Association analysis on the groceries data

Example: Identifying Frequently-Purchased Groceries —-

Step 1: Collecting Data

Our market basket analysis will utilize the purchase data collected from one month of operation at a real-world grocery store. The data contains 9,835 transactions or about 327 transactions per day (roughly 30 transactions per hour in a 12-hour business day), suggesting that the retailer is not particularly large, nor is it particularly small. The dataset used here was adapted from the Groceries dataset in the arules R package.

Step 2: Exploring and preparing the data

load the grocery data into a sparse matrix

**library**(arules)

**library**(arulesViz)

**library**(colorspace)

Theory:

* [Rules](https://rstudio-pubs-static.s3.amazonaws.com/356691_173bf83f376f444d8e840056bf47b3dd.html#rules)
  + [Correlation analysis](https://rstudio-pubs-static.s3.amazonaws.com/356691_173bf83f376f444d8e840056bf47b3dd.html#correlation-analysis)
* [Inspect top 5 rules](https://rstudio-pubs-static.s3.amazonaws.com/356691_173bf83f376f444d8e840056bf47b3dd.html#inspect-top-5-rules)
* [Induction](https://rstudio-pubs-static.s3.amazonaws.com/356691_173bf83f376f444d8e840056bf47b3dd.html#induction)
* [Jaccard Index](https://rstudio-pubs-static.s3.amazonaws.com/356691_173bf83f376f444d8e840056bf47b3dd.html#jaccard-index)
* [Advcanced graphics](https://rstudio-pubs-static.s3.amazonaws.com/356691_173bf83f376f444d8e840056bf47b3dd.html#advcanced-graphics)
* [Treemap](https://rstudio-pubs-static.s3.amazonaws.com/356691_173bf83f376f444d8e840056bf47b3dd.html#treemap)
* [Less than likely? - Lift < 1](https://rstudio-pubs-static.s3.amazonaws.com/356691_173bf83f376f444d8e840056bf47b3dd.html#less-than-likely---lift-1)

Association rules are statements that help to find patterns in seemingly unrelated data or a relational database (information repository). Easy example of such would be: If I buy milk, there is 80% probability that I will also buy yogurt"

An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the if/then statements have been found to be true. In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in shopping basket data analysis, product clustering, catalog design and store layout.

data(Groceries)

transactions <- Groceries

summary(transactions)

nrow(transactions)

**Interpret:** Density of 0.026 means that there are 2.6% non zero cells in the matrix. Matrix has 9835 times 169 = 1662115 cells. Since 2.6% of that are non-zero cells, so 4.336710^{4} items were purchased.

Average transaction consisted of 4.409456 items, whereas only one item have been bought in 2159 transactions. Maximum number of items bought was 32.

Let us proceed to frequency plots. The more frequent the item will be in transaction the higher its bar. Morover there are plots with different support levels. Support is the frequency of the pattern in the rule, therefore it being set to 0.1 means that the item must occur at least 10 times in 100 transactions. That is why the second plot has more items. Other way of selecting desired number of elements is to provide not support, but just the desired number. This is presented on the third graph.

On an average, each itemset or basket contains 4 to 5 items. In other words, basket having less than 5 items is more frequent as compare to baskets having more than 15 items. Buyers generally come to purchase fewer items from the shop. Support being set to .01 means that plot only includes item set having more than 1 repetition in each 100 transactions. Anything less than that is ignored for the study.

itemFrequencyPlot(transactions, support=0.1, cex.names=0.8)

itemFrequencyPlot(transactions, support=0.05, cex.names=0.8)

itemFrequencyPlot(transactions, topN=20)

**Correlation analysis**

The lift score . Lift = 1 ??? A and B are independent . Lift > 1 ??? A and B are positively correlated . Lift < 1 ??? A and B are negatively correlated

Firstly let us try the eclat algorithm - to see most frequent itemsets. Below we will see the list of the most common items together with their individual support.

freq.itemsets <- eclat(transactions, parameter=list(supp=0.075, maxlen=15))

inspect(freq.itemsets)

# Apriori Algorithm

rules <- apriori(Groceries, parameter = list(support = 0.009, confidence = 0.25, minlen = 2))

summary(rules)

We obtained a set of 224 rules, where mean support is equal to 16% and mean confidence is 37%. These are not bad values. It means that mean rule occurrs in 16% transactions and its implication has 37% power.

# Inspect top 5 rules

inspect(head(sort(rules, by ="lift"),5))

# Let us see rules that have high support and high confidence.

inspect(sort(sort(rules, by ="support"),by ="confidence")[1:5])

**#Visualization**

inspect(sort(sort(rules, by ="support"),by ="confidence")[1:5])

#Induction

Below analyses depend on choosing one product and checking which products it implies or by which products it is implied.

**Beverages:**

milk.rules <- sort(subset(rules, subset = rhs %**in**% "whole milk"), by = "confidence") summary(milk.rules)

inspect(milk.rules)

is.significant(milk.rules, transactions)

is.maximal(milk.rules)

is.redundant(milk.rules)

plot(milk.rules, measure=c("support", "confidence"), shading="lift")

coke.rules <- sort(subset(rules, subset = rhs %**in**% "soda"), by = "confidence") summary(coke.rules)

inspect(coke.rules)

is.significant(coke.rules, transactions)

plot(coke.rules, measure=c("support", "confidence"), shading="lift")

**Inference:**

1. Analysis was aimed to see what makes people buy milk (what products to be exact). To do se we should choose subset of rules that has whole milk (or soda) in right hand side of a rule.
2. It turns out that most popular baskets are curd, yoghurt or fruits and vegetables. Seems like the most popular one-week ahead groceries we do.
3. On the other hand it seems that soda is mostly bought with either sweets (chocolate) or with beverages/meat. Looks like a party ahead!
4. Most of the rules are significant (Fisher’s exact test) apart from some of the least confident rules of milk buying.
5. We can also see on the scatter plot of rules for milk that the higher the confidence the higher lift, which was not observed before. It also occurs on Coke rules plot, but is not that visible.

**#Meat rules:**

In case of meat, we search whether meats like: beef, chicken (poultry) or sausage show up in the left hand sides of rules.

Let’s see what people buy after they have put meat (sausage or beef) to the basket. It turns out that the most popular option associated with meat is milk! It is a little bit confusing, because only in lift column we see how popular option is. The real winner here are root vegetables that are 3 times more likely to be put into the basket than other products. Rest of the products are just regular grocery stuff.

meat.rules <- sort(subset(rules, subset = lhs %**in**% "beef"|lhs %**in**% "sausage" |lhs %**in**% "chicken"), by = "confidence") summary(meat.rules)

inspect(meat.rules)

is.significant(meat.rules, transactions)

# **Yogurt rules:**

yog.rules <- sort(subset(rules, subset = lhs %**in**% "yogurt"), by = "confidence") summary(yog.rules)

inspect(yog.rules)

is.significant(yog.rules, transactions)

Same as above we subset only these rules that have yogurt in left hand side of a rule.

Most of the times someone buys yogurt he will also put milk or vegetables into his basket - with greater correlation to ‘other vegetables’. There is not much variation, nothing changes with the lowering confidence.

**Visualization: Some Visualization for above subrules**

*# plot for subrules* plot(meat.rules,method="graph",interactive=FALSE,shading="lift") title(main = "Meat")

plot(milk.rules,method="graph",interactive=FALSE,shading="lift") title(main = "Milk")

plot(yog.rules,method="graph",interactive=FALSE,shading="lift") title(main = "Yogurt")

plot(coke.rules,method="graph",interactive=FALSE,shading="lift") title(main = "Coke")

**Interpretation**: Above we can see graphs for the previously interpreted rules that have the same conclusions as previously. The reddier the circle the more probable is the client to buy two of those items than any other items and the bigger the circle the more probable is to buy two of those items. Moreover the arrow points to the direction of a possible basket rule. Therefore in case of Coke, we can notice bottled water and soda as the rule with highest support.

More complicated conclusions can be drawn from the meat rules plot. We can see that the sausage is the mostly supported additional product for milk.

**Jaccard Index**

For the set of milk rules let’s calculate the Jaccard Index. It is the representation of how much likely are two items to be bought together.

trans.sel<-transactions[,itemFrequency(transactions)>0.1] *# selected transactions* dissimilarity(trans.sel, which="items")

Because I have picked such high minimal frequency we have not much items, but moreover Jaccard Index seems to have high values telling us that most of those products do not overlap. Such an array as presented above tells that the higher the values of Jaccard Index the more likely are two products to be in the same transaction. Highest percentage is between root vegetables and tropical fruits.

**#Advanced Graphs**

plot(meat.rules, method="grouped", measure="support", control=list(col=sequential\_hcl(100)))

We can even show dependencies with parallel coordinates plot. We can see that the mostly red arrow (each of them represents one rule) connects beef and root vegetables. Moreover most of the arrows connect sausage on the first position, as previously stated.

plot(meat.rules, method="paracoord", control=list(reorder=TRUE))

**Matrix Plots:**

**Motivation:** Apart from the analytical study of the created rulesets and research of the rules for particular items, we can present more advanced graphics to more thoroughly analyze ruleset.

Let’s present the ruleset for meat but in a matrix form. Each of the matrix cells can have different blue shade depending on the lift value. Numbers on the axes are corresponding to the items listed before the matrix. For example the most blue cell corresponds to the rule {beef} -> {root vegetables}, hence (as previously mentioned) root vegetables are most likely to be bought with beef. On the second place is the chicken and for the rest of antecedent items there is no significant lift at all (it is too small to be presented on the graph). Such a graph is only confirmation fo the conclusions drawn before, but in a simplier form.

plot(meat.rules, method="matrix", measure=c("support","confidence"), control=list(reorder=TRUE, col=sequential\_hcl(200)))

II. A better way to present this is a Grouped matrix plot. It has the same data on the axes as before, but moreover it shows the support of the rules. It can be noted that rules connected with sausages have the biggest support (among listed). The previously concluded biggest lift for {beef} -> {root vegetables} is also noticeable.

plot(meat.rules, method="grouped", measure="support", control=list(col=sequential\_hcl(100)))

We can even show dependencies with parallel coordinates plot. We can see that the mostly red arrow (each of them represents one rule) connects beef and root vegetables. Moreover most of the arrows connect sausage on the first position, as previously stated.

plot(meat.rules, method="paracoord", control=list(reorder=TRUE))

Treemap (Avaliability Of Products)

If we want to look into data deeper, we can create interesting plots that show us how many products of each type are available to buy in the grocery store. Here I made two treemaps, including one with deeper segmentation that present which products are on the lists of the shop. Moreover it can explain why there are so many connections with milk and fresh products, whereas just a little with coke.

install.packages("magrittr") # package installations are only needed the first time you use it

install.packages("dplyr") # alternative installation of the %>%

library(magrittr) # needs to be run every time you start R and want to use %>%

library(dplyr) # alternatively, this also loads %>%

## Installing the package and calling the package in R## install.packages("treemap")

library(treemap)

Note: The pipe operator %>% was introduced to "decrease development time and to improve readability and maintainability of code."

treemap(occur2,index=c("level1", "level2"),vSize="n",title="",palette="Dark2",border.col="#FFFFFF")

Each of these charts have different level of deepth. First only shows the bigger group names (like aisles in the shop). Second shows deeper segmentation intro product types (for example within aisle). The last chart presents each of the products available - it does give us less information that the previous one.

**#Less than likely? - Lift < 1**

Interesting part of the study would be checking for items that are less than likely to be bought together. These would be described by lift < 1.

inspect(tail(sort(rules, by = "lift")))

There is only one item in our rules set, that has lift less than 1. It is a connection between whole milk and bottled beer. It means that we are less likely to buy milk than any other product in dataset, while already having beer in basket. Maybe that’s a hint that beeroholics don’t drink milk?