STA380HW2_Q3

Andrea You

08/18/2017

R Markdown

Question 3 Grocery (Association Rule Mining)

Overview: In this question, the main goal is to find interesting association rules for shopping baskets and the key is to pick our own thresholds for lift and confidence. Here, we define "interesting" rules as rules that could be used to take some pratical actions especially in business settings.

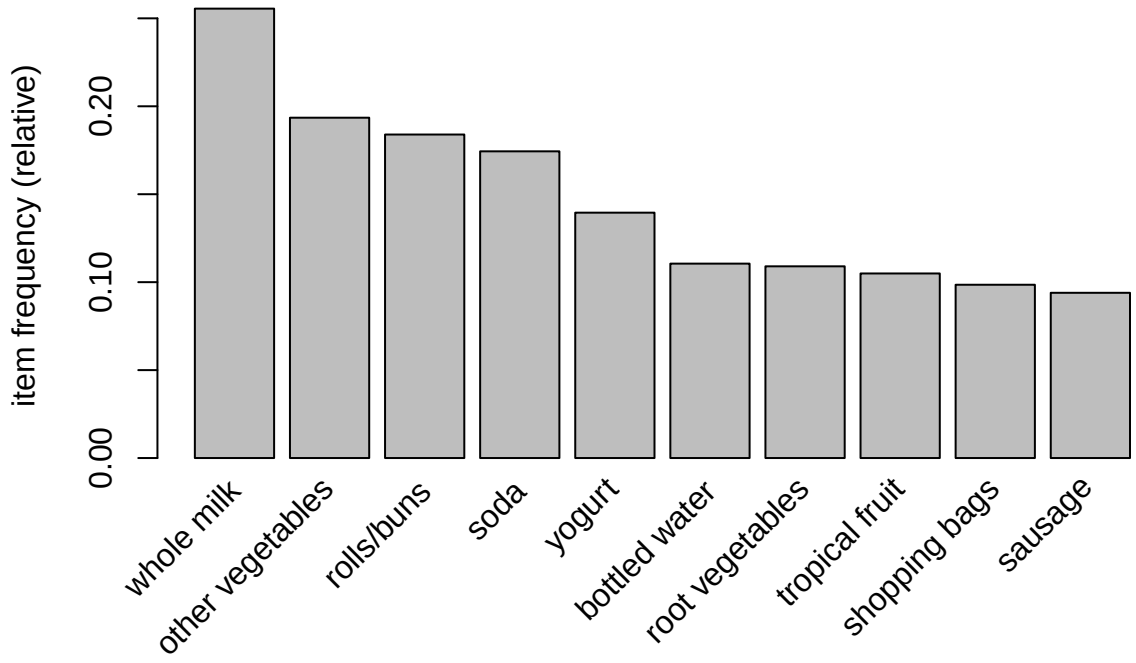
Aftering reading in the text file and basic explorating, we first selected out rules that meet min required support and confidence threholds, followed by sorting according to confidence. And then we approached association rules with multiple ways, all with visulization, interpretation and discussion.Finally, we ended up the discussion with potential application and commercial suggestions.

```
# Load the libraries
library(arules)
library(arulesViz)
```

```
# Read in the text file as a format accessible for "arules" package
library(arules)
grocery <- read.transactions('https://raw.githubusercontent.com/jgscott/STA380/master/data/groceries.tx
summary(grocery)
```

```
## transactions as itemMatrix in sparse format with
##  9835 rows (elements/itemsets/transactions) and
##  169 columns (items) and a density of 0.02609146
##
## most frequent items:
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt          (Other)
##             1372            34055
##
## element (itemset/transaction) length distribution:
## sizes
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
## 2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55
##   16   17   18   19   20   21   22   23   24   26   27   28   29   32
##   46   29   14   14    9   11    4    6    1    1    1    1    3    1
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   4.409   6.000  32.000
##
## includes extended item information - examples:
##                  labels
## 1 Instant food products
## 2              UHT-milk
## 3      abrasive cleaner
```

```
# Plot top 10 frequent appearing items in grocery
itemFrequencyPlot(grocery,topN=10)
```



###General settings

```
# To get as many as rules from the start, we set the min support to 0.001
# To get high confidence rules, we set the min confidence to 0.5
groceryrules <- apriori(grocery, parameter = list(support = 0.001, confidence = 0.5))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.5    0.1    1 none FALSE            TRUE       5   0.001      1
##  maxlen target   ext
##      10  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 [0.01s].
## writing ... [5668 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```

```
# Show the top 10 rules
inspect(groceryrules[1:10])
```

```
##      lhs                    rhs                support     confidence
```

```
## [1]  {honey}             => {whole milk}       0.001118454 0.7333333
## [2]  {tidbits}           => {rolls/buns}       0.001220132 0.5217391
## [3]  {cocoa drinks}      => {whole milk}       0.001321810 0.5909091
## [4]  {pudding powder}    => {whole milk}       0.001321810 0.5652174
## [5]  {cooking chocolate} => {whole milk}       0.001321810 0.5200000
## [6]  {cereals}           => {whole milk}       0.003660397 0.6428571
## [7]  {jam}               => {whole milk}       0.002948653 0.5471698
## [8]  {specialty cheese}  => {other vegetables} 0.004270463 0.5000000
## [9]  {rice}              => {other vegetables} 0.003965430 0.5200000
## [10] {rice}              => {whole milk}       0.004677173 0.6133333
##      lift
## [1]  2.870009
## [2]  2.836542
## [3]  2.312611
## [4]  2.212062
## [5]  2.035097
## [6]  2.515917
## [7]  2.141431
## [8]  2.584078
## [9]  2.687441
## [10] 2.400371
```

```
# Get summary of groceryrules we have got from above
summary(groceryrules)
```

```
## set of 5668 rules
##
## rule length distribution (lhs + rhs):sizes
##    2    3    4    5    6
##   11 1461 3211  939   46
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    3.00    4.00    3.92    4.00    6.00
##
## summary of quality measures:
##     support           confidence          lift
##  Min.   :0.001017   Min.   :0.5000   Min.   : 1.957
##  1st Qu.:0.001118   1st Qu.:0.5455   1st Qu.: 2.464
##  Median :0.001322   Median :0.6000   Median : 2.899
##  Mean   :0.001668   Mean   :0.6250   Mean   : 3.262
##  3rd Qu.:0.001729   3rd Qu.:0.6842   3rd Qu.: 3.691
##  Max.   :0.022267   Max.   :1.0000   Max.   :18.996
##
## mining info:
##     data ntransactions support confidence
##  grocery          9835   0.001        0.5
```

As the summary shows, there are in total 5668 rules generated. And most frequent rules appearing are 4-item ones.

```
# Sort groceryrules by confidence
groceryrules<-sort(groceryrules, by="confidence")
inspect(groceryrules[1:5])
```

```
##     lhs                    rhs              support confidence     lift
## [1] {rice,
```

```
##      sugar}              => {whole milk} 0.001220132          1 3.913649
## [2] {canned fish,
##      hygiene articles}   => {whole milk} 0.001118454          1 3.913649
## [3] {butter,
##      rice,
##      root vegetables}    => {whole milk} 0.001016777          1 3.913649
## [4] {flour,
##      root vegetables,
##      whipped/sour cream} => {whole milk} 0.001728521          1 3.913649
## [5] {butter,
##      domestic eggs,
##      soft cheese}        => {whole milk} 0.001016777          1 3.913649
```

```
summary(groceryrules)
```

```
## set of 5668 rules
##
## rule length distribution (lhs + rhs):sizes
##    2    3    4    5    6
##   11 1461 3211  939   46
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    3.00    4.00    3.92    4.00    6.00
##
## summary of quality measures:
##     support           confidence          lift
##  Min.   :0.001017   Min.   :0.5000   Min.   : 1.957
##  1st Qu.:0.001118   1st Qu.:0.5455   1st Qu.: 2.464
##  Median :0.001322   Median :0.6000   Median : 2.899
##  Mean   :0.001668   Mean   :0.6250   Mean   : 3.262
##  3rd Qu.:0.001729   3rd Qu.:0.6842   3rd Qu.: 3.691
##  Max.   :0.022267   Max.   :1.0000   Max.   :18.996
##
## mining info:
##     data ntransactions support confidence
##  grocery          9835   0.001        0.5
```

Exploration using threholds

There are mainly three objective measures: support, confidence and lift.

```
# Generally explore rules
# Choose subset according to certain lift and confidence threholds (we use their mean in this case)
inspect(subset(groceryrules,subset=lift>3.262)[1:5])
```

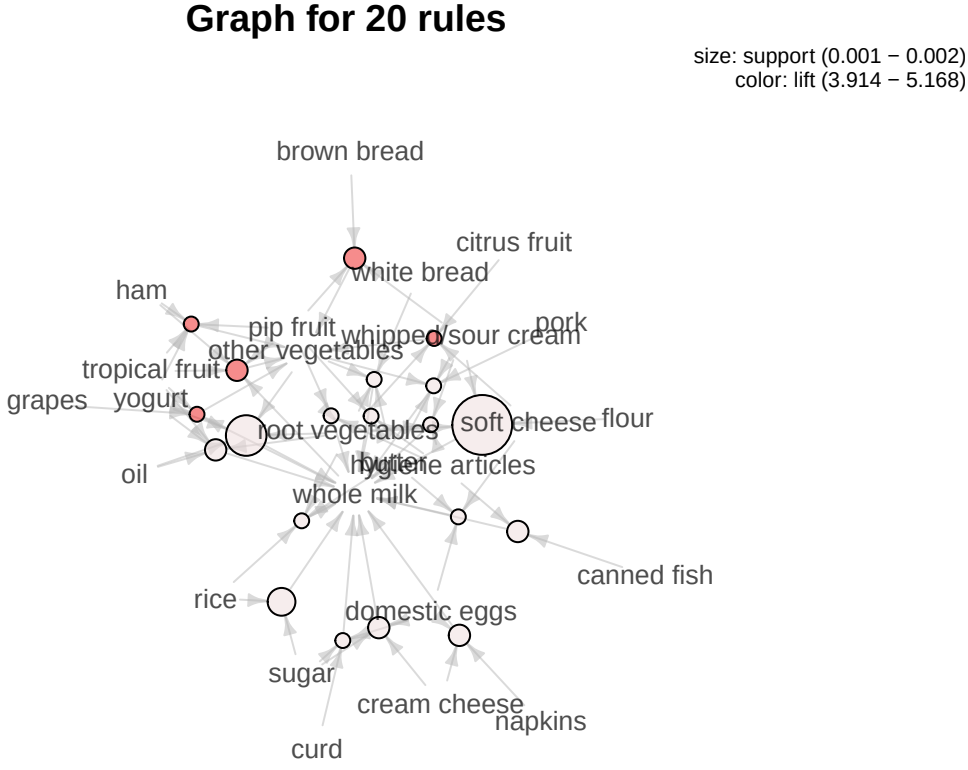
```
##     lhs                    rhs              support confidence     lift
## [1] {rice,
##      sugar}              => {whole milk} 0.001220132          1 3.913649
## [2] {canned fish,
##      hygiene articles}   => {whole milk} 0.001118454          1 3.913649
## [3] {butter,
##      rice,
##      root vegetables}    => {whole milk} 0.001016777          1 3.913649
## [4] {flour,
##      root vegetables,
```

```
##      whipped/sour cream} => {whole milk} 0.001728521          1 3.913649
## [5] {butter,
##      domestic eggs,
##      soft cheese}        => {whole milk} 0.001016777          1 3.913649
```

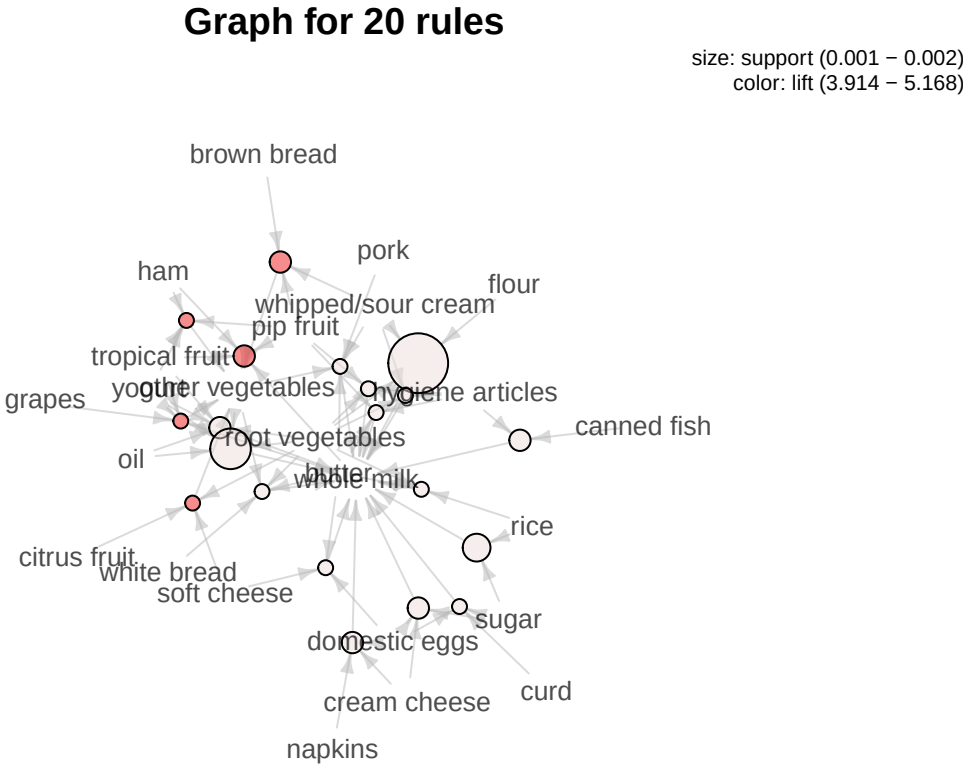
```
inspect(subset(groceryrules, subset=confidence > 0.6250)[1:5])
```

```
##     lhs                     rhs              support confidence     lift
## [1] {rice,
##      sugar}              => {whole milk} 0.001220132          1 3.913649
## [2] {canned fish,
##      hygiene articles}   => {whole milk} 0.001118454          1 3.913649
## [3] {butter,
##      rice,
##      root vegetables}    => {whole milk} 0.001016777          1 3.913649
## [4] {flour,
##      root vegetables,
##      whipped/sour cream} => {whole milk} 0.001728521          1 3.913649
## [5] {butter,
##      domestic eggs,
##      soft cheese}        => {whole milk} 0.001016777          1 3.913649
```

```
plot(head(subset(groceryrules,subset=lift>3.262), 20), method = "graph", control=list(cex=.8))
```




```
plot(head(subset(groceryrules, subset=confidence > 0.6250), 20), method = "graph", control=list(cex=.8))
```



Observation:

In the case of lift is higher than its mean, there are some associations related to whole milk. Lift could be thought as how much more likely an item is to be purchased given that it is known that another item has been purchased relative to its general purchase rate. For example, with rice and sugar,it is almost four times more likely that whole milk is going to be purchased than in the general grocery purchase.

In the case of confidence is higher than its mean, there are also some associations related to whole milk.Confidence represents how likely a rule is. For example,rice,sugar associated with whole milk is a rule that has a confidence "1".

Interpretation and discussion: In this case, it is their means that are as threholds because we targeted rules with above_than_average threholds level.

Pontential application and suggestions: Associations that could be used are like rice, sugar with wholemilk. Grocery stores could position these three product items closely.

However,as when lift is high, it could be the case that support is low, which means that the itemsets are rare in all grocery transactions.And rules that hold 100% of the time may not have the highest possible lift. As a result, method above has somewhat problmatic.

Exploration with subjective selection and objective measure (contradicting or actionable)

```
inspect(subset(groceryrules, subset=support > 0.01 & confidence > 0.5 & lift>3))
```

```
##     lhs                   rhs                    support confidence     lift
## [1] {citrus fruit,
##      root vegetables} => {other vegetables} 0.01037112  0.5862069 3.029608
## [2] {root vegetables,
##      tropical fruit}  => {other vegetables} 0.01230300  0.5845411 3.020999
```

Observation:

After trying different combinations of threholds, we chose the above one. In this case, there are associations of other vegetables with citrus fruit,root vegetables, and with root vegetables,tropical fruit.

Interpretation and discussion: As in this case, the corresponding association rule is actionable, we use these threholds.

Pontential application and suggestions: Associations that could be used are like other vegetables with citrus fruit,root vegetables, and with root vegetables,tropical fruit. Grocery stores could position these these product items closely.