CS 422: Data Mining

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

Spring 2018: Homework 1 (10 points)

Due date: Wed, Feb 07, 2018 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.

Additional parts to this homework will be added as new material is covered.

1 **Recitation problems** (total points TBD)

1.1 Tan, Chapter 1

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

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.

2 **Practicum problems** (total points TBD)

2.1 Problem 1

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 , it can be viewed in Excel or a text editor.

(a) Use the read.csv() function to read the data into R . Call the loaded data ‘college’ . Make sure that you have the directory set to the correct location for the data. 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) Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don’t really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```
      > rownames(college) <- college[,1]
      > fix(college)
```

You should see that there is now a row.names column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try:

```
      > college <- college [ , -1]
      > fix(college)
```

Now you should see that the first data column is Private . Note that another column labeled row.names now appears before the Private column. However, this is not a data column but rather the name that R is giving to each row.

(c)

i. Use the summary() function to produce a numerical summaryof 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 plot() function to produce side-by-side boxplots of Outstate versus Private .

iv. 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. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite .

v. 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.

vi. Continue exploring the data, and provide a brief summary of what you discover.