Statistical Learning

Exercises sheet 3 – Classification

1. Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergrad GPA (meaning *grade point average*), and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

   (a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

   (b) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

2. 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.

   (a) Create a binary variable, *mpg01*, that contains a 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.

   (b) Explore the data graphically in order to investigate the association between *mpg01* and the other features. Which of the other features seem most likely to be useful in predicting *mpg01*? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

   (c) Split the data into a training set and a test set.

   (d) Perform LDA on the training data in order to predict *mpg01* using the variables that seemed most associated with *mpg01* in (b). What is the test error of the model obtained?

   (e) Perform QDA on the training data in order to predict *mpg01* using the variables that seemed most associated with *mpg01* in (b). What is the test error of the model obtained?

   (f) Perform logistic regression on the training data in order to predict *mpg01* using the variables that seemed most associated with *mpg01* in (b). What is the test error of the model obtained?

   (g) Perform KNN on the training data, with several values of K, in order to predict *mpg01*. Use only the variables that seemed most associated with *mpg01* in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

3. Using the *Boston* data set, fit classification models in order to predict whether a given suburb has a crime rate above or below the median. Explore logistic regression, LDA, and KNN models using various subsets of the predictors. Describe your findings.