retail

Joyce Woznica

October 19,2019

This is a tutorial uses R package “arules” to analyze the market basket data.

# Load the libraries

# install.packages(arules)  
# install.packages(arulesViz)  
library(arules)

## Loading required package: Matrix

##   
## Attaching package: 'arules'

## The following objects are masked from 'package:base':  
##   
## abbreviate, write

library(arulesViz)

## Loading required package: grid

## Registered S3 method overwritten by 'seriation':  
## method from   
## reorder.hclust gclus

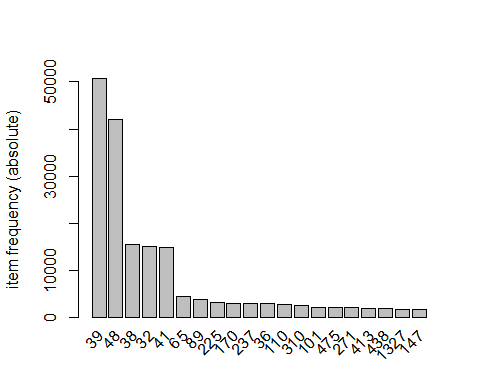
# Load the dataset

retail <- read.transactions("C:/Users/Joyce/Desktop/Syracuse/IST707/Classwork/week3/retail.csv", format="basket", sep=",")

# We can explore the data before we make any rules. The following script is going to create an item frequency plot for the top 20 items.

# “type” indicates whether item frequencies should be displayed relative or absolute

itemFrequencyPlot(retail,topN=20,type="absolute")



# Mine rules with the Association Rule algorithm.

# It is required to set the minimum support and confidence values.

rules <- apriori(retail, parameter = list(supp = 0.001, conf = 0.8,maxlen=3))

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport maxtime support minlen  
## 0.8 0.1 1 none FALSE TRUE 5 0.001 1  
## maxlen target ext  
## 3 rules FALSE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 88   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[16470 item(s), 88162 transaction(s)] done [0.53s].  
## sorting and recoding items ... [2117 item(s)] done [0.01s].  
## creating transaction tree ... done [0.05s].  
## checking subsets of size 1 2 3

## Warning in apriori(retail, parameter = list(supp = 0.001, conf = 0.8,  
## maxlen = 3)): Mining stopped (maxlen reached). Only patterns up to a length  
## of 3 returned!

## done [0.07s].  
## writing ... [355 rule(s)] done [0.01s].  
## creating S4 object ... done [0.01s].

# Show the top 5 rules, rounding with 2 digits

# From the top 5 rules, we could notice that the rule {bottled beer, soups} => {whole milk} is strong, indicating bottled beer and soups are frequently bought together with whole milk. So we would suggest to place these items close to each other.

options(digits=2)  
inspect(rules[1:5])

## lhs rhs support confidence lift count  
## [1] {3854} => {38} 0.0011 0.91 5.2 94   
## [2] {1045} => {32} 0.0011 0.91 5.3 97   
## [3] {4030} => {48} 0.0010 0.83 1.7 90   
## [4] {1473} => {39} 0.0012 0.80 1.4 108   
## [5] {1727} => {38} 0.0018 0.93 5.3 162

# Get summary info about all rules

summary(rules)

## set of 355 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 2 3   
## 42 313   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.00 3.00 3.00 2.88 3.00 3.00   
##   
## summary of quality measures:  
## support confidence lift count   
## Min. :0.001 Min. :0.80 Min. : 1 Min. : 89   
## 1st Qu.:0.001 1st Qu.:0.82 1st Qu.: 1 1st Qu.: 106   
## Median :0.002 Median :0.84 Median : 2 Median : 144   
## Mean :0.003 Mean :0.88 Mean : 12 Mean : 275   
## 3rd Qu.:0.003 3rd Qu.:0.96 3rd Qu.: 5 3rd Qu.: 241   
## Max. :0.084 Max. :1.00 Max. :296 Max. :7366   
##   
## mining info:  
## data ntransactions support confidence  
## retail 88162 0.001 0.8

# Sort rules so that we can view the most relevant rules first. For example, sort rules with “confidence”:

rules<-sort(rules, by="confidence", decreasing=TRUE)  
subset.matrix <- is.subset(rules, rules)  
subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA

## Warning in `[<-`(`\*tmp\*`, as.vector(i), value = NA): x[.] <- val: x is  
## "ngTMatrix", val not in {TRUE, FALSE} is coerced; NA |--> TRUE.

redundant <- colSums(subset.matrix, na.rm=T) >= 1  
rules.pruned <- rules[!redundant]  
rules<-rules.pruned

# Visualize the rules

#plot(rules,method="graph",interactive=TRUE,shading=NA)