Exercise 6 2021-02-04

2. Bias and variance of ridge regression.

Ridge regression:

$$\hat{\beta}_\tau = \arg\min_\beta \|X\beta - y\|_2^2 + \tau\|\beta\|_2^2 \quad , \quad \tau \geq 0$$

True model $y = X\beta^* + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$.

Prove that

$$\mathbb{E}[\hat{\beta}_\tau] = S_\tau^{-1} S \beta^* \quad , \quad \mathrm{Cov}[\hat{\beta}_\tau] = S_\tau^{-1} S S_\tau^{-1} \sigma^2$$

where

$$S = X^T X \quad , \quad S_\tau = X^T X + \tau \mathbb{1}.$$

By taking the derivative of the loss function with respect to β, one can obtain

$$\frac{\partial \text{Loss}}{\partial \beta} = 2\left[X^T(y - X\beta) + \tau\beta\right] \overset{!}{=} 0 ,$$

for which we can solve for the zero point to in turn obtain the optimal $\hat{\beta}_\tau$ as per

$$\hat{\beta}_\tau = (X^T X + \tau \mathbb{1})^{-1} X^T y .$$

Then we have the expectation value as

$$\begin{aligned}
E[\hat{\beta}_\tau] &= E\left[(X^T X + \tau \mathbb{1})^{-1} X^T \overset{y = X\beta^* + \varepsilon}{y}\right] \\
&= E\left[(X^T X + \tau \mathbb{1})^{-1} X^T (X\beta^* + \varepsilon)\right] \\
&= E\left[(X^T X + \tau \mathbb{1})^{-1} X^T X \beta^*\right] + E\left[(X^T X + \tau \mathbb{1})^{-1} X^T \varepsilon\right] \\
&= E\left[S_\tau^{-1} S \beta^*\right] + E\left[S_\tau^{-1} X^T\right] \underbrace{E[\varepsilon]}_{=0 \text{ since } \varepsilon \sim N(0,\sigma^2)} \\
&= S_\tau^{-1} S \, E[\beta^*] \\
&= S_\tau^{-1} S \beta^*
\end{aligned}$$

And then the variance becomes

$$\mathrm{Cov}[\hat{\beta}_\tau] = E[(\hat{\beta}_\tau - E[\hat{\beta}_\tau])^2]$$

$$= E[(\hat{\beta}_\tau - E[\hat{\beta}_\tau])(\hat{\beta}_\tau - E[\hat{\beta}_\tau])^T]$$

$$= E[\hat{\beta}_\tau \hat{\beta}_\tau^T - \hat{\beta}_\tau E[\hat{\beta}_\tau]^T - E[\hat{\beta}_\tau]\hat{\beta}_\tau^T + E[\hat{\beta}_\tau]E[\hat{\beta}_\tau]^T]$$

$$= E[\hat{\beta}_\tau \hat{\beta}_\tau^T] - E[\hat{\beta}_\tau]E[\hat{\beta}_\tau]^T - \cancel{E[\hat{\beta}_\tau]E[\hat{\beta}_\tau]^T} + \cancel{E[\hat{\beta}_\tau]E[\hat{\beta}_\tau]^T}$$

$$= E[(S_\tau^{-1}X^T y)(S_\tau^{-1}X^T y)^T] - (S_\tau^{-1}S\beta^*)(S_\tau^{-1}S\beta^*)^T$$

$$= E[(S_\tau^{-1}X^T(X\beta_* + \varepsilon))(S_\tau^{-1}X^T(X\beta^* + \varepsilon))^T] - (S_\tau^{-1}S\beta_*)(S_\tau^{-1}S\beta^*)^T$$

$$\left[\begin{aligned} &(S_\tau^{-1}X^T(X\beta^* + \varepsilon))(S_\tau^{-1}X^T(X\beta^* + \varepsilon))^T = \\ &= (S_\tau^{-1}X^TX\beta^*)(S_\tau^{-1}X^TX\beta^*)^T + S_\tau^{-1}X^T\varepsilon(S_\tau^{-1}X^TX\beta^*)^T \\ &+ S_\tau^{-1}X^TX\beta^*(S_\tau^{-1}X^T\varepsilon)^T + S_\tau^{-1}X^T\varepsilon(S_\tau^{-1}X^T\varepsilon)^T \end{aligned}\right]$$

$$= E[(S_\tau^{-1}S\beta^*)(S_\tau^{-1}S\beta^*)^T] + E[S_\tau^{-1}X^T]\underbrace{E[\varepsilon]}_{=0}E[S_\tau^{-1}S\beta^*]^T$$

$$+ E[S_\tau^{-1}S\beta^*]\underbrace{E[\varepsilon]^T}_{=0}E[S_\tau^{-1}X^T] + E[S_\tau^{-1}X^T\varepsilon(S_\tau^{-1}X^T\varepsilon)^T]$$

$$- (S_\tau^{-1}S\beta^*)(S_\tau^{-1}S\beta^*)^T$$

$$= \cancel{(S_\tau^{-1}S\beta^*)(S_\tau^{-1}S\beta)^T} + E[S_\tau^{-1}X^T]E[\varepsilon\varepsilon^T]E[X(S_\tau^{-1})^T]$$

$$- \cancel{(S_\tau^{-1}S\beta^*)(S_\tau^{-1}S\beta^*)}$$

$$= S_\tau^{-1}X^T \underbrace{E[\varepsilon\varepsilon^T]}_{=\sigma^2} X \underbrace{(S_\tau^{-1})^T}$$

Covariance matrix symmetric

$$\Rightarrow (S_\tau^{-1})^T = S_\tau^{-1}$$

$$= S_\tau^{-1}X^TX S_\tau^{-1}\sigma^2$$

$$= S_\tau^{-1} S S_\tau^{-1}\sigma^2$$

$\blacksquare$