ST790: Homework 5

Due: 10/22/2019

General Instructions:

- All the HW files (except the R code) should be saved as PDF, and named in the form Lastname_Firstname_hw1.pdf".
- The code should be saved as "Lastname_Firstname_hw1_prob1_code.r".
- Test your R code before submission to make sure it can be executed successfully by the "source()" function.
- It is ok to turn in multiple files.

1. (**LDA and Logistic Regression for Binary Classification**)

   (a) Fit the LDA for the training data in Scenario 1 (HW4). Report the training and testing errors.

   (b) Fit the logistic regression for the training data in Scenario 1. Report the training and testing errors.

   (c) Compare the results (a) and (b) with those of Bayes rule and linear model fit (you have done in HW 4). Make a summary.

2. (**Two-Class Classification Problem: Scenario 2**) (Textbook page 17)

   Generate a training set of $n = 200$ from a mixture data as follows.

step 1: Generate 10 points $\mu_k, k = 1, ..., 10$ from a bivariate Gaussian distribution $N((1,0)^T, \mathbf{I})$. They will be used as means (centers) to generate the **Green** class for both training and test data.

step 2: Generate 10 points $\nu_k, k = 1, ..., 10$ from a bivariate Gaussian distribution $N((0,1)^T, \mathbf{I})$. They will be used as means (centers) to generate the **Red** class.

step 3: For the **Green** class, generate 100 observations as follows: for each observation, randomly pick a μ_k with probability 1/10, and then generate a point from $N(\mu_k, \mathbf{I}/5)$.

step 4: For the **Red** class, generate 100 observations as follows: for each observation, randomly pick a ν_k with probability 1/10, and then generate a point from $N(\nu_k, \mathbf{I}/5)$.

   (a) Generate the training set.

   (b) Draw the scatter plot of the training set, using different labels/colors for two classes.

   (c) Generate a test set, with 500 observations from each class, using *set.seed(2014)*. The same center parameters are used in the training and test sets. Save the test set for future use.

   Submit the scatter plot.

3. (**Linear, LDA and QDA Methods for Classification in Scenario 2**)

   (a) Train the linear regression model, using the function "lm(y∼x)'," with the training set.

   (b) Add the linear decision boundary to the scatterplot.

   (c) Report the training and test errors for this linear classification rule.

   (d) Fit the LDA and QDA for the training data in Scenario 2. Report the corresponding training and testing errors.

4. (**k-Nearest Neighbor for Classification: Scenario 2**)

   (a) Fit k-nearest neighbor classifier with a range of values k for the training data generated under Scenario 2, $k = \{1, 4, 7, 10, 13, 16, 30, 45, 60, 80, 100, 150, 200\}$. Report both training and testing errors for each k-NN classifier. Plot two curves: the training error vs the degree of freedom n/k, and the testing error vs n/k, in one same figure (Similar to Figure 2.4 in the textbook).

   (b) Based on the plots obtained in (a), how should you choose the best k?

5. Classify the 1's, 2's, 3's for the zip code data in the textbook.

   (a) Use the k-nearest neighbor classification with $k = 1, 3, 5, 7, 15$. Show both the training and test error for each choice.

   (b) Implement the LDA and QDA methods, and report there training and testing errors.

   Note: Before carrying out the LDA analysis, you are suggested to delete variable 16 first from the data, since the variable takes a constant value and it can cause the singularity of the covariance matrix. In general, a constant variable does not have a discriminating power to separate two classes.