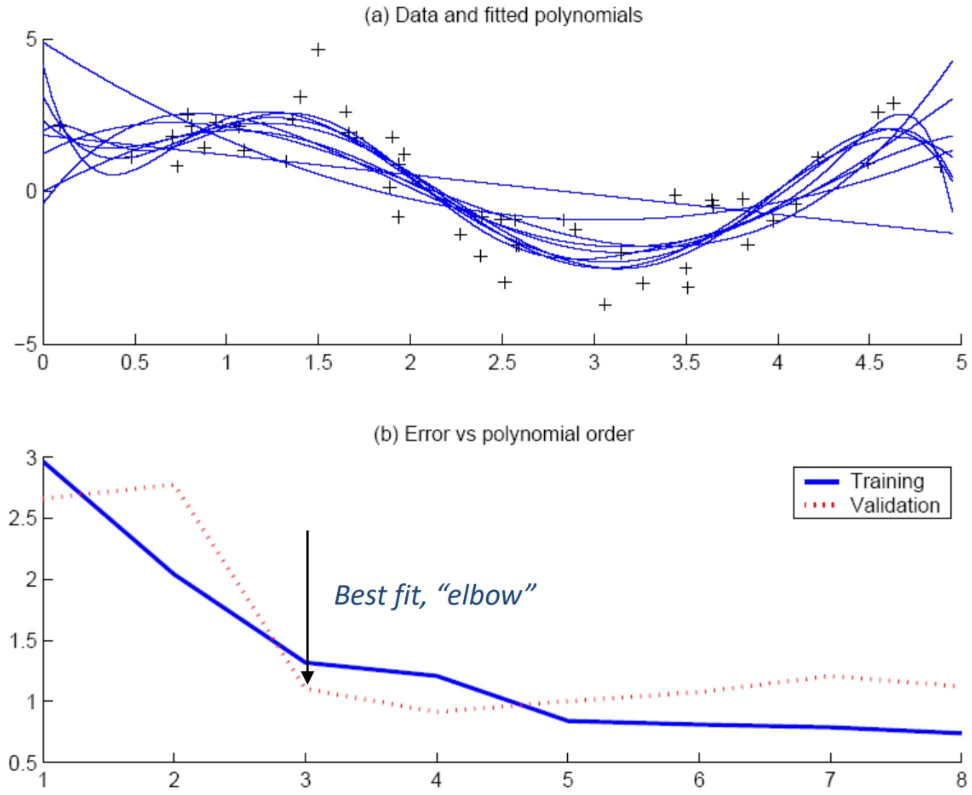
Polynomial Regression



2

Model Selection





Cross Validation

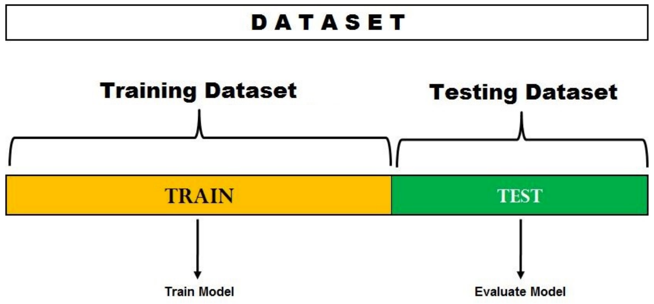
To estimate generalization error, we need data unseen during training.

We split the data as

- Training set (50%)
- Validation set (25%)
- Test (publication) set (25%)

Measure generalization accuracy by testing on data unused during training

Hold out



Regularization

Penalize complex models

- E'=error on data + λ*model complexity

* If λ increases, variance decreases, but bias increases

In regression...

Regularization (L2):

$$E(\mathbf{w} \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g(x^t \mid \mathbf{w})\right]^2 + \lambda \sum_i w_i^2$$

3. Regularization

Distance

- Distance measure

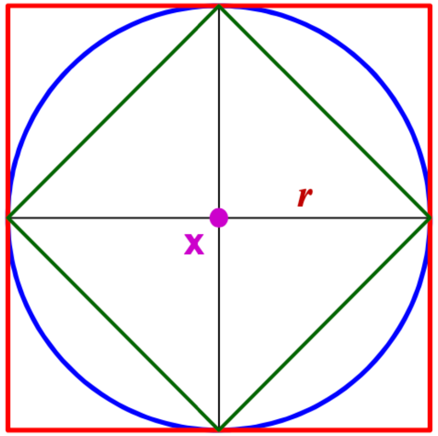
$$d(\mathbf{u},\mathbf{v}) = (\sum|u_i - v_i|^2)^{\frac{1}{2}} = ||\mathbf{u}-\mathbf{v}||_2 \qquad \textit{Euclidean (L2 norm)}$$

$$d(\mathbf{u},\mathbf{v}) = \sum|u_i - v_i| = ||\mathbf{u}-\mathbf{v}||_1 \qquad \textit{Manhattan (L1 norm)}$$

$$d(\mathbf{u},\mathbf{v}) = (\sum|u_i - v_i|^p)^{\frac{1}{p}} = ||\mathbf{u}-\mathbf{v}||_p \qquad \textit{Minkowski (Lp norm)}$$

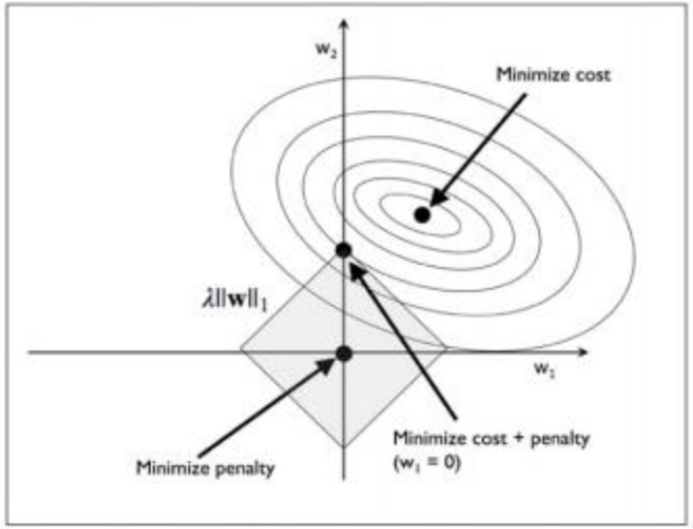
$$d(\mathbf{u},\mathbf{v}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})} \qquad \textit{Mahalanobis Distance}$$

Distance



- Green: All points **y** at distance $\mathbf{L_1(x, y) = r}$ from point **x**
- Blue: All points **y** at distance $\mathbf{L_2(x, y) = r}$ from point **x**
- Red: All points **y** at distance $\mathbf{L_\infty(x, y) = r}$ from point **x**

Lasso Regression

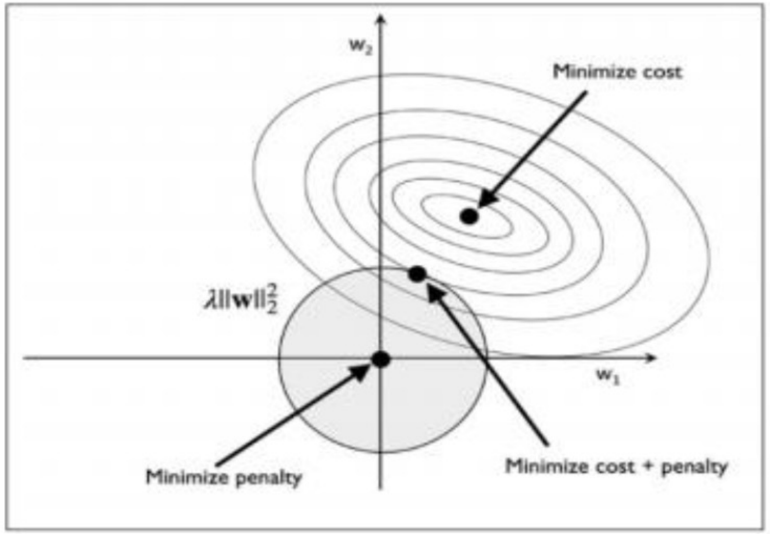


- LASSO (Least Absolute Shrinkage and Selection Operator)

$$(\widehat{\boldsymbol{\beta}}^{\lambda,1} =)\ \widehat{\boldsymbol{\beta}}_{LASSO} = \underset{\boldsymbol{\beta}}{argmin}\ (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\ ||\boldsymbol{\beta}||_1$$

where $||\boldsymbol{\beta}||_1 = \sum_j^p |\beta_j|$

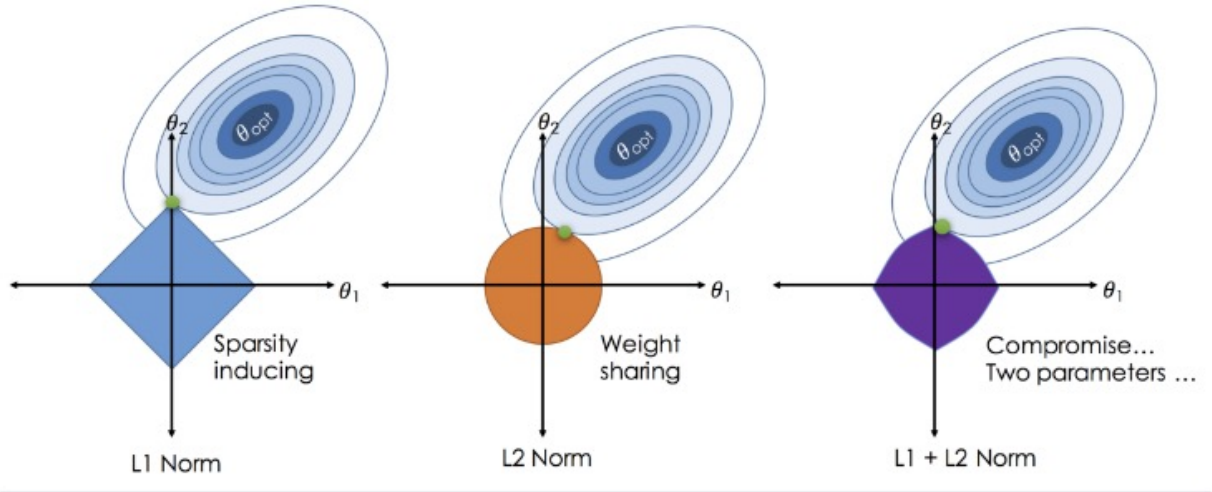
Ridge Regression



Ridge Regression solves

$$\min_{\boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta}||_2^2 \qquad (L2\ penalty)$$

Elastic-Net Regression



$$\frac{\sum_{i=1}^{n}(y_i - x_i^J \hat{\beta})^2}{2n} + \lambda \left(\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{\beta}_j^2 + \alpha \sum_{j=1}^{m} |\hat{\beta}_j| \right)$$

수고하셨습니다!

해당 세션자료는 KUBIG Github에서 보실 수 있습니다!

다음은 이번 주차 과제 설명이 있습니다!