Association Rule Mining:

Market Basket Analysis

QSTP: Data Analysis with R

Associative Rule Mining

In a given transaction with multiple items, it tries to find the rules that govern how or why such items are often bought together.

For example, butter and jam are often bought together because a lot of people like to make sandwiches.

Apriori Algorithm

Apriori algorithm, a classic algorithm, is useful in mining frequent itemsets and relevant association rules. Usually, you operate this algorithm on a database containing a large number of transactions.

Example is the items customers buy at a supermarket. It helps the customers buy their items with ease, and enhances the sales performance of the departmental store.

- Support
- Confidence
- Lift

SUPPORT

It gives the occurence of item or group of items as compared to all transactions.

For example: If in data set, there are total 8 transactions. Two items X and Y are together for 4 transactions.

Support(X=>Y)= 4\8

Support(X=>Y)

= Frequency(X,Y)/N

Where N is the total no of transactions.

CONFIDENCE

It gives how likely an item Y is purchased when item X is purchased (X=>Y).

For example: If in data set, there are total 8 transactions. Two items X and Y are together for 4 transactions and X is brought for 6 transactions.

Confidence(X=>Y)= 4/6

Confidence (X=>Y)

= Support(X=>Y) /Support(X)

LIFT

It gives how likely an item Y is purchased when item X is purchased (X=>Y) taking into consideration occurence of Y as well.

For example: If in data set, there are total 8 Transactions. Items X and Y are together for 4 , X for 6 and Y for 6 transactions.

Lift(X=>Y)= (4/8)/((6/8)*(6/8))

=8/9

Lift(X=>Y)

= Support(X=>Y)/ (Support(X)*Support(Y))

Implementation in R

Packages used are:

- ➔ **arules**
- ➔ **arulesViz**

Install these packages in RStudio.

Execute these statements then run other functions.

```
library(arules)

library(arulesViz)
```

apriori function:

apriori(data, parameter)

data: data in form of transaction objects
parameter: parameter can be given us list of confidence and support values.

Note:

Please note that you cannot load data frames in place of data in apriori function directly created by csv files. You need to make transaction object for that. That function we will discuss later on.

Example:

data("Adult")

r1 <-
apriori(Adult,parameter=list(support=0.5, confidence=0.5))
You can also specify minlen or maxlen in parameter ,that is minimum or maximum no of items that one group can have.

Note:

There are some inbuilt data sets in R which you can run directly like shown. No need to create objects separately.

inspect function:

Summarises all statistics.

Example:
inspect(head(r1))

Summarises statistics of first 6 transactions.

Ālternative to inspect:

We can also convert rules to a dataframe and then use View()

Example: df_r1<-as(r1,"data.frame")
View(df_r1)

sort function:

There are various parameters to sort() function, the important ones are:
what is to be sorted
decreasing: True sorts in reverse order
by: custom value by which you want to sort

Example:

sort(r1, by="lift", decreasing=”True”)

Sorts r1 on basis of lift values of every possible group created by apriori function in decreasing order.

plot function:

plot(r1)
It just gives plot between confidence, support and lift.

plot(r1,method=”grouped”)
It plots according to the group of items.

plot function:

plot(r1[1:20], method=”graph”, control = list(type=items))

The size of graph nodes is based on support levels and the color on lift ratios. The incoming lines show the LHS and the RHS is represented by names of items.

Assignment 2

Use inbuilt Groceries data set (like the Adult shown in slides) for this assignment.

Run apriori function 3 times by creating variables rule1, rule2 ,rule3 (you are free to choose any variable names)

- ➔ **rule1: support=0.002,confidence=0.5**
- ➔ **rule2: support=0.002,confidence=0.5,minlen=5**
- ➔ **rule3: support=0.007,confidence=0.6**

Assignment 2 (Contd..)

For all three inspect its first 4 and last 4 elements.

Sort them on basis of lift values in increasing order and inspect first 4 and last 4 elements.

Plot all three by grouped and graph method.

Analyze the plots.

Can you conclude something significant from plot of rule3 ?