Statistical and Machine Learning (Spring 2019)

Mini Project 3

Instructions:

- Due date: March 6, 2019.
- Total points = 20.
- Submit a typed report.
- It is OK to discuss the project with other students in the class, but each student must write their own code and answers. If your submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job.
- You must use the following template for your report:

  Mini Project #
  Name
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.

---

1. Consider the business school admission data available on eLearning. The admission officer of a business school has used an "index" of undergraduate grade point average (GPA, X_1) and graduate management aptitude test (GMAT, X_2) scores to help decide which applicants should be admitted to the school's graduate programs. This index is used categorize each applicant into one of three groups — admit (group 1), do not admit (group 2), and borderline (group 3). We will take the last five observations in each category as test data and the remaining observations as training data.

   (a) Perform an exploratory analysis of data by examining appropriate plots and comment on how helpful these predictors may be in predicting response.

   (b) Perform an LDA and provide an equation for the decision boundary. Superimpose the decision boundary on an appropriate display of the data. Does the decision boundary seem sensible? In addition, compute the confusion matrix and overall misclassification rate based on both training and test data. What do you observe?

   (c) Repeat (b) using QDA.

   (d) Repeat (b), with the exception of providing the decision boundary equation, using KNN with K chosen optimally using the test data.

   (e) Compare the results in (b)–(d). Which classifier would you recommend? Justify your conclusions.

2. Annual financial data are collected for bankrupt firms approximately two years prior to their bankruptcy and for financially sound firms at about the same time. The data on four variables, $X_1 = CF/TD$ =(cash flow)/(total debt), $X_2 = NI/TA$ =(net income)/(total assets), $X_3 = CA/CL$ =(current assets)/(current liabilities), and $X_4 = CA/NS$ =(current assets)/(net sales) are given on eLearning. This is a binary classification problem. Take bankrupt firm as "+" response (indicated as 0 in the data) and nonbankrupt firm as "–" response (indicated as 1 in the data). Use all the data as training data.

(a) Perform an exploratory analysis of data by examining appropriate plots and comment on how helpful these predictors may be in predicting response.

(b) Build an appropriate logistic regression model for these data. Interpret the estimated regression coefficients in the final proposed model.

3. Consider the bankruptcy data of the previous problem.

(a) Use the logistic regression model built in the previous problem to provide an equation for the decision boundary for the classification problem. In addition, compute the confusion matrix, sensitivity, specificity, and overall misclassification rate, and plot the ROC curve. What do you observe?

(b) Repeat (a) with all predictors (not just those appearing the final proposed model from the previous problem). Comment on whether there is any benefit in performing variable selection over using all predictors.

(c) Repeat (a) with all predictors using LDA.

(d) Repeat (a) with all predictors using QDA.

(e) Compare the results in (a)–(d). Which classifier would you recommend? Justify your conclusions.