groceries

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)

## Warning: package 'arules' was built under R version 3.2.4

## Loading required package: Matrix

##   
## Attaching package: 'arules'

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

library(arulesViz)

## Warning: package 'arulesViz' was built under R version 3.2.3

## Loading required package: grid

## Warning: replacing previous import by 'utils::head' when loading  
## 'arulesViz'

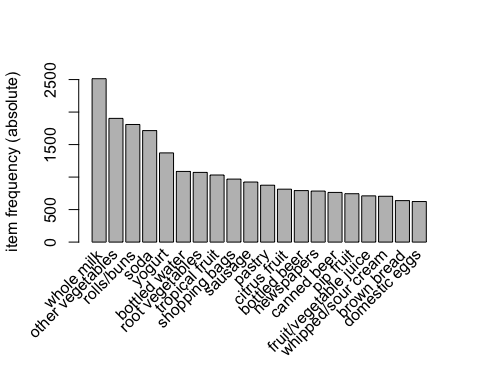
# Load the dataset

groceries <- read.transactions("/Users/byu/Desktop/Data/groceries.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(groceries,topN=20,type="absolute")



# Mine rules with the Association Rule algorithm.

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

rules <- apriori(groceries, parameter = list(supp = 0.001, conf = 0.8))

## Apriori  
##   
## Parameter specification:  
## confidence minval smax arem aval originalSupport support minlen maxlen  
## 0.8 0.1 1 none FALSE TRUE 0.001 1 10  
## target ext  
## rules FALSE  
##   
## Algorithmic control:  
## filter tree heap memopt load sort verbose  
## 0.1 TRUE TRUE FALSE TRUE 2 TRUE  
##   
## Absolute minimum support count: 9   
##   
## set item appearances ...[0 item(s)] done [0.00s].  
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.00s].  
## sorting and recoding items ... [157 item(s)] done [0.00s].  
## creating transaction tree ... done [0.00s].  
## checking subsets of size 1 2 3 4 5 6 done [0.02s].  
## writing ... [410 rule(s)] done [0.00s].  
## creating S4 object ... done [0.00s].

# 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  
## 1 {liquor,red/blush wine} => {bottled beer} 0.0019 0.90 11.2  
## 2 {cereals,curd} => {whole milk} 0.0010 0.91 3.6  
## 3 {cereals,yogurt} => {whole milk} 0.0017 0.81 3.2  
## 4 {butter,jam} => {whole milk} 0.0010 0.83 3.3  
## 5 {bottled beer,soups} => {whole milk} 0.0011 0.92 3.6

# Get summary info about all rules

summary(rules)

## set of 410 rules  
##   
## rule length distribution (lhs + rhs):sizes  
## 3 4 5 6   
## 29 229 140 12   
##   
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.0 4.0 4.0 4.3 5.0 6.0   
##   
## summary of quality measures:  
## support confidence lift   
## Min. :0.00102 Min. :0.80 Min. : 3.1   
## 1st Qu.:0.00102 1st Qu.:0.83 1st Qu.: 3.3   
## Median :0.00122 Median :0.85 Median : 3.6   
## Mean :0.00125 Mean :0.87 Mean : 4.0   
## 3rd Qu.:0.00132 3rd Qu.:0.91 3rd Qu.: 4.3   
## Max. :0.00315 Max. :1.00 Max. :11.2   
##   
## mining info:  
## data ntransactions support confidence  
## groceries 9835 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)

# If we want to target items to generate rules. For example, the frequently bought items with “whole milk”:

# "minlen" is to avoid empty left hand side items.

rules<-apriori(data=groceries, parameter=list(supp=0.001,conf = 0.08, minlen=2),   
 appearance = list(default="lhs",rhs="whole milk"),  
 control = list(verbose=F))  
rules<-sort(rules, decreasing=TRUE,by='confidence')  
inspect(rules[1:5])

## lhs rhs support confidence lift  
## 1 {rice,   
## sugar} => {whole milk} 0.0012 1 3.9  
## 2 {canned fish,   
## hygiene articles} => {whole milk} 0.0011 1 3.9  
## 3 {butter,   
## rice,   
## root vegetables} => {whole milk} 0.0010 1 3.9  
## 4 {flour,   
## root vegetables,   
## whipped/sour cream} => {whole milk} 0.0017 1 3.9  
## 5 {butter,   
## domestic eggs,   
## soft cheese} => {whole milk} 0.0010 1 3.9

# Likewise, we could set "whole milk" on the left hand side as well.

rules<-apriori(data=groceries, parameter=list(supp=0.001,conf = 0.15,minlen=2),   
 appearance = list(default="rhs",lhs="whole milk"),  
 control = list(verbose=F))  
rules<-sort(rules, decreasing=TRUE,by="confidence")  
inspect(rules[1:5])

## lhs rhs support confidence lift  
## 6 {whole milk} => {other vegetables} 0.075 0.29 1.5   
## 5 {whole milk} => {rolls/buns} 0.057 0.22 1.2   
## 4 {whole milk} => {yogurt} 0.056 0.22 1.6   
## 2 {whole milk} => {root vegetables} 0.049 0.19 1.8   
## 1 {whole milk} => {tropical fruit} 0.042 0.17 1.6

# Visualize the rules

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