STA 545 Statistical Data Mining I, Fall 2020

Homework 6, due: Wednesday 10/14/2020 (1PM)

Please submit the pdf file generated by R markdown in UBlearns. Please use tables, figures, or a few sentences to answer data analysis questions.

1. (25 points) Consider a binary classification problem with only one input variable. For the first class, the input variable is generated from a normal distribution with mean 1 and standard deviation 1. For the second class, the input variable is generated from a normal distribution with mean -1 and standard deviation 1. Assume that the prior probability of each class is 0.5. Please derive the Bayes Rule and calculate the Bayes error.

2. (25 points) Please show that $\sum_{k=1}^{K} \hat{f}_k(x) = 1$ if we use an intercept term in the multivariate linear regression model for a K-class classification problem.

3. (50 points) Evaluate the classification performance of LDA on the zipcode data. In particular, consider the 2's, 3's, and 4's. Show both the training and test error. The zipcode data are available from the book website www-stat.stanford.edu/ElemStatLearn. Please extract 2's, 3's, and 4's from the training data and the test data downloaded from the above link as the training and test datasets for this homework question. **For the LDA method, please write your own R function rather than using any package for LDA in R.**