MATH 189 Homework 4

Due Feb 10th, 2023

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set. This data can be found in the ISLR package.

```
> library(ISLR)
> data(Auto)
```

(1) Create a binary variable, mgp01, that contains 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.

(2) Explore the data graphically in order to investigate the association between mgp01 and the other features. Which of the other features seem most likely to be useful in predicting mgp01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

(3) Split the data into a training set of size 300 and a test set of size 92.

(4) Perform LDA on the training data in order to predict mgp01 using the variables that seemed most associated with mgp01 in (b). What is the test error of the model obtained?

(5) Perform QDA on the training data in order to predict mgp01 using the variables that seemed most associated with mgp01 in (b). What is the test error of the model obtained?

Hint:

(a) The test error rate associated with a set of test observations $\{(x_i^{\mathrm{t}}, y_i^{\mathrm{t}})\}_{i=1}^{n_{\mathrm{t}}}$ is

$$\frac{1}{n_{\mathrm{t}}} \sum_{i=1}^{n_{\mathrm{t}}} I(y_i^{\mathrm{t}} \neq \hat{y}_i^{\mathrm{t}}),$$

where $\hat{y}_i^{\mathrm{t}}$ is the predicted class label that results from applying the classifier to the test observation with predictor x_i^{t}.

(b) You may use R built-in packages to solve the problem. Check the functions that implement LDA and QDA in the R package MASS.