Lecture 16 : Ridge Regression

• Linear regression based on a version of $\hat{\mu}^{MLE}$. We usually assume an n-dim vector $\underline{y} = (y_1, \dots, y_n)^T$ from linear model

$$\underline{y} = X \cdot \beta + \underline{\epsilon}$$

$X^{n \times p}$ known as design/structure matrix ; β is an unknown p-dimensional parameter vector & noise vector $\underline{\epsilon}$ is assumed to have uncorrelated components & constant variance. Thus $\epsilon \sim (\underline{0}, \sigma^2 \cdot I_n)$

Identity matrix of size n

$$\begin{bmatrix} 1 & & 0 \\ & 1 & \\ & 1\,1\,1 & \\ 0 & & 1 \end{bmatrix}$$

Diagonal elements

The least squares estimator (Gauss & Legendre in the early 1800s!) minimizes SSE

$$\hat{\beta} = \underset{\beta}{\arg\min} \{ \| \underline{y} - X\beta \|^2 \}$$

$$-2(y - X\beta) \cdot X = 0$$

$p \times \cancel{n}$ $\cancel{n} \times 1$

$$(y - X\beta) \cdot X = 0 \qquad X^T \cdot y$$

$$X^T y - X^T \cdot X \cdot \beta = 0$$

$$X^T X \beta = X^T \cdot y \; ; \; \hat{\beta} = (X^T X)^{-1} \cdot X^T y$$

$$\text{Let } S = (X^T X) \Rightarrow \hat{\beta} = S^{-1} X^T y$$

$$\hat{\beta}(\lambda) = \underset{\beta}{\arg\min} \{ \|y - X \cdot \beta\|^2 + \lambda \cdot \|\beta\|^2 \}$$

$$\boxed{\text{Recall } \|\underline{\beta}\|^2 = \beta_1^2 + \beta_2^2 + \dots + \beta_p^2}$$

Then, do similar math

$$-2(y - X^T \beta) \cdot X + 2 \cdot \lambda \cdot \beta = 0$$

$$-X^T y + X^T X \cdot \beta + \lambda \beta = 0$$

$$\beta (X^T X + \lambda \cdot I_n) = X^T y$$

$$\boxed{\hat{\beta}^{Ridge} = \underbrace{(X^T X + \lambda \cdot I_n)}^{-1} \cdot X^T y}$$

Standard errors on $\hat{\beta}$

$$\hat{\beta}^{OLS} \sim \left(\beta, \sigma^2 \cdot S^{-1}\right)$$

$$\hat{\beta}^{Ridge} \sim \left((S+\lambda \cdot I_n)^{-1} \cdot S \cdot \beta, \quad \cdot \quad \right)$$

$$\sigma^2 \cdot (S+\lambda \cdot I_n)^{-1} \cdot S \cdot (S+\lambda \cdot I_n)^{-1}$$