Introduction

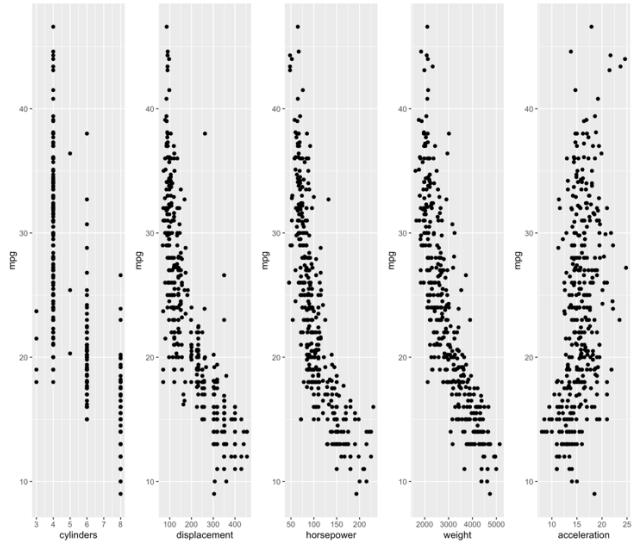
We want to predict whether a given car gets high or low gas mileage based 7 car attributes such as cylinders, displacement, horsepower, weight, acceleration, model year and origin. This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition

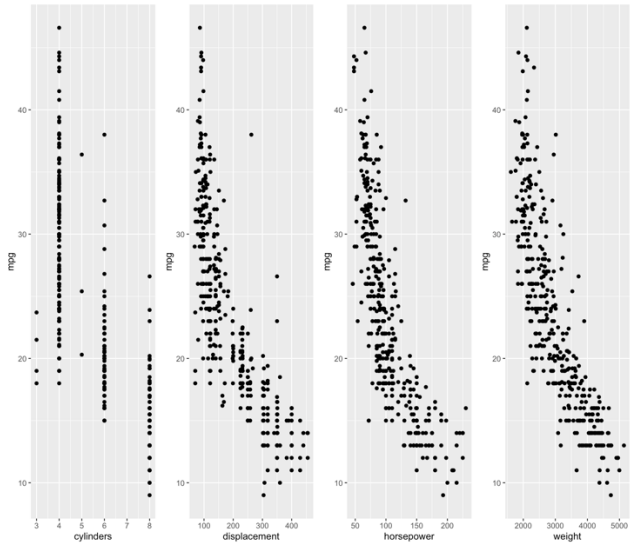
This data set consists of records for 392 cars manufactured between 1970 and 1982. Their weights vary between 1613 and 5140 lbs, their horsepower vary between 46 and 230, and their displacement vary between 68 and 455. 245 of these cars were manufactured in the USA, 68 in Asia and 79 in Europe. Each record consist of 8 variables: mpg01 (millage per gallon of gas (high or low)), cylinders, displacement, horsepower, weight, acceleration, model year and origin.

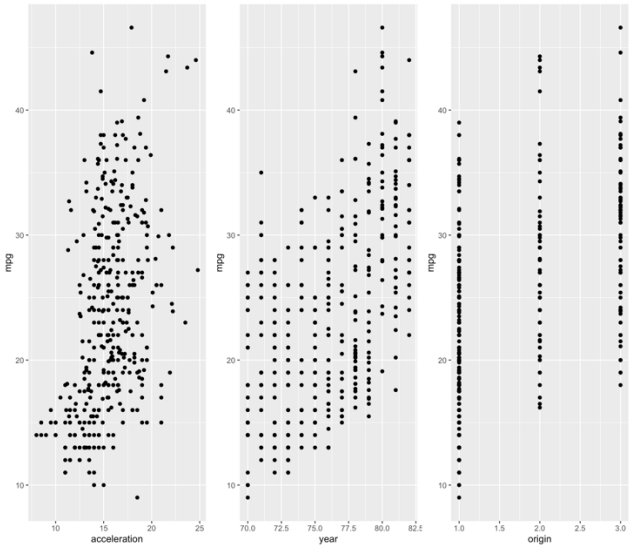
Exploratory (preliminary) Data Analysis

From the initial data analysis for the data, you can notice that the high correlation between the displacement and all other variables, for example the correlation between the displacement and the number of cylinders is 0.9508233, and between the displacement and weight of the car is 0.9329944. Also, you can notice that the correlation between the displacement and all other variables is higher that the correlation between the number of cylinders and all other variables. But some algorithms will build unstable models if two or more highly correlated variables are included in the model, and others will just slow down. Either way, it is a good idea to remove highly (linearly) correlated variables. Therefore, removing the displacement from the models might be a good action. Also, you can notice that the correlation between mpg01 and the acceleration (0.3468215), is the lowest correlation between the mpg01 and all other variables. So, the acceleration doesn't seem to be a good predictor. I think that cylinder, horsepower, weight, year, and origin seem most likely to be useful in predicting mpg01.

Explore the data graphically in order to investigate the association between mpg01 and the other features.







Methods

(i) Linear Discriminant Analysis.

(ii) Quadratic Discriminant Analysis

(iii) Naive Bayes Classifier

(iv) Logistic Regression

(v) K-Nearest Neighbors

Results

For these exercises, we implemented several classification methods with a training data set and those models were used to predict the response in a testing data set.

Model	Avg-TE
Linear Discriminant Analysis	0.08926
Quadratic Discriminant Analysis	0.09879
Naive Bayes Classifier	0.09865
Logistic Regression	0.09326
KNN 1	0.11603
KNN 3	0.12686
KNN 5	0.14694

Model	Mean-TE	Var-TE
Linear Discriminant Analysis	0.09102564	0.001837906
Quadratic Discriminant Analysis	0.09846154	0.002081565
Naive Bayes Classifier	0.09564103	0.002149769
Logistic Regression	0.09487179	0.001613771
KNN 1	0.1338462	0.002185962
KNN 3	0.1266667	0.002295406
KNN 5	0.1269231	0.002395752

Findings

Intuitively, the linear discriminant classifier has the smaller mean test error value than all other methods over B = 100 iterations. An important statistical question here is whether this is statistically significant or not. That is, whether the 100 test error values of linear discriminant classifier are stochastically smaller than those of other methods or not, say, at the significant level $\alpha = 5\%$.
We can verify that linear discriminant classifier shows reasonable significance compared to all other methods except Linear Regression. In fact, if we plot the boxplots of testing errors for linear discriminant classifier and the linear regression models, then it turns out that both have very similar distributions of testing errors for 100 iterations. (The difference between the two looks like normal noise).

The Naive Bayes classifier and the quadratic discriminant classifier have better prediction than all test KNN classifiers, but they are worse than both of the linear discriminant classifier and the linear regression. Although, the Naive Bayes classifier is slightly better than the quadratic discriminant classifier.

For KNN classifiers, you can notice that when k = 3, it out perform the KNN when k = 1 and when k = 5, and when k = 7, it out perform the KNN when k = 5 and when k = 9. Also, you can notice that the best results were obtained when k=51, therefore, the optimal k value is between 25 and 75 with high probability (it could any number between 1 and 353), but testing all of these values is time consuming, and it is not important since the linear discriminant classifier clearly way outperform the KNN classifiers.