Stat 462/862 Assignment 2

(Due on 11:50pm eastern time on Oct 25, 2023)

1. Consider the data set *Auto* in the R package *ISLR*. We wish to develop a model to predict whether or not a given car gets high or low gas mileage.

   (a) Create a binary variable, *mpg*01, that contains a 1 if *mpg* contains a value above this median, and a 0 contains a value below its median.

   (b) Explore the data graphically in order to investigate the association between *mpg*01 and the other features. Which of the other features seem most likely to be useful in predicting *mpg*01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

   (c) Split the data into a training set and a test set.

   (d) Perform LDA on the training data to predict *mpg*01 using the variables that seemed most associated with *mpg*01 in (b). Report all the parameter estimates. What is the test error of the model obtained?

   (e) Perform QDA on the training data to predict *mpg*01 using the variables that seemed most associated with *mpg*01 in (b). Report all the parameter estimates. What is the test error of the model obtained?

   (f) Using LDA, and QDA to estimate the probability that a dodge challenger se car with the following setting (cylinders = 6, displacement = 400, horsepower = 110, weight = 3000, acceleration = 15, year = 75, origin = 2) gets high gas mileage.

2. Consider the SPAM datasets which include training data and test data. See the files SPAM_training.csv and SPAM_test.csv. In both files, the first 57 columns are the predictors and the last column is the response.

   (a) Fit the data using the first-order linear model. Obtain the estimated regression coefficients.

   (b) Fit the data using ridge regression. Obtain the corresponding estimated regression coefficients.

   (c) Obtain the test errors and their standard errors for the first order linear model and the ridge regression.

3. For dataset **Boston** in the R package *MASS*, consider the followings.

   (a) Apply the best subset selection method and propose a model.

   (b) Apply the ridge regression method and propose a model.

4. Generate the data $\{X_{i1}, X_{i2}, X_{i3}, Y_i\}_{i=1}^{50}$ using the model $Y_i = 10 - 5X_{i1} + 20.6X_{i2} + 3.4X_{i3} + \epsilon_i$, where $X_{i1} \sim Unif(1, 10), X_{i2} \sim Unif(-1, 10), X_{i3} \sim Unif(0, 4), \epsilon_i \sim N(0, 2)$. Now add five more predictor variables $Z_1 = -2.1X_1X_2, Z_2 = -5.9X_1X_3, Z_3 = 6X_2X_3, Z_4 \sim N(20, 50), Z_5 \sim N(5, 1)$.

   (a) Show the Lasso solution path?

   (b) What is the tuning parameter that minimizes the cross-validation error? What is the corresponding minimum cross-validation error?

   (c) What are the significant variables chosen by the Lasso? Interpret the result.

   (d) What is the fitted model?

5. (Graduate students only) Consider a dataset (X, Y) in which Y is output and X represents inputs. Let n be the number of observations and p be the number of input variables in the dataset. Consider a special case $n = p = 1$. The ridge regression aims to minimize

$$\sum_{i=1}^{n}(Y_i - \beta_0 - (\sum_{j=1}^{p} X_{ij}\beta_j))^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1}$$

while the lasso minimizes

$$\sum_{i=1}^{n}(Y_i - \beta_0 - (\sum_{j=1}^{p} X_{ij}\beta_j))^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

Suppose in both ridge regression and lasso, the intercept is omitted from model. Thus there is only one regression coefficient denoted by β.

   (a) Choose a few random values of Y and λ, plot (1) and (2) as a function of β, and find their minima on the graphs. Verify that these minima are attached at

$$\hat{\beta}_{ridge} = \frac{Y}{1 + \lambda}$$

   and

$$\hat{\beta}_{lasso} = \begin{cases} Y - \frac{\lambda}{2}, & \text{if } Y > \frac{\lambda}{2} \\ Y + \frac{\lambda}{2}, & \text{if } Y < -\frac{\lambda}{2} \\ 0, & \text{otherwise.} \end{cases}$$

(b) Choose a few random values of Y, and for each value of Y, plot $\hat{\beta}_{ridge}$ and $\hat{\beta}_{lasso}$ on the same axes, as functions of λ. Describe the observations from the plots.