Statistical Learning and Data Mining, Midterm Exam 2018/11/20

1. Consider the "Data1" data set. Let y be the response variable and V_1, V_2, V_3, V_4, V_5 be the predictors.

   (a) Find the best linear regression model using the best subset selection method with the C_P criteria for the training data. Based on the predictors selected above, include their interaction terms and rebuild the linear regression model using the best subset selection method with C_p criteria. Write down the final fitted model, and its training and test mean square errors.

   (b) Based on the five predictors $V_1, \ldots, V_5$ and their interaction terms, find the lasso regression model with the tuning parameter λ which minimizes the 10-fold cross validation error. Write down the fitted lasso regression model, the training and test mean square errors.

   (c) Based on the results of (a)–(b), which model will you suggest to use for future prediction? Justify your answer.

2. Consider the "Data2" data set. Let Y be the categorical response variable and $V_1, \ldots, V_{495}$ be the predictors.

   (a) Use the LDA method with the 100 most correlated predictors to predict Y for the training data. Find the training and test classification errors.

   (b) Use the QDA method with the 100 most correlated predictors to predict Y for the training data. Find the training and test classification errors.

   (c) Using the 100 most correlated predictors, perform KNN on the training data. Provide a table of the leave-one-out cross validation error and test error for $1 \leq K \leq 10$. Which value of K attains the smallest cross validation classification error? Plot the training, cross validation and test classification errors versus K ($1 \leq K \leq 10$) in a single figure. Make sure to label the training, cross validation and test error curves. (Hint : Use `knn.cv()`)

   (d) Find lasso logistic model for Y with the tuning parameter λ minimizing the 10-fold cross validation classification error. What are the training and test errors? (Hint : Use `cv.glmnet()` with `family="binomial"` and predict the class with the maximum probability)

   (e) Based on the results of (a)–(d), which model will you suggest to use for future prediction? Justify your answer.

3. Consider the "Data3" data set. Let $y=$ ViolentCrimesPerPop be the response variable and rest of the variables be the predictors.

   (a) Fit the lasso regression model using the tuning parameter λ which minimizes the 10-fold cross validation error. What are the training and test mean square errors?

   (b) Fit the ridge regression model using the tuning parameter λ which minimizes the 10-fold cross validation error. What are the training and test mean square errors?

   (c) Based on the results of (a)–(b), which model will you suggest? Why?

4. Consider the "Data4" data set. Let y be the response variable and $V_1, \ldots, V_{5000}$ be the predictors.

   (a) First find the 100 most correlated predictors based on all data. Then divide the data into 10-folds, and build the linear regression model of y on these 100 predictors to evaluate the 10-fold cross validation error.

   (b) First divide the data into 10-folds. Then find the 100 most correlated predictors based on the training set of each fold. Finally evaluate the 10-fold cross validation error.

   (c) Which result would you prefer if the goal is to evaluate the prediction error of the linear regression model of y on the 100 most correlated predictors. Why?