CS446: Machine Learning, Fall 2017, Homework 2

Name: Lanxiao Bai (lbai5)

Worked individually

Problem 1

Solution: Since we have model

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \varepsilon$$

by letting

$$\frac{\partial ||\mathbf{y} - \mathbf{X}\mathbf{w}||_2^2}{\partial \mathbf{w}} = 0$$

,
we can get the least square solution

$$\hat{\mathbf{w_{LS}}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Problem 2

Solution: In order to check if the least square estimation is biased, we can calculate the bias of the estimation

$$\begin{aligned}
\text{bias}(\hat{\mathbf{w_{LS}}}) &= \mathbb{E}[\hat{\mathbf{w_{LS}}}] - \mathbf{w} \\
&= \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] - \mathbf{w} \\
&= \mathbb{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w} + \varepsilon)] - \mathbf{w} \\
&= \mathbb{E}[(\mathbf{X}^{-1}(\mathbf{X}^T)^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w} + (\mathbf{X}^{-1}(\mathbf{X}^T)^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w}\varepsilon] - \mathbf{w} \\
&= \mathbb{E}[\mathbf{w}] + (\mathbf{X}^{-1}(\mathbf{X}^T)^{-1}\mathbf{X}^T\mathbb{E}[\varepsilon] - \mathbf{w} \\
&= \mathbf{w} + \mathbf{0} - \mathbf{w} = \mathbf{0}
\end{aligned}$$

So we see that the least square estimator is unbiased.

Problem 3

Solution:

$$
\begin{aligned}
Var(\hat{\mathbf{w_{LS}}}) &= Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\
&= Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w}+\varepsilon)) \\
&= Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w}) + Var((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\varepsilon)) \\
&= Var(\mathbf{w}) + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Var(\varepsilon)((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2 I((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}
$$

Problem 4

Solution: Now that we have the objective function

$$\arg\min_{\mathbf{w}} ||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_2^2$$

similarly, we let

$$\frac{\partial(||\mathbf{y}-\mathbf{X}\mathbf{w}||_2^2 + \lambda||\mathbf{w}||_2^2)}{\partial\mathbf{w}} = 0$$

and get

$$\hat{\mathbf{w}_{\text{ridge}}} = (\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Problem 5

Solution: Similarly, we calculate the bias of $\hat{\mathbf{w}_{\text{ridge}}}$,

$$
\begin{aligned}
\text{bias}(\hat{\mathbf{w}_{\text{ridge}}}) &= \mathbb{E}[\hat{\mathbf{w_{ridge}}}] - \mathbf{w} \\
&= \mathbb{E}[(\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] - \mathbf{w} \\
&= \mathbb{E}[(\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w}+\varepsilon)] - \mathbf{w} \\
&= \mathbb{E}[(\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w}] - \mathbf{w} \\
&= (\mathbf{I}_D + \lambda(\mathbf{X}^T\mathbf{X})^{-1})\mathbf{w} && \textit{Sta}\ (2006)
\end{aligned}
$$

Hence, the estimator is biased.

Problem 6

Solution:

$$
\begin{aligned}
Var(\hat{\mathbf{w}_{\text{ridge}}}) &= Var((\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}) \\
&= Var((\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w}) + \sigma^2((\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)((\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= \sigma^2((\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)((\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= \sigma^2(\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}((\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1})^T
\end{aligned}
$$

Problem 7

Solution:

$$\begin{aligned}
tr(Var(\hat{\mathbf{w_{LS}}})) &= tr(\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \\
&= \sigma^2 tr((\mathbf{X}^T\mathbf{X})^{-1}) \\
&= \sigma^2 tr((\mathbf{V}\mathbf{S}^2\mathbf{V}^T)^{-1}) \\
&= \sigma^2 tr(\mathbf{V}(\mathbf{S}^2)^{-1}\mathbf{V}^{-1}) \\
&= \sigma^2 tr((\mathbf{S}^2)^{-1}) \\
&= \sigma^2 \sum_{i=1}^{n} \frac{1}{s_i^2}
\end{aligned}$$

and

$$\begin{aligned}
tr(Var(\hat{\mathbf{w_{ridge}}})) &= tr(\sigma^2(\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}((\lambda\mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1})^T) \\
&= \sigma^2 tr(\mathbf{V}(\lambda\mathbf{I}_D + \mathbf{S}^2)^{-1}\mathbf{V}^T(\mathbf{V}\mathbf{S}^2\mathbf{V}^T)(\mathbf{V}(\lambda\mathbf{I}_D + \mathbf{S}^2)^{-1}\mathbf{V}^T)^T) \\
&= \sigma^2 tr(\mathbf{V}(\lambda\mathbf{I}_D + \mathbf{S}^2)^{-1}\mathbf{V}^T(\mathbf{V}\mathbf{S}^2\mathbf{V}^T)\mathbf{V}^T((\lambda\mathbf{I}_D + \mathbf{S}^2)^{-1})^T\mathbf{V}) \\
&= \sigma^2 tr((\lambda\mathbf{I}_D + \mathbf{S}^2)^{-1}\mathbf{S}^2(\lambda\mathbf{I}_D + \mathbf{S}^2)^{-1}) \\
&= \sigma^2 tr(((\lambda\mathbf{I}_D + \mathbf{S}^2)^{-1})^2\mathbf{S}^2) \\
&= \sigma^2 \sum_{i=1}^{n} \frac{s_i^2}{\lambda + s_i^2}
\end{aligned} \tag{1}$$

Problem 8

Solution: We see that the bias of $\hat{\mathbf{w_{LS}}}$ is smaller than that of $\hat{\mathbf{w_{ridge}}}$ but the variance of of $\hat{\mathbf{w_{LS}}}$ is larger than that of $\hat{\mathbf{w_{ridge}}}$.

References

(2006). Regularization: Ridge regression and the lasso.
URL http://statweb.stanford.edu/~tibs/sta305files/Rudyregularization.pdf