MSSC 6250 Machine Learning Homework 2

Ridge, Lasso, and Splines

Dr. Cheng-Han Yu

- Deadline: **Friday, Mar 3 11:59 PM**
- Homework presentation date: **Tuesday, Mar 7**
- Please submit your work in **one PDF** file to **D2L > Assessments > Dropbox**. *Multiple files or a file that is not in pdf format are not allowed.*
- Any relevant code should be attached.
- Read **ISL** Chapter 5.1, 6.2, and 7.

Exercises required for all students

1. **ISL** Sec. 5.4: 3

Solution:

a-) k-fold cross-validation is done by taking the set of observations and splitting into k non-overlapping random groups. Each one of these groups then are used as a validation, and the one left as the training set. Then to estimate the test error, we do the average of the k MSE estimates.

b-)

i. The validation set strategy is easier to implement, since it only requires splitting the training data into two groups.However, depending on which observations are included in the training and validation sets, the estimate of the test error rate might vary a lot.

ii. LOOCV is a specific instance of k-fold cross-validation. LOOCV is the approach that is the most expensive computationally wise, since the model needs to be fit n times. Besides that, LOOCV has lower bias but larger variance than the common k-fold cross-validation.

2. **ISL** Sec. 6.6: 3

Solution:

a-)

i. Decrease steadily. As s increases the constraint on beta decreases until it is non-existent. Then the RSS reduces until it becomes the least squares answer.

b-)

ii. Decrease initially, and then eventually start increasing in a U shape. When $s = 0$, all β s are 0, and we have the simplest model possible, which will have higher RSS. As we increase s, $beta$ s become non-zero values and model starts fitting well on test data, which makes RSS decreases. And if we continue increasing s we will end-up overfitting.

c-)

iii. Steadily increase. When $s = 0$ the model is just the intercept, which has very small variance. Once the values of βs start to increase, we will get higher and higher values for variance, since it will start to overfit the model.

d-)

iv. Steadily decrease. When $s = 0$ the model is just the intercept, and likely to be far from the true value. Thus bias is high. As s increases, more β s become non-zero, making our model better at fitting the training data, which then decreases the bias.

e-)

v. Remains constant. The error is irreducible, since it is not dependent on the model, so changing the the βs values, by changing s, will not increase or decrease the irreducible error.

3. **ISL** Sec. 6.6: 4

Solution:

a-)

iii. Steadily increase. As λ increases, βs decreases from the OLS βs until it reaches 0. Then the RSS will increase until it reaches a maximum value, when βs reach 0.

b-)

ii. Decrease initially, and then eventually start increasing in a U shape. When $\lambda = 0$, all β s are the same as OLS. When lambda increases, the fit's flexibility decreases, which lowers the variance of predictions for causing an increase in bias. The test RSS will initially fall as a result of the improved prediction accuracy. Predictions will become more skewed when lambda rises above the ideal point because there is a considerably higher since we will start to overfit.

c-

iv. Steadily decreases. When $\lambda = 0$, The actual estimates depend more on the training data, which means that variance is high. when λ increases, β s start decreasing and model is simplified. The larger λ grows, all βs start to reach 0, which predictions are constant with no variance.

d-)

iii. Steadily increases. When $\lambda = 0$, we have a OLS, which mean the least bias. When λ increases, β s start going towards zero, the model fits less accurately to training data, which causes bias to increase. The larger λ grows, all βs start to reach 0, which predictions are constant and bias is the largest.

e-)

v. Remains constant. The error is irreducible, since it is not dependent on the model, so changing the the βs values, by changing λ, will not increase or decrease the irreducible error.

4. **ISL** Sec. 6.6: 6

Solution:

a-)

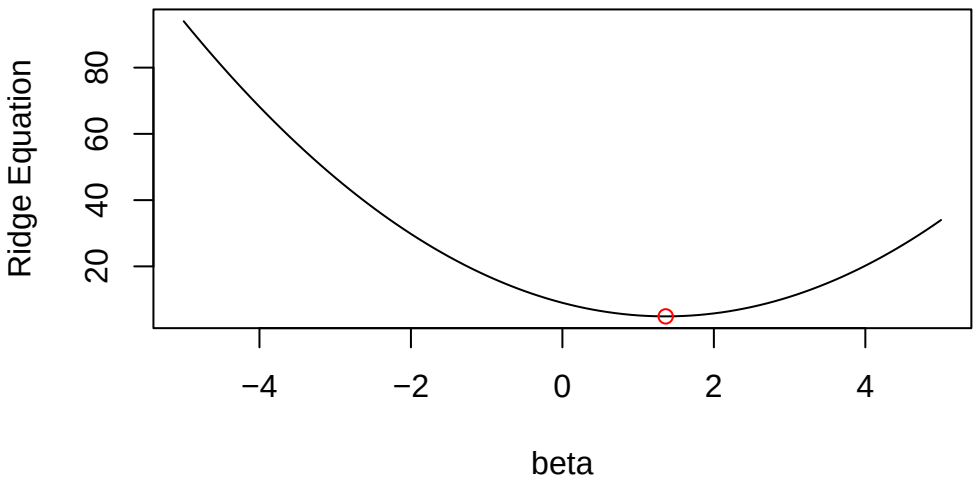
When $p = 1$, (6.12) is equal to $(y - \beta)^2 + \lambda\beta^2$. We plot this function for $y = 2, \lambda = 2$.

```
y = 3
beta = seq(-5,5,0.1)
lambda = 1.2

eq_6.12 = (y - beta)2 + lambda*(beta2)


est_beta_6.14 = y/(1 + lambda)
est_val = (y - est_beta_6.14)2 + lambda*(est_beta_6.142)

plot(beta, eq_6.12, xlab="beta", ylab="Ridge Equation",type="l")
points(est_beta_6.14, est_val,col="red",type ="p")
```



As we can see, the red point show that the minimum value occurs at $\beta = y/(1+\lambda)$.

5. **ISL** Sec. 6.6: 9 (a)-(d)

6. **ISL** Sec. 7.9: 9

7. Suppose for a linear regression problem with $n = p$, $\mathbf{X} = \mathbf{I}$ and no intercept. Show that

   (a) The least squares problem can be simplified to finding $\beta_1, \ldots, \beta_p$ that minimize $\sum_{j=1}^{p} (y_j - \beta_j)^2$. What is least squares estimator b_j?

   (b) The ridge estimator is $\hat{\beta}_j^r = \frac{y_j}{1+\lambda} = \arg\min \sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$.

   (c) The Lasso solution of $\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$ is

$$\hat{\beta}_j^l = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2 \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2 \\ 0 & \text{if } |y_j| \leq \lambda/2 \end{cases}$$

Exercises required for MSSC PhD students

1. **ISL** Sec. 6.6: 5

2. **ISL** Sec. 7.9: 11

3. Define the objective of the **elastic net** problem $J_1(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1$ and the objective of Lasso $J_2(\boldsymbol{\beta}) = \|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|^2 + c\lambda_1\|\boldsymbol{\beta}\|_1$ where $c = (1+\lambda_2)^{-1/2}$, and

$$\tilde{\mathbf{X}} = c\begin{pmatrix}\mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I}_p\end{pmatrix}, \quad \tilde{\mathbf{y}} = \begin{pmatrix}\mathbf{y} \\ \mathbf{0}_{p\times 1}\end{pmatrix}.$$

Show that $J_1(c\boldsymbol{\beta}) = J_2(\boldsymbol{\beta})$. Therefore the elastic net problem can be solved using algorithms for Lasso on modified data.

Optional Exercises

Let the ridge estimator be $\hat{\boldsymbol{\beta}}^r = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$. Show that

1. $\hat{\boldsymbol{\beta}}^r = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})\mathbf{b}$, and hence $\hat{\beta}_j^r = \frac{b_j}{1+\lambda}$ is a special case when $\mathbf{X}'\mathbf{X} = \mathbf{I}$.
2. The bias $\mathrm{E}\left(\hat{\boldsymbol{\beta}}^r\right) - \boldsymbol{\beta} = \left[(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X}) - \mathbf{I}\right]\boldsymbol{\beta}$, and hence $\mathrm{Bias}(\hat{\beta}_j^r) = \frac{-\lambda}{1+\lambda}\beta_j$ when $\mathbf{X}'\mathbf{X} = \mathbf{I}$.
3. $\sum_{j=1}^p \left(\hat{\beta}_j^r - \beta_j\right)^2 = \lambda^2\boldsymbol{\beta}'\left(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}\right)^{-2}\boldsymbol{\beta} = \sum_{j=1}^p \frac{\lambda^2}{(d_j^2+\lambda)^2}\beta_j^2$.
4. $\mathrm{Var}\left(\hat{\beta}_j^r\right) = \sigma^2\frac{d_j^2}{\left(d_j^2+\lambda\right)^2}$ where d_j is the j-th singular value of $\mathbf{X}$. Hence $\mathrm{Var}\left(\hat{\beta}_j^r\right) = \sigma^2\frac{1}{(1+\lambda)^2}$ when $\mathbf{X}'\mathbf{X} = \mathbf{I}$.
5. $\hat{\mathbf{y}}^r = \mathbf{X}\hat{\boldsymbol{\beta}}^r = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} = \sum_{j=1}^p \mathbf{u}_j\left(\frac{d_j^2}{d_j^2+\lambda}\mathbf{u}_j'\mathbf{y}\right)$ where $\mathbf{u}_j$ is the j-th column of the SVD of $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}'$. Hence, a greater amount of shrinkage is applied to the coordinates of basis vectors with smaller d_j^2.
6. The trace of the projection matrix $\mathbf{H}_\lambda = \mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'$ defines the **effective degrees of freedom**. In general, $\mathrm{tr}(H) = \sum_{j=1}^p \frac{d_j^2}{d_j^2+\lambda}$. For linear regression where $\lambda = 0$, $\mathrm{tr}(H) = p$.