Exercise 5

June 20, 2023

1 Bias and variance of ridge regression

Ridge regression solves the regularized least squares problem

$$\widehat{\beta}_\tau = \operatorname{argmin}_\beta (y - X\beta)^\top (y - X\beta) + \tau \beta^\top \beta$$

with regularization parameter $\tau \geq 0$. Assume that the true model is $y = X\beta^* + \epsilon$ with zero mean Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and centered features $\frac{1}{N}\sum_i X_i = 0$ (note that these assumptions imply that y is also centered in expectation).
First we calculated the derivative for β,

$$\frac{\partial}{\partial \beta}\left((y - X\beta)^\top (y - X\beta) + \tau \beta^\top \beta\right) = -2X^\top (y - X\beta) + 2\tau\beta \stackrel{!}{=} 0.$$

Thus, we have

$$\begin{aligned}
&X^\top X\beta + \tau\beta = X^\top y \\
&(X^\top X + \tau \mathbb{I}_D)\beta = X^\top y \\
&\widehat{\beta}_\tau = (X^\top X + \tau \mathbb{I}_D)^{-1} X^\top y.
\end{aligned}$$

- **Claim:**

  $$\mathbb{E}\left[\widehat{\beta}_\tau\right] = S_\tau^{-1} S \beta^* = V \operatorname{diag}\left(\frac{\lambda_j^2}{\lambda_j^2 + \tau}\right) V^\top \beta^*,$$

  where V comes form Singular Value Decomposition, λ_j is the j-th singular value of X. As we assumed (on lecture notes) that $X = U\Lambda V^\top$. S and S_τ are the ordinary and regularized scatter matrices:

  $$S = X^\top X \qquad S_\tau = X^\top X + \tau \mathbb{I}_D.$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}_\tau\right] &= \mathbb{E}\left[(X^\top X + \tau \mathbb{I}_D)^{-1} X^\top X \beta^*\right] + \mathbb{E}\left[(X^\top X + \tau \mathbb{I}_D)^{-1} X^\top \epsilon\right] \\
&= \mathbb{E}\left[(X^\top X + \tau \mathbb{I}_D)^{-1} X^\top X \beta^*\right] + \mathbb{E}\left[(X^\top X + \tau \mathbb{I}_D)^{-1} X^\top\right] \cdot \mathbb{E}[\epsilon] \\
&= \mathbb{E}\left[(X^\top X + \tau \mathbb{I}_D)^{-1} X^\top X \beta^*\right] \\
&= (X^\top X + \tau \mathbb{I}_D)^{-1} X^\top X \beta^* \\
&= S_\tau^{-1} S \beta^*.
\end{aligned}$$

In second line, we use the property of expectation for two independent random variables ($\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$). Since error term $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then $\mathbb{E}[\epsilon] = 0$.
Now we prepare SVD for $X^\top X$ and $(X^\top X + \tau \mathbb{I}_D)^{-1}$.

$$X^\top X = (U \Lambda V^\top)^\top (U \Lambda V^\top) = V \Lambda^2 V^\top. \tag{1}$$

Then we perform some basic matrix calculations:

$$\begin{aligned}
(X^\top X + \tau \mathbb{I}_D)^{-1} &= \left(V \Lambda^2 V^\top + V \tau \mathbb{I}_D V^\top\right)^{-1} \\
&= \left(V (\Lambda^2 + \tau \mathbb{I}_D) V^\top\right)^{-1} \\
&= V (\Lambda^2 + \tau \mathbb{I}_D)^{-1} V^\top.
\end{aligned}$$

Here we have the second preparation result.

$$(X^\top X + \tau \mathbb{I}_D)^{-1} = V (\Lambda^2 + \tau \mathbb{I}_D)^{-1} V^\top. \tag{2}$$

Now, let's continue our proof,

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}_\tau\right] &= (X^\top X + \tau \mathbb{I}_D)^{-1} X^\top X \beta^* \\
&= V (\Lambda^2 + \tau \mathbb{I}_D)^{-1} V^\top \cdot V \Lambda^2 V^\top \beta^* \\
&= V (\Lambda^2 + \tau \mathbb{I}_D)^{-1} \mathbb{I}_D \Lambda^2 V^\top \beta^* \\
&= V (\Lambda^2 + \tau \mathbb{I}_D)^{-1} \Lambda^2 V^\top \beta^* \\
&= V \operatorname{diag}\left(\frac{\lambda_j^2}{\lambda_j^2 + \tau}\right) V^\top \beta^*.
\end{aligned}$$

When $\tau = 0$, we have

$$\mathbb{E}\left[\widehat{\beta}_\tau\right] = V \operatorname{diag}\left(\frac{\lambda_j^2}{\lambda_j^2 + 0}\right) V^\top \beta^* = V V^\top \beta^* = \beta^*.$$

□

- **Claim:**

$$\operatorname{Cov}\left[\widehat{\beta}_\tau\right] = S_\tau^{-1} S S_\tau^{-1} \sigma^2 = V \operatorname{diag}\left(\frac{\lambda_j^2}{(\lambda_j^2 + \tau)^2}\right) V^\top \sigma^2.$$

Proof. By definition of covariance, we have

$$\operatorname{Cov}\left[\widehat{\beta}_\tau\right] = \mathbb{E}\left[\left(\widehat{\beta}_\tau - \mathbb{E}\left[\widehat{\beta}_\tau\right]\right)\left(\widehat{\beta}_\tau - \mathbb{E}\left[\widehat{\beta}_\tau\right]\right)^\top\right],$$

then we need to calculate that

$$\begin{aligned}
\widehat{\beta}_\tau - \mathbb{E}\left[\widehat{\beta}_\tau\right] &= (X^\top X + \tau \mathbb{I}_D)^{-1} X^\top y - (X^\top X + \tau \mathbb{I}_D)^{-1} X^\top X \beta^* \\
&= (X^\top X + \tau \mathbb{I}_D)^{-1} X^\top (X \beta^* + \epsilon) - (X^\top X + \tau \mathbb{I}_D)^{-1} X^\top X \beta^* \\
&= (X^\top X + \tau \mathbb{I}_D)^{-1} X^\top \epsilon.
\end{aligned}$$

Here we will show the first desired result,

$$\begin{aligned}
\mathrm{Cov}\left[\widehat{\beta}_\tau\right] &= \mathbb{E}\left[\left(\widehat{\beta}_\tau - \mathbb{E}\left[\widehat{\beta}_\tau\right]\right)\left(\widehat{\beta}_\tau - \mathbb{E}\left[\widehat{\beta}_\tau\right]\right)^\top\right] \\
&= \mathbb{E}\left[\left((X^\top X + \tau \mathbb{I}_D)^{-1} X^\top \epsilon\right)\left((X^\top X + \tau \mathbb{I}_D)^{-1} X^\top \epsilon\right)^\top\right] \\
&= \mathbb{E}\left[(X^\top X + \tau \mathbb{I}_D)^{-1} X^\top \epsilon \epsilon^\top X (X^\top X + \tau \mathbb{I}_D)^{-\top}\right] \\
&= \mathbb{E}\left[(X^\top X + \tau \mathbb{I}_D)^{-1} X^\top \sigma^2 \mathbb{I}_D X (X^\top X + \tau \mathbb{I}_D)^{-1}\right] \\
&= \sigma^2 \mathbb{E}\left[(X^\top X + \tau \mathbb{I}_D)^{-1} X^\top X (X^\top X + \tau \mathbb{I}_D)^{-1}\right] \\
&= \mathbb{E}\left[(X^\top X + \tau \mathbb{I}_D)^{-1} X^\top X (X^\top X + \tau \mathbb{I}_D)^{-1}\right] \sigma^2 \\
&= (X^\top X + \tau \mathbb{I}_D)^{-1} X^\top X (X^\top X + \tau \mathbb{I}_D)^{-1} \sigma^2 \\
&= S_\tau^{-1} S S_\tau^{-1} \sigma^2.
\end{aligned}$$

As for the second desired result, just do SVD.

$$\begin{aligned}
\mathrm{Cov}\left[\widehat{\beta}_\tau\right] &= (X^\top X + \tau \mathbb{I}_D)^{-1} X^\top X (X^\top X + \tau \mathbb{I}_D)^{-1} \sigma^2 \\
&= V(\Lambda^2 + \tau \mathbb{I}_D)^{-1} V^\top \cdot V \Lambda^2 V^\top \cdot V (\Lambda^2 + \tau \mathbb{I}_D)^{-1} V^\top \sigma^2 \\
&= V(\Lambda^2 + \tau \mathbb{I}_D)^{-1} \mathbb{I}_D \Lambda^2 \mathbb{I}_D (\Lambda^2 + \tau \mathbb{I}_D)^{-1} V^\top \sigma^2 \\
&= V(\Lambda^2 + \tau \mathbb{I}_D)^{-1} \Lambda^2 (\Lambda^2 + \tau \mathbb{I}_D)^{-1} V^\top \sigma^2 \\
&= V \,\mathrm{diag}\left(\frac{\lambda_j^2}{(\lambda_j^2 + \tau)^2}\right) V^\top \sigma^2.
\end{aligned}$$

When $\tau = 0$, we have

$$\begin{aligned}
\mathrm{Cov}\left[\widehat{\beta}_\tau\right] &= V \,\mathrm{diag}\left(\frac{\lambda_j^2}{(\lambda_j^2 + 0)^2}\right) V^\top \sigma^2 \\
&= V \,\mathrm{diag}\left(\frac{1}{\lambda_j^2}\right) V^\top \sigma^2 \\
&= V \Lambda^{-2} V^\top \sigma^2 \\
&= (X^\top X)^{-1} \sigma^2 \\
&= S^{-1} \sigma^2.
\end{aligned}$$

□

2 LDA-Derivation from the Least Squares Error

We will start from

$$\frac{\partial}{\partial \beta} \sum_{i=1}^{N} (y_i^* - X_i \cdot \beta)^2 \overset{!}{=} 0$$

to show that

$$\Sigma \cdot \beta + \frac{1}{4} (\mu_1 - \mu_{-1})^\top \cdot (\mu_1 - \mu_{-1}) \cdot \beta = \frac{1}{2} (\mu_1 - \mu_{-1})^\top .$$

Proof.

$$\begin{aligned}
\frac{\partial}{\partial \beta} \sum_{i=1}^{N} (y_i^* - X_i \cdot \beta)^2 &\stackrel{!}{=} 0 \\
\sum_{i=1}^{N} -2X_i^\top (y_i^* - X_i \cdot \beta) &= 0 \\
\frac{1}{N} \sum_{i=1}^{N} X_i^\top y_i^* &= \frac{1}{N} \sum_{i=1}^{N} X_i^\top X_i \beta
\end{aligned}$$

Here for left hands side(LHS), we have

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} X_i^\top y_i^* &= \frac{1}{N} \left[\sum_{i:y_i^*=1} X_i^\top \cdot (+1) + \sum_{i:y_i^*=-1} X_i^\top \cdot (-1) \right] \\
&= \frac{1}{N} (N_1 \cdot \mu_1 - N_{-1} \cdot \mu_{-1})^\top \\
&= \frac{1}{2} (\mu_1 - \mu_{-1})^\top .
\end{aligned}$$

For right hands side(RHS), notice that $\mu_1 + \mu_{-1} = 0$, then we have

$$\begin{aligned}
&\frac{1}{N} \sum_{i=1}^{N} X_i^\top X_i \\
=&\frac{1}{N} \sum_{i=1}^{N} X_i^\top \left(X_i - \frac{1}{2}(\mu_1 + \mu_{-1}) \right) \\
=&\frac{1}{N} \left[\sum_{i:y_i^*=1} X_i^\top \left(X_i - \frac{1}{2}(\mu_1 + \mu_{-1}) \right) \right] + \frac{1}{N} \left[\sum_{i:y_i^*=-1} X_i^\top \left(X_i - \frac{1}{2}(\mu_1 + \mu_{-1}) \right) \right] \\
=&\frac{1}{N} \left[\sum_{i:y_i^*=1} X_i^\top \left(X_i - \mu_1 + \frac{1}{2}(\mu_1 - \mu_{-1}) \right) \right] + \frac{1}{N} \left[\sum_{i:y_i^*=-1} X_i^\top \left(X_i - \mu_{-1} - \frac{1}{2}(\mu_1 - \mu_{-1}) \right) \right] . \\
=&\frac{1}{N} \left[\sum_{i=1}^{N} X_i^\top (X_i - \mu_{y_i}) \right] + \frac{1}{N} \left[\sum_{i:y_i^*=1} X_i^\top - \sum_{i:y_i^*=-1} X_i^\top \right] \frac{1}{2}(\mu_1 - \mu_{-1}) \\
=&\frac{1}{N} \left[\sum_{i=1}^{N} X_i^\top (X_i - \mu_{y_i}) \right] + \frac{1}{2}(\mu_1 - \mu_{-1})^\top \cdot \frac{1}{2}(\mu_1 - \mu_{-1}) \\
=&\frac{1}{N} \left[\sum_{i=1}^{N} X_i^\top (X_i - \mu_{y_i}) \right] + \frac{1}{4}(\mu_1 - \mu_{-1})^\top \cdot (\mu_1 - \mu_{-1})
\end{aligned}$$

Now we have two parts, the first part is a bit tricky. Maybe you are curious about the meaning of μ_{y_i}, it is an indicator parameter. We give an explanation of μ_{y_i}, where

$$\mu_{y_i} = \begin{cases} \mu_1 & \text{if } i : y_i^* = 1, \\ \mu_{-1} & \text{if } i : y_i^* = -1. \end{cases}$$

Since $\sum_{i=1}^{N} X_i = \sum_{i=1}^{N} \mu_{y_i}$, then $\frac{1}{N}\left[\sum_{i=1}^{N} -(X_i - \mu_{y_i})\mu_{y_i}^{\top}\right] = 0$. Then the first part add a zero still equals itself.

$$
\begin{aligned}
&\frac{1}{N}\left[\sum_{i=1}^{N} X_i^{\top}\left(X_i - \mu_{y_i}\right)\right] + \frac{1}{N}\left[\sum_{i=1}^{N} -(X_i - \mu_{y_i})\mu_{y_i}^{\top}\right] \\
=&\frac{1}{N}\left[\sum_{i=1}^{N} \left(X_i - \mu_{y_i}\right)^{\top}\left(X_i - \mu_{y_i}\right)\right] \\
=&\frac{1}{N}\left[\sum_{i:y_i^*=-1} (X_i - \mu_{-1})^T \cdot (X_i - \mu_{-1}) + \sum_{i:y_i^*=1} (X_i - \mu_1)^T \cdot (X_i - \mu_1)\right] \\
=&\Sigma.
\end{aligned}
$$

Thus we have proved the RHS

$$
\frac{1}{N}\sum_{i=1}^{N} X_i^{\top} X_i = \Sigma + \frac{1}{4}(\mu_1 - \mu_{-1})^{\top} \cdot (\mu_1 - \mu_{-1}).
$$

Combine the LHS and RHS, we have

$$
\frac{1}{2}\left(\mu_1 - \mu_{-1}\right)^{\top} = \Sigma \cdot \beta + \frac{1}{4}(\mu_1 - \mu_{-1})^{\top} \cdot (\mu_1 - \mu_{-1}) \cdot \beta.
$$

□