DATA 558
SPRING QUARTER 2022

Homework # 3
Due Via Online Submission to Canvas: Wed, May 11 at 5PM

Instructions:

You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, code without written answers will receive no credit. To receive credit, you must explain your answers and show your work. All plots should be appropriately labeled and legible, with axis labels, legends, etc., as needed.

On this assignment, some of the problems involve random number generation. Be sure to set a random seed (using the command `set.seed()`) before you begin.

1. In this problem, we'll see a (very!!) simple simulated example where a least squares linear model is "too flexible".

   (a) First, generate some data with $n = 100$ and $p = 10,000$ features, and a quantitative response, using the following `R` commands:

   ```
   y <- rnorm(100)
   x <- matrix(rnorm(10000*100), ncol=10000)
   ```

   Write out an expression for the model corresponding to this data generation procedure. For instance, it might look something like

   $$Y = 2X_1 + 3X_2 + \epsilon, \qquad \epsilon \sim N(0, 1).$$

   (b) What is the value of the irreducible error?

   (c) Consider a very simple model-fitting procedure that just predicts 0 for every observation. That is, $\hat{f}(x) = 0$ for all x.

      i. What is the bias of this procedure?
      ii. What is the variance of this procedure?
      iii. What is the expected prediction error of this procedure?
      iv. Use the validation set approach to estimate the test error of this procedure. What answer do you get?

v. Comment on your answers to (iii) and (iv). Do your answers agree with each other? Explain.

(d) Now use the validation set approach to estimate the test error of a least squares linear model using $X_1, \ldots, X_{10{,}000}$ to predict Y. What is the estimated test error?

Hint: If you fit a least squares linear model to predict Y using $X_1, \ldots, X_p$ where $p \geq n$, then only the first $n - 1$ coefficients will be assigned values. The rest will show up as `NA` because those coefficients aren't needed to obtain a perfect (i.e. zero) residual sum of squares on the training data. You can see all of the coefficient values by applying the `coef()` command to the output of the linear model.

(e) Comment on your answers to (c) and (d). Which of the two procedures has a smaller estimated test error? higher bias? higher variance? In answering this question, be sure to think carefully about how the data were generated.

2. In lecture during Week 5, we discussed "Option 1" and "Option 2": two possible ways to perform the validation set approach for a modeling strategy where you identify the q features most correlated with the response, and then fit a least squares linear model to predict the response using just those q features. If you missed that lecture, then please familiarize yourself with the lecture notes (posted on `Canvas`) before you continue.

   Here, we are going to continue to work with the simulated data from the previous problem, in order to illustrate the problem with Option 1.

   (a) Calculate the correlation between each feature and the response. Make a histogram of these correlations. What are the values of the 10 largest absolute correlations?

   (b) Now try out "Option 1" with $q = 10$. What is the estimated test error?

   (c) Now try out "Option 2" with $q = 10$. What is the estimated test error?

   (d) Comment on your results in (b) and (c). How does this relate to the discussion of Option 1 versus Option 2 from lecture? Explain how you can see that Option 1 gave you a useless (i.e. misleading, inaccurate, wrong) estimate of the test error.

3. In this problem, you will analyze a (real, not simulated) dataset of your choice with a quantitative response Y, and $p \geq 50$ quantitative predictors.

   (a) Describe the data. Where did you get it from? What is the meaning of the response, and what are the meanings of the predictors?

   (b) Fit a least squares linear model to the data, and provide an estimate of the test error. (Explain how you got this estimate.)

(c) Fit a ridge regression model to the data, with a range of values of the tuning parameter λ. Make a plot like the left-hand panel of Figure 6.4 in the textbook.

(d) What value of λ in the ridge regression model provides the smallest estimated test error? Report this estimate of test error. (Also, explain how you estimated test error.)

(e) Repeat (c), but for a lasso model.

(f) Repeat (d), but for a lasso model. Which features are included in this lasso model?

In this problem, you may use the function in the `glmnet` package that performs cross-validation.

4. Consider using the `Auto` data set to predict `mpg` using polynomial functions of `horsepower` in a least squares linear regression.

(a) Perform the validation set approach, and produce a plot like the one in the right-hand panel of Figure 5.2 of the textbook. Your answer won't look *exactly* the same as the results in Figure 5.2, since you'll be starting with a different random seed. Discuss your findings. What degree polynomial is best, and why?

(b) Perform leave-one-out cross-validation, and produce a plot like the one in the left-hand panel of Figure 5.4 of the textbook. Discuss your findings. What degree polynomial is best, and why?

(c) Perform 10-fold cross-validation, and produce a plot like the one in the right-hand panel of Figure 5.4 of the textbook. Discuss your findings. What degree polynomial is best, and why?

(d) Fit a least squares linear model to predict `mpg` using polynomials of degrees from 1 to 10, using all available observations. Make a plot showing "Degree of Polynomial" on the x-axis, and "Training Set Mean Squared Error" on the y-axis. Discuss your findings.

(e) Fit a least squares linear model to predict `mpg` using a degree-10 polynomial, using all available observations. Using the `summary` command in `R`, examine the output. Comment on the output, and discuss how this relates to your findings in (a)–(d).

5. **Extra Credit!** Let's consider doing least squares and ridge regression under a very simple setting, in which $p = 1$, and $\sum_{i=1}^n y_i = \sum_{i=1}^n x_i = 0$. We consider regression without an intercept. (It's usually a bad idea to do regression without an intercept, but if our feature and response each have mean zero, then it is okay to do this!)

(a) The least squares solution is the value of $\beta \in \mathbb{R}$ that minimizes

$$\sum_{i=1}^n (y_i - \beta x_i)^2.$$

Write out an analytical (closed-form) expression for this least squares solution. Your answer should be a function of $x_1, \dots, x_n$ and $y_1, \dots, y_n$. *Hint: Calculus!!*

(b) For a given value of λ, the ridge regression solution minimizes

$$\sum_{i=1}^{n} (y_i - \beta x_i)^2 + \lambda \beta^2.$$

Write out an analytical (closed-form) expression for the ridge regression solution, in terms of $x_1, \dots, x_n$ and $y_1, \dots, y_n$ and λ.

(c) Suppose that the true data-generating model is

$$Y = 3X + \epsilon,$$

where ϵ has mean zero, and X is fixed (non-random). What is the expectation of the least squares estimator from (a)? Is it biased or unbiased?

(d) Suppose again that the true data-generating model is $Y = 3X + \epsilon$, where ϵ has mean zero, and X is fixed (non-random). What is the expectation of the ridge regression estimator from (b)? Is it biased or unbiased? Explain how the bias changes as a function of λ.

(e) Suppose that the true data-generating model is $Y = 3X + \epsilon$, where ϵ has mean zero and variance σ^2, and X is fixed (non-random), and also $\text{Cov}(\epsilon_i, \epsilon_{i'})$= 0 for all $i \neq i'$. What is the variance of the least squares estimator from (a)?

(f) Suppose that the true data-generating model is $Y = 3X + \epsilon$, where ϵ has mean zero and variance σ^2, and X is fixed (non-random), and also $\text{Cov}(\epsilon_i, \epsilon_{i'})$= 0 for all $i \neq i'$. What is the variance of the ridge estimator from (b)? How does the variance change as a function of λ?

(g) In light of your answers to parts (d) and (f), argue that λ in ridge regression allows us to control model complexity by trading off bias for variance.

Hint: For this problem, you might want to brush up on some basic properties of means and variances! For instance, if $Cov(Z, W) = 0$, then $Var(Z + W) = Var(Z) + Var(W)$. And if a is a constant, then $Var(aW) = a^2 Var(W)$, and $Var(a + W) = Var(W)$.