STAT 6340 (Statistical and Machine Learning, Fall 2024)

Mini Project 3

Instructions:

- Due date: Oct 2, 2024 (Use Submission Link on eLearning to upload by 11:59 pm)
- Total points = 30
- Submit a typed report.
- It is OK to consult with others regarding this project or use any publicly available resources, but you must write code and answers on your own. Violation of this policy will be considered academic dishonesty, and you will referred to appropriate university authorities. If you don't do the work by yourself, how will you learn?
- 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: 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 *three* pages. Also, only those software outputs should be provided in this section that are referred to in the report.
- You can code in `R` or `Python`.
- You are encouraged to use `R Markdown` or `Jupyter Notebook` to write your report.
- Bonus (4 points): Use *both* `R` and `Python` to code, compare the results, and provide an explanation if the results are not identical. This work should be described on a separate page titled Bonus Work with two sections — one to describe the comparison and the other for the code. The first section should be limited to one page.

In this project, we will work with the Pima Indians Diabetes dataset from Mini Project 2 with one change: all the 768 observations in the data will be taken as the training data. There is no separate test set. Standardize all predictors before doing any analysis.

1. (10 points) Analyze the data as follows.

   (a) (7 points) Recall the exploratory data analysis that you performed previously. 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.

   (b) (3 points) Provide coefficient estimates of the final model in a table. Interpret the estimated coefficients of at least two predictors. Provide training error rate for the model.

2. (20 points) Analyze the data as follows. You may use `caret` package in `R` or something similar in `python` to fit the model and compute performance measures.

(a) (2 points) Fit a logistic regression model using all predictors in the data. Provide its error rate, sensitivity, and specificity based on training data. Also draw its ROC curve and present an estimate of AUC based on training data.

(b) (2 points) Write your own code to estimate the test error rate of the model in (a) using LOOCV.

(c) (2 points) Verify your results in (b) using a package. Make sure the two results match.

(d) (2 points) Perform an LDA of the data. Provide its error rate, sensitivity, and specificity based on training data. Also draw its ROC curve and present an estimate of AUC based on training data. Estimate the test rate using LOOCV.

(e) (2 points) Repeat (d) using QDA.

(f) (2 points) Repeat (d) using naive Bayes classifier.

(g) (2 points) Repeat (d) for the logistic regression model you proposed in #1.

(h) (3 points) Fit a KNN with K chosen optimally using the LOOCV estimate of the test error rate. Repeat (d) for the optimal KNN. (You may explore `tune.knn` function in `R` or something similar in `python` for finding the optimal value of K but this is not required.)

(i) (3 points) Compare the results from various classifiers. Which classifier would you recommend? Justify your answer.

loocv b and c?