STA 545 Statistical Data Mining I, Fall 2020

Homework 1, due: Wednesday 9/9/2020 (1PM)

1. (40 points) Read the materials in the R/Rstudio/Rmarkdown folder in UBlearns.

   (a) Install R and RStudio on your laptop.

   (b) In RStudio, install the packages ISLR and rmarkdown.

   (c) Download the R markdown template file in the Materials about R/Rstudio/Rmarkdown folder in UBlearns.

   (d) Change the template to show the information about the wage dataset used in the ISLR textbook. In the ISLR package, the name of this dataset is Wage.

   (e) Run the R markdown file to generate a pdf file.

   Please submit a printed copy of the pdf file generated from R markdown. In addition, please write the solutions of the following homework questions in R markdown.

2. (30 points) Explain whether each scenario is a classification or regression problem. In addition, please provide the sample size n and the number of independent variables p for each scenario.

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   (b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

(c) We are interest in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

3. (30 points) You will now think of some real-life applications for statistical learning.

   (a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Please do not use the examples shown in the slides.

   (b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Please do not use the examples shown in the slides.

   (c) Describe three real-life applications in which cluster analysis might be useful. Please do not use the examples shown in the slides.