CS 422: Data Mining

Department of Computer Science
Illinois Institute of Technology
Vijay K. Gurbani, Ph.D.

Spring 2019: Homework 1 (10 points)

Due date: Sun, Feb 17, 2019 11:59:59 PM Chicago Time

Please read all of the parts of the homework carefully before attempting any question. If you detect any ambiguities in the instructions, please let me know right away instead of waiting until after the homework has been graded.

This is the final version of the homework assignment.

1 Exercises (total points 3)

1.1 Tan, Chapter 1 (1 point divided evenly among the questions)

Besides the lecture, make sure you read Chapter 1. After doing so, answer the following questions at the end of the chapter: 1, 3.

1.2 Tan, Chapter 2 (1 point divided evenly among the questions)

Besides the lecture, make sure you read Chapter 2, sections 2.1 – 2.3. After doing so, answer the following questions at the end of the chapter: 2, 3, 7, 12.

1.3 ISLR 7e (Gareth James, et al.) (1 point divided evenly among the questions)

Section 3.7 (Exercises), page 120: Exercises 1, 3, 4-a.

2 Programming Problems (total points 7, divided as indicated below)

2.1 Problem 1 (3 points)

This exercise relates to the College data set, which can be found in the file College.csv. It contains a number of variables for 777 different universities and colleges in the US. The variables are

- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled
- Top10perc : New students from top 10 % of high school class
- Top25perc : New students from top 25 % of high school class
- F.Undergrad : Number of full-time undergraduates

- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

Before reading the data into R , you should view in a text editor or imported in a spreadsheet.

(a) (0.1 points) Use the read.csv() function to read the data into an R dataframe, called 'college.df' . Make sure that you have the directory set to the correct location for the data. Print the top 6 observations in this dataframe.

(You can use the setwd() function to set the current working directory. For example, assume that College.csv is located in /home/vkg/CS422/Homework-1. Then, setwd("/home/vkg/CS422/Homework-1") will set the current working directory appropriately. Once you do this, you can simply invoke read.csv("College.csv", …) instead of prefxing the path name.)

(b) (0.1 points) Count and print the number of private colleges and the number of public colleges. You may use the table() function to build a contingency table of counts at each combination of factor levels. Recall that a factor in R is a constrained type that can take only a certain predefined values (an enum in Java or C).

(c) (0.7 points) Create two new data frames: one contains all public colleges and the other contains all private colleges. Plot the histogram of the PhD holders in private colleges and public colleges. (You will produce 2 histograms). For each plot, overlay the plot with a density curve. Make sure you use color and labeling to make your graph look attractive. Provide some comments on the histograms (i.e., are private colleges top-heavy with respect to PhD faculty? Are public colleges top-heavy?)

(Hint: Use dplyr::filter() to create the data frames. To create a histogram, use the hist() function. To overlay a density curve on top of a histogram, use lines(density(…)). The lines() command plots the result on the existing plot, i.e., it will overlay the line on top of the existing plot.

(d) (0.6 points) Create a new data frame that contains the data in the college.df dataframe, but sorted on the attribute "Grad.Rate". Then, print the top 5 colleges that have the minimum graduation rates (print the name and graduation rate), and the top 5 colleges that have the maximum graduation rate (again, print the name and graduation rate).

(Hint: To sort the data frame on attributes, use the dplyr::arrange() method. To print the two attributes --- name and graduation rate --- use the dplyr::select() method.)

(e) (1.5 points divided evenly)

i. Use the summary() function to produce a numerical summary of the variables in the data set.

ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10] .

iii. Use the boxplot() function to produce side-by-side boxplots that help answers the following question: Which alumni donate more to their colleges --- those who go to public schools or those who go to private schools?

To do this, you will use the boxplot() function, but you will group the attribute you are interested in by a control group. The control group here is the attribute Private (which is 'Yes' or 'No'), and the attribute of interest here is perc.alumni. The format is boxplot(x, data=), where x is a formula and data= denotes the data frame providing the data. An example of a formula is y~group where a separate boxplot for numeric variable y is generated for each value of group. Label the X- and Y-axes appropriately and provide a main= parameter to the boxplot() command for a graph title. You should see two boxplots if all works.

As an added resource, take a look at https://www.statmethods.net/graphs/boxplot.html

iv Use the boxplot() function to produce side-by-side boxplots that help answers the following question: Which colleges --- public or private --- employ more Ph.D.'s?

v. Create a new qualitative variable, called Elite by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classes exceeds 50 %.

```
> Elite <- rep("No", nrow(college))
> Elite[college$Top10perc > 50] <- "Yes"
> Elite <- as.factor(Elite)
> college <- data.frame(college, Elite)
```

Use the summary() function to see how many elite universities there are.

vi. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

vii. Continue exploring the data, and provide a brief summary of what you discover. (Yes, I know this is an open ended question, but discovery is an important part of data mining.)

For Parts 2.2 and 2.3 below, install the ISLR package in your R environment. Then load it and use the Auto dataset from the package.

2.2 Problem 2: Linear Regression I (2 points)

ISLR 7e (Gareth James, et al.), Applied questions (page 121), question 8.

2.3 Problem 3: Linear Regression II (2 points)

ISLR 7e (Gareth James, et al.), Applied questions (page 122), question 9(a) - 9(d).