Homework 3

Due on 03/25/2022

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the dataset "auto.csv". The dataset contains 392 observations. The response variable is mpg_cat, which indicates whether the miles per gallon of a car is high or low. The predictors are:

- cylinders: Number of cylinders between 4 and 8
- displacement: Engine displacement (cu. inches)
- horsepower: Engine horsepower
- weight: Vehicle weight (lbs.)
- acceleration: Time to accelerate from 0 to 60 mph (sec.)
- year: Model year (modulo 100)
- origin: Origin of car (1. American, 2. European, 3. Japanese)

Split the dataset into two parts: training data (70%) and test data (30%).

(a) Produce some graphical or numerical summaries of the data.

(b) Perform a logistic regression using the training data. Do any of the predictors appear to be statistically significant? If so, which ones? Compute the confusion matrix and overall fraction of correct predictions using the test data. Briefly explain what the confusion matrix is telling you.

(c) Train a multivariate adaptive regression spline (MARS) model using the training data.

(d) Perform LDA using the training data. Plot the linear discriminants in LDA.

(e) Which model will you use to predict the response variable? Plot its ROC curve using the test data. Report the AUC and the misclassification error rate.