QMST 3339 Homework 3

Tahir Ekin[1]

[1]McCoy College of Business, Texas State University

Spring 2018

Guidelines

1. A hard copy (print-out) of your report that summarizes the results and includes the supporting graphs is due to the beginning of the lecture on 3/19/2018, Monday. Please also upload your document and your R code to TRACS/Assignments.

2. Do not include data in your report.

3. This is a group-based homework. However, all team members are expected to understand the assignment, relevant script and the submission thoroughly.

4. This is graded out of 5 points, and accounts for 5% of your overall grade. See the syllabus for more details.

5. Submission of any work, for which unauthorized help has been received, is termed as academic dishonesty and will be grounds for a failing grade in the course. See "Academic Honesty" section of the syllabus for more details.

1 Association

1.1 Association Concepts: 2 pts

A sample of transaction data has 7 transactions

- Bread, Milk, Chips, Mustard
- Beer, Diaper, Bread, Eggs
- Beer, Coke, Diaper, Milk
- Beer, Bread, Diaper, Milk, Chips
- Coke, Bread, Diaper, Milk
- Beer, Bread, Diaper, Milk, Mustard
- Coke, Bread, Diaper, Milk

The minimum support threshold for items is set as 0.6. Use the apriori algorithm (by hand) to find the frequent items and a list of rules. Show all steps and compute the support, confidence and lift values for the constructed rules.

1.2 Association Apriori Algorithm : 3 pts

Online radio (e.g., Lastfm, Pandora) keeps track of everything you play. It uses this information for recommending music you are likely to enjoy and supports focused marketing that sends you advertisements. You are given data from such a site. It consists of the information of every artist that the user has downloaded and the demographic data of the user. Your objective is to build a recommender system. That is, a system that recommends new music to users in this community. The file is a rather large data set with close to 300,000 records of song (artist) selections made by 15,000 users. Each row of the data set contains the name of the artist to whom the user has listened.

1.2.1 Data Pre-Processing (0.5 pts in total, 0.1 pts each)

Download the lastfm.csv from the TRACS site. Use the read.table function to read the dataset into R.

1. Explore the dataset. Write the number of records and variables.

2. Write the number of unique countries, users and artists that are represented in this dataset. You can use dim(table()) or levels() functions in R.

3. Select records number 200 and 201. Write the name of the bands, the name of the artists and the name of the country.

4. We need to organize this dataset into an incidence matrix by user. User is coded as an integer. Write the R function to convert and replace "user" into a categorical (factor) variable, permanently.

5. Before moving to the data analysis, ensure you have successfully completed the transformations below. You do not need to report output for this step.

You need to manipulate the data into an itemMatrix before using arules package in R. Using the split function, split the data in the vector x named *lastfm* into groups defined by vector f (*users*). In supermarket terms, think of users as shoppers and artists as items bought.

playlist <- split(x=lastfm[,"artist"],f=user)

Then, we remove artist duplicates since an artist may be chosen by the same user more than once.

playlist <- lapply(playlist,unique)

For instance, the first two listeners (numbered 1 and 3) listen to the following bands:

playlist[1:2];

Then, we convert this into a list of "transactions" which is the data class defined in arules.

playlist <- as(playlist,"transactions")

Now, the data is transformed into an incidence matrix as given in *playlist*, where each listener represents a row, with 0 and 1s across the columns indicating whether or not he or she has downloaded a certain artist.

1.2.2 Data Analysis (2.5 pts in total, 0.3 pts each unless otherwise specified)

1. Use the itemFrequency() function to list the support for the bands. Write the top five most frequently chosen bands.

2. Present a horizontal barplot for the bands that have at least a support of 0.10 (10% of the users) or greater.

3. Report the lower, mid and upper quartiles for the distribution of item relative frequency (support).

4. Find the rules that have support which is greater than the upper quartile value of item relative frequency and confidence of at least 0.40. You do not need to present them. What is the number of such rules?

5. Inspect the rules and show the list of 5 rules with the highest lift

6. Among the 5 rules you found in the previous question, select the one with the highest confidence. Explicitly write down its support, confidence and lift values. Explain their meaning in this context. 1 pt