

Statistical Machine Learning

3주차

담당: 15기 박지우

Regression



1. Linear Model

2. Linear Regression

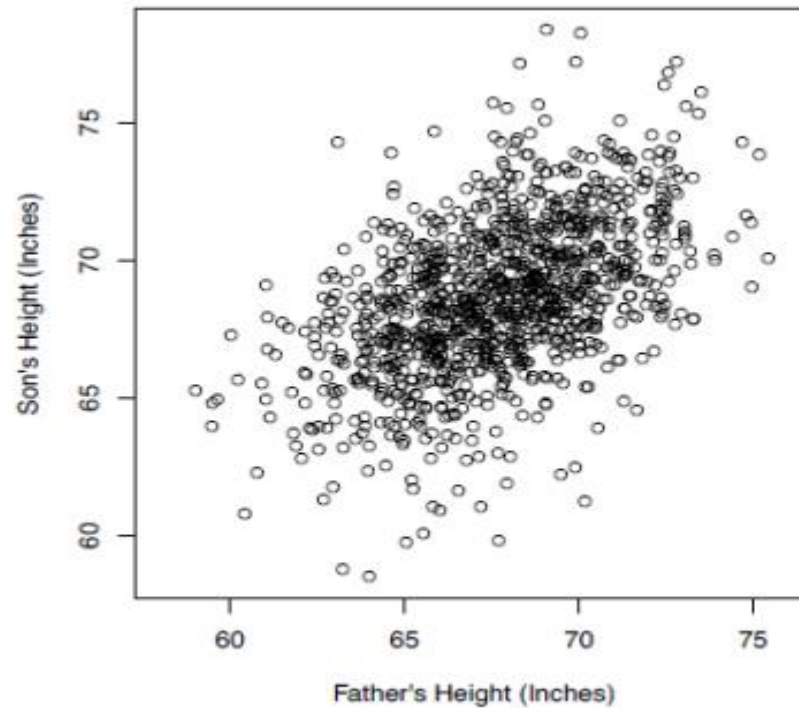
3. MSE



4. Regularization

Linear Model

What is Regression



Linearity

- 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 Model

- Linearity? \longrightarrow Linear Model

$$Y_i \stackrel{\text{ind}}{\sim} (\mu_i(\mathbf{X}_i), \sigma^2) \quad \text{where} \quad E[Y_i] = \mu_i(\mathbf{X}_i)$$

$$\mu_i(\mathbf{X}_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} = \boldsymbol{\beta}^T \mathbf{X}_i$$

$$\boldsymbol{\mu}(\mathbf{X}) = \mathbf{X} \boldsymbol{\beta}$$

Linear Regression

Linear Regression

- Least Square Estimator

$$\sum \epsilon_i^2 = \sum (Y_i - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})^2$$

$$\frac{\partial}{\partial \beta_0} \sum (Y_i - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})^2 \stackrel{set}{=} 0$$

$$\frac{\partial}{\partial \beta_1} \sum (Y_i - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})^2 \stackrel{set}{=} 0$$

⋮

$$\frac{\partial}{\partial \beta_p} \sum (Y_i - \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi})^2 \stackrel{set}{=} 0$$

Linear Regression

- Error term?
 - Mean 0
 - Identical, Independent
 - Normal?

MSE

Linear Regression and likelihood function

- Normal distribution

$$\log L(\mu) \approx -\frac{\sum_{i=1}^n (y_i - \mu)}{\sigma^2}$$

Likelihood function and Loss function

- Binary Cross Entropy
- Categorical Cross Entropy
- MSE

Generalized Linear Model

	Normal	Poisson	Binomial	Gamma	Inv Gaussian
Notation	$N(\mu, \sigma^2)$	$P(\mu)$	$B(n, \pi)/n$	$G(\mu, v)$	$IG(\mu, \sigma^2)$
Support	$(-\infty, \infty)$	$\{0, 1, \dots\}$	$\{0, \dots, n\}/n$	$(0, \infty)$	$(0, \infty)$
$a(\phi)$	$\phi = \sigma^2$	1	$1/m$	v^{-1}	σ^2
$b(\theta)$	$\theta^2/2$	e^θ	$\log(1 + e^\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$b'(\theta) = E(Y)$	θ	e^θ	$\frac{e^\theta}{1+e^\theta}$	$-1/\theta$	$(-2\theta)^{-1/2}$
$(b')^{-1}(\mu) = g(\mu)$	μ	$\log(\mu)$	$\log \frac{\mu}{1-\mu}$	μ^{-1}	μ^{-2}
$b''(\theta)$	1	μ	$\mu(1 - \mu)$	μ^2	μ^3

Table: Summary of some popular GLM models.

Regularization

Stein's Paradox

- Let $\mathbf{X} = [X_1, \dots, X_p]^T \sim N_p(\boldsymbol{\theta}, I)$
- The UMVUE and MLE of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}_{MLE,UMVUE} = \mathbf{X}$$

- Using squared error loss, the risk of $\hat{\boldsymbol{\theta}}_{MLE,UMVUE}$ is

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{UMVUE}) = E[||\mathbf{X} - \boldsymbol{\theta}||^2] = p$$

Stein's Paradox

- James and Stein (1961) Estimator

$$\hat{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{p-2}{\|\mathbf{X}\|^2}\right) \mathbf{X}$$

- When $p \geq 3$,

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS}) = p - (p-2)E\left(\frac{1}{\|\mathbf{X}\|^2}\right) < p$$

Stein's Paradox

- Proof

$$\begin{aligned} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS}) &= E \left[\left\| \mathbf{X} - \boldsymbol{\theta} - \frac{(p-2)\mathbf{X}}{\|\mathbf{X}\|^2} \right\|^2 \right] \\ &= p - 2(p-2) \sum_j^p E \left(\frac{X_j(X_j - \theta_j)}{\|\mathbf{X}\|^2} \right) + (p-2)^2 E \left(\frac{1}{\|\mathbf{X}\|^2} \right) \\ &= p - (p-2) E \left(\frac{1}{\|\mathbf{X}\|^2} \right) \end{aligned}$$

Since $\sum_j^p E \left(\frac{X_j(X_j - \theta_j)}{\|\mathbf{X}\|^2} \right) = (p-2) E \left(\frac{1}{\|\mathbf{X}\|^2} \right)$

Stein's Paradox

JS estimator shrinks each component of X towards the origin, and thus the biggest improvement comes when

$\| \theta \|$ is close to zero

Normality assumption is not critical, and similar results can be shown for a wide class of distributions.

Ridge Regression

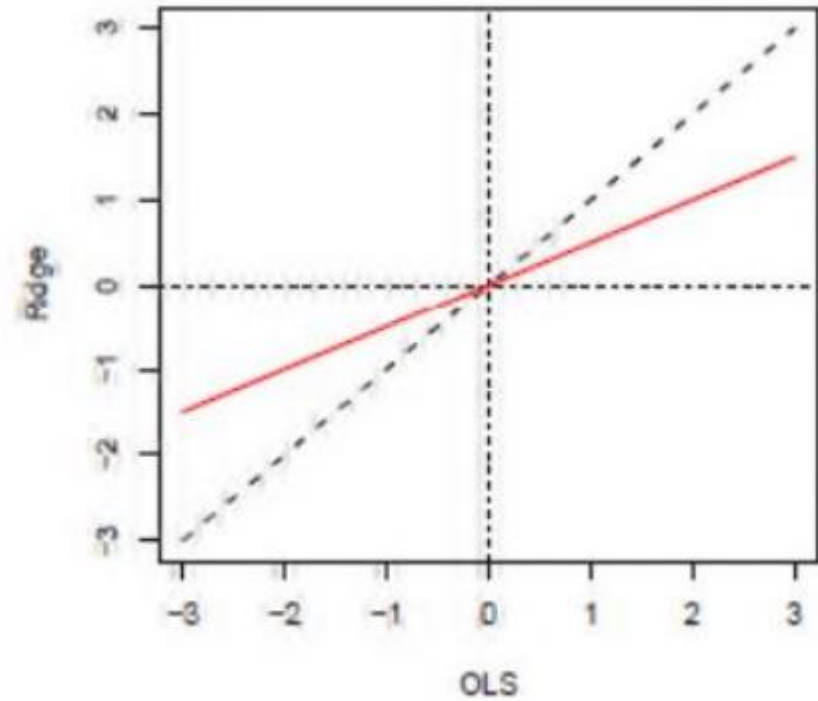
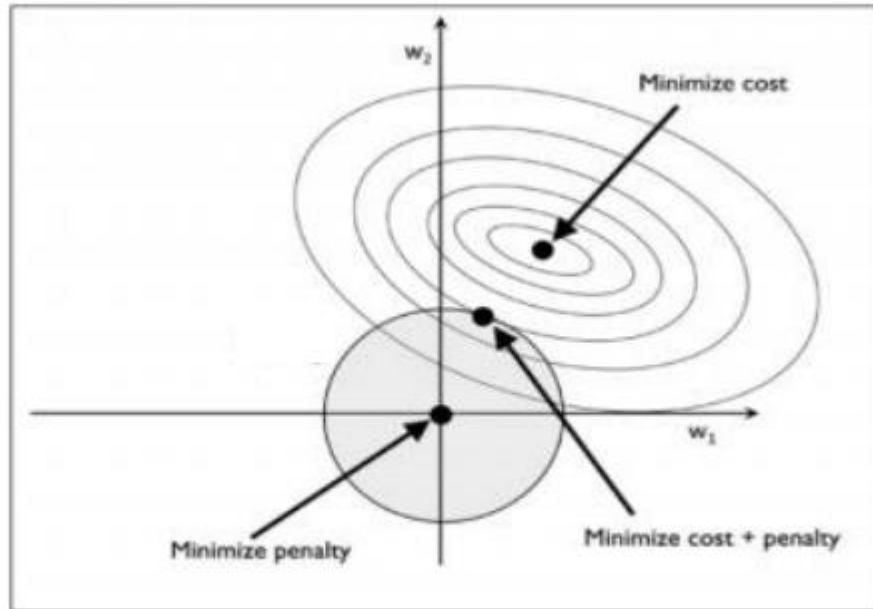
- We can consider

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}$$

- Ridge estimator is

$$\hat{\boldsymbol{\beta}}_{Ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta}$$

Ridge Regression



Lasso Regression

- Ridge Regression solves

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

- LASSO Regression solves

$$\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta) + \lambda ||\beta||_1 \quad (L1 \text{ penalty})$$

Lasso Regression

- LASSO (Least Absolute Shrinkage and Selection Operator)

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

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

Lasso Regression

