CS 422-04: Data Mining (CRN: 19881)

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

Fall 2017: Homework 3 (10 points)

Due date: Wednesday, Oct 25, 2017 11:59:59 PM Chicago Time

1 Exercises (2.5 points divided evenly among the questions)

1.1 Chapter 5

Exercises 7(a), 7(b), 8.

1.2 Chapter 6

Exercise 2, 6.

1.3 Chapter 8

Exercise 2, 11.

2 Practicum problems

2.1 Problem 1: Naive Bayes classification (2.5 points)

This problem uses the ILPD database that you used for Homework 2. As before, set the seed to 100 (i.e., set.seed(100)) first. Then, divide the dataset into two parts: train and test with a random 60/40 split (i.e., 60% of the data is used for training and 40% for testing).

Take a look at 2.1(f) from Homework 2, which asked you to choose the best model. Using the same predictor variables you used for the best model in Homework 2, create a Naive Bayes classification model.

Then, test your model and report the (a) ROC curve using the ROCR package, and (b) the AUC. Does the Naive Bayes classifier perform better or worse than your best tree-based classifier from Homework 2?

2.2 Problem 2: Association rules (2.5 points)

In the file groceries.csv you will find transactions related to market-basket analysis. There are 9,835 rows (transactions) and 169 columns (items). Read in this file using read.transaction() and run the **Apriori** (or **Ecalat***) association rule algorithm on it to answer the following questions:

(a) Which item is the most frequently bought and what is its frequency?
(b) Which item is the least frequently bought and what is its frequency?
(c) At what level of support are the rules generated?
(d) What are the top 5 rules, sorted by *support*?
(e) What are the top 5 rules, sorted by *confidence*?
(f) What are the top 5 rules, sorted by *lift*?

(g) What are the bottom 5 rules, sorted by *support*?
(h) What are the bottom 5 rules, sorted by *confidence*?
(i) What are the bottom 5 rules, sorted by *lift*?

* If you use Eclat, make sure you invoke ruleInduction() as follows to get the rules:

```
> trans <- read.transactions(…)
> ecl <- eclat(trans, …)
> eclat.rules <- ruleInduction(ecl)
```

2.3 Problem 3: Clustering (2.5 points)

HARTIGAN is a dataset directory that contains test data for clustering algorithms. The data files are all simple text files, and the format of the data files is explained on the web page at https://people.sc.fsu.edu/~jburkardt/datasets/hartigan/hartigan.html

Perform k-means clustering on file19.txt on the above web page. This file contains a multivariate mammals dataset; there are 9 columns and 66 rows.

(a) Data cleanup

(i) Think of what attributes, if any, you may want to omit from the dataset when you do the clustering. Indicate all of the attributes you removed before doing the clustering.

(iii) Does the data need to be standardized?

(iii) You will have to clean the data to remove multiple spaces and make the comma character the delimiter. Please make sure you include your cleaned dataset in the archive file you upload.

(b) Clustering

(i) Determine how many clusters are needed by running the WSS or Silhouette graph. Plot the graph using fviz_nbclust().

(ii) Once you have determined the number of clusters, run k-means clustering on the dataset to create that many clusters. Plot the clusters using fviz_cluster().

(iii) How many observations are in each cluster?

(iv) What is the total SSE of the clusters?

(v) What is the SSE of each cluster?

(vi) Perform an analysis of each cluster to determine how the mammals are grouped in each cluster, and whether that makes sense? For example, to get the indices of all animals in cluster 1, you would execute:

```
> which(k$cluster == 1)
```

assuming *k* is the variable that holds the output of the kmeans() function call.

For the above question, put your answer in a text file and include the text file in the uploaded archive.