STAT 4360 (Introduction to Statistical Learning, Fall 2021)

Mini Project 2

Instructions:

- Due date: Sep 22, 2021.
- Total points = 30
- 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.

- 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 wine data available on eLearning as `wine.txt`. The data come from a study of Pinot Noir wine quality. The dataset contains 38 observations and 7 variables: `Quality`, `Clarity`, `Aroma`, `Body`, `Flavor`, `Oakiness`, and `Region`. We will take `Quality` as the response variable and the remaining variables (which represent features of the wine) as predictors. Be sure to treat `Region` as a qualitative predictor. The goal is to develop a model that relates the quality of Pinot Noir with its features. The model can potentially be used to predict the quality of the wine.

   (a) Perform an exploratory analysis of data. Comment on findings that interest you.

   (b) Is `Quality` appropriate as a response variable or a transformation is necessary? In case a transformation is necessary, find a simple transformation and use it for the rest of this problem.

   (c) Do part (a) of Exercise 15 in Chapter 3 for these data.

   (d) Do part (b) of Exercise 15 in Chapter 3 for these data.

   (e) Build a "reasonably good" multiple regression model for these data. Be sure to explore interactions of `Region` with other predictors. Carefully justify all the choices you make in building the model and verify the model assumptions.

   (f) Write the final model in equation form, being careful to handle the qualitative predictors and interactions (if any) properly.

   (g) Use the final model to predict the `Quality` of a wine from `Region` 1 with other predictors set equal to their sample means. Also provide a 95% prediction interval for the response and a 95% confidence interval for the mean response. Interpret the results.

2. Consider the business school admission data available on eLearning as `admission.csv`. 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 first five observations in each category as test data and the remaining observations as training data. Read the handout regarding drawing multiclass decision boundaries before proceeding any further.

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

   (b) Perform an LDA using the training data. 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) Compare the results in (b) and (c). Which classifier would you recommend? Justify your conclusions.

3. Consider the diabetes dataset available on eLearning as `diabetes.csv`. These data are from `https://www.kaggle.com/johndasilva/diabetes?select=diabetes.csv`. You can read more about the data, including a description of the variables, on this website. We will take `Outcome` as the response, the other variables as predictors, and all the data as training data.

   (a) Perform an exploratory analysis of the data. Comment on findings that interest you.

   (b) Perform an LDA of the data. Compute the confusion matrix, sensitivity, specificity, and overall misclassification rate based on 0.5 cutoff for the posterior probability. Plot the ROC curve. What do you observe?

   (c) Repeat (a) using QDA.

   (d) Compare the results from (a) and (b). Which classifier would you recommend? For the recommended classifier what posterior probability cutoff would you suggest? Justify your answer.