Statistical Data Mining I
Homework 3

1) The insurance company benchmark data set gives information on customers. Specifically, it contains 86 variables on product-usage data and socio-demographic data derived from zip area codes. There are 5,822 customers in the training set and another 4,000 in the test set. The data were collected to answer the following questions: Can you predict who will be interested in buying a caravan insurance policy and give an explanation why? Compute the OLS estimates and compare them with those obtained from the following variable-selection algorithms: Forwards Selection, Backwards Selection, Lasso regression, and Ridge regression. Support your answer.

   (The data can be downloaded from https://kdd.ics.uci.edu/databases/tic/tic.html.)

2) We have seen that as the number of features used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.

   Generate a data set with p = 20 features, n = 1, 000 observations, and an associated quantitative response vector generated according to the model

   $Y = X\beta + \varepsilon$

   where β has some elements that are exactly equal to zero. Split your data set into a training set containing 800 observations and a test set containing 200 observations.

   Perform best subset selection or forward stepwise selection -- on the training set, and plot the training set MSE associated with the best model of each size. Plot the test set MSE associated with the best model of each size.

   For which model size does the test set MSE take on its minimum value? Comment on your results. How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the coefficient values.

   Hint: you can simulate data in R for the 20 features. Uniform distribution (>runif) or Normal distribution (>rnorm). Be sure to set the seed!

3) Consider the iris data. Divide the data into test and training.

   data(iris)
   > ?iris

   a) Perform k-nearest neighbor on the data for a range of k values. Plot the error rate as a function of k. Report the confusion matrix for the optimal model. Comment on the ability of kNN to discriminate the various species.
      Hint: "confusionMatrix" function can be used from the "caret" package.

b) Perform k-nearest neighbor on the first two principal components. Plot the error rate as a function of k. Report the confusion matrix. Plot the scores for the first two principal components and color the samples by class (Species). How does error rate compare to Part (A)?