Homework 1

Due April 10, 2020 by 11:59pm

Instructions: **all code exercises must be completed using Python.** No upload to Canvas is required for the reading and studying exercises. Upload your answers to the exercises to Canvas. Submit the answers to the questions (including the relevant output of the code) in a PDF file and your code in a (single) separate file. Be sure to comment your code to indicate which lines of your code correspond to which question part.

1. Read Chapter 2 in *An Introduction to Statistical Learning*.

2. Study Computer Lab. 1 in *canvas.uw.edu/courses/1371621/pages/course-materials* .

3. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

   (a) You want to predict whether a particular customer is going to click on an online advertisement or not. You have information on whether or not they clicked on 200 other ads, in addition to whether the ad was in the same category, whether the ad was shown during regular working hours, whether the ad was shown on a weekend, and the percent of all customers who had previously clicked on the ad.

   (b) Suppose it is the end of the quarter and you wish to predict your score on the final exam. You have data from 20 classes you have previously taken, consisting of your final exam scores, your average scores on the midterms (i.e., one average midterm score per class), your average homework scores (i.e., one average homework score per class), and whether the final exam was take-home or not.

   (c) You work for an ice cream shop and are in charge of determining what factors affect how much ice cream is sold each day. For 600 days you have information on how much ice cream the shop sold, in addition to whether the day was sunny or not, what the temperature was, whether school is in session or not, whether your most popular flavor was available that day, and whether you had recently run any advertisements.

4. In this problem you will brainstorm real-life applications for statistical learning. Your answers aren't allowed to be the same as any of the examples in the other homework problems.

   (a) Describe three real-life applications in which classification might be useful, **one from political science, one from sports, and one from an area of your choice**. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(b) Describe three real-life applications in which regression might be useful, **one from agriculture, one from business, and one from an area of your choice**. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(c) Describe three real-life applications in which cluster analysis might be useful, **one from education, one from meteorology, and one from an area of your choice**. Be sure to describe why it would be useful.

5. This exercise involves the `Auto` data set found here: `http://www-bcf.usc.edu/~gareth/ISL/Auto.data`. Make sure that the missing values have been removed from the data.

   (a) Which of the predictors are quantitative, and which are qualitative?

   (b) What is the range of each quantitative predictor?

   (c) What is the mean and standard deviation of each quantitative predictor?

   (d) Now remove the last 50 observations. What is the range, mean, and standard deviation of each quantitative predictor in the subset of the data that remains?

   (e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

   (f) Suppose that we wish to predict gas mileage (`mpg`) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting `mpg`? Justify your answer.