STAT 6340 (Statistical and Machine Learning, Spring 2020)

Mini Project 3

Instructions:

- Due date: Mar 4, 2020.
- Total points = 30.
- Submit a typed report.
- It is OK to discuss the project with others and to search on the internet, 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.

- Section 1 of the report must be limited to **five** pages. Also, only those output should be provided in this section that are referred to in the report.

---

1. Consider the German credit dataset from Mini Project 2. We will take `Default` as the response and would like to understand how it is related with other variables in the data.

   (a) Perform an exploratory analysis of data.

   (b) Build a "reasonably good" logistic regression model for these data. There is no need to explore interactions. Carefully justify all the choices you make in building the model.

   (c) Write the final model in equation form. Provide a summary of estimates of the regression coefficients, the standard errors of the estimates, and 95% confidence intervals of the coefficients. Interpret the estimated coefficients of at least two predictors. Provide training error rate for the model.

2. Consider the German credit dataset from #1. Although in the class we have discussed `cv.glm` for computing test error rates using cross-validation, you may use the `caret` package (`https://topepo.github.io/caret/`) for doing so as it is not restricted to the GLMs.

   (a) Fit a logistic regression model using all predictors in the data. Provide its error rate, sensitivity, and specificity based on training data.

   (b) Write your own code to estimate the test error rate of the model in (a) using LOOCV.

   (c) Verify your results in (b) using a package. Make sure the two results match.

   (d) For the logistic regression model you proposed in #1, estimate the test error rate using LOOCV.

   (e) Repeat (d) using LDA from Mini Project #2.

(f) Repeat (d) using QDA from Mini Project #2.

(g) Fit a KNN with K chosen optimally using the LOOCV test error rate. Repeat (d) for the optimal KNN. (You may explore `tune.knn` function for finding the optimal value of K but this is not required.)

(h) Compare the results from the various classifiers. Which classifier would you recommend? Justify your answer.

3. Consider the wine quality dataset from `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`. We will focus only on the data concerning white wines (and not red wines). Dichotomize the `quality` variable as `good`, which takes the value 1 if `quality` ≥ 7 and the value 0, otherwise. We will take `good` as response and all the 11 physiochemical characteristics of the wines in the data as predictors. Use 10-fold cross-validation for estimating the test error rates below and compute the estimates using `caret` package with seed set to `1234` before each computation.

(a) Fit a KNN with K chosen optimally using test error rate. Report error rate, sensitivity, specificity, and AUC for the optimal KNN based on the training data. Also, report its estimated test error rate.

(b) Repeat (a) using logistic regression.

(c) Repeat (a) using LDA.

(d) Repeat (a) using QDA.

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