Statistical Data Mining I
Homework 1

Due: Friday September 14th (11:59 pm)
40 points

Directions: Submit all source codes with write up. Please see UB Learns homework guidelines

1) (10 points) Consider the Student Performance Data Set on the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/student+performance). Suppose that you are getting this data in order to build a predictive model for First Period Grades. Using the full dataset, investigate the data using exploratory data analysis such as scatterplots, and other tools we have discussed in class. Pre-process this data and justify your choices (elimination of outliers, elimination of variables, variable transformations, etc.) in your write up. Submit the cleaned dataset as an *.RData file.

2) (10 points) Perform a multiple regression on the dataset you pre-processed in question one. The response are the first period grades. Use the lm() function in R.
   a) Which predictors appear to have a significant relationship to the response.
   b) What suggestions would you make to a first-year student trying to achieve good grades.
   c) Use the * and : symbols to fit models with interactions. Are there any interactions that are significant?

3) (10 points) ISL textbook exercise 2.10 modified: This exercise concerns the boston housing data in the MASS library (>library(MASS) >data(Boston)).
   a) Make pairwise scatterplots of the predictors, and describe your findings.
   b) Are any of the predictors associated with per capita crime rate?
   c) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.
   d) In this data set, how many of the suburbs average more than seen rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

4) (10 points) ESL textbook exercise 2.8 modified: Compare the classification performance of linear regression and k-nearest neighbor classification on the *zipcode* data. In particular, consider only the 2's and 3's for this problem, and k = 1,3,5,7,9,11, 13,15. Show both the training and the test error for each choice of k. The *zipcode* data is available in the ElemStatLearn package. Note that you do not have to divide the data into test and training as it is done for you.
   > ls("package:ElemStatLearn")
   >?zip.test
   >?zip.train