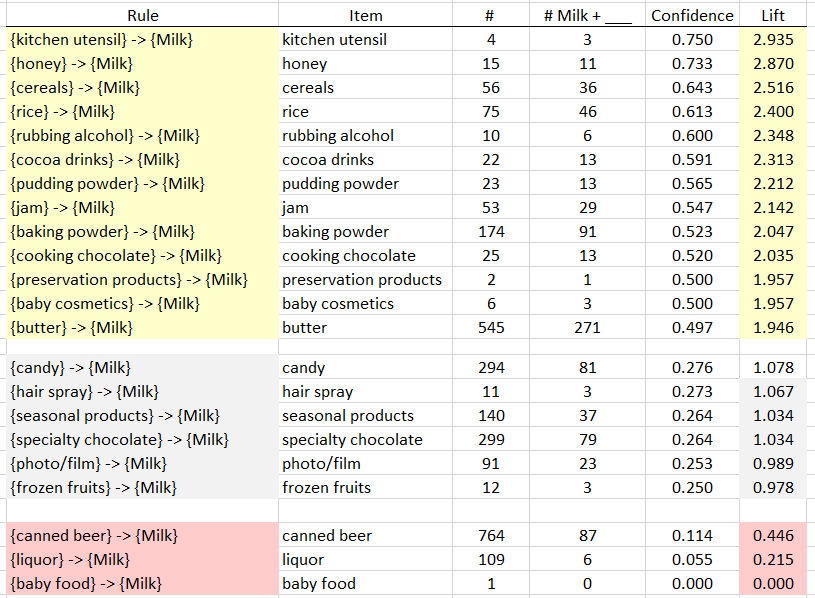
Handout 6: Association Rules

Market Basket Analyses are a common application of association rules. One goal of a market basket analysis is to understand the association between items purchases. The relationship between items purchased at a grocery store will be considered in this handout.

|  |  |
| --- | --- |
| http://imagesus.homeaway.com/mda01/f4e2379a-d720-459f-b9c0-9999fc559022.1.6 |  |

An association rule highlights the fact that some items are more (or less) indicative of the purchase of others. For example, purchasing cereal increases the likelihood of purchasing milk. These types of analyses may also reveals that liquor and milk are rarely purchased together.



**Association Rules** are used to uncover associations or relationships that exist between items. Often these rules are constructed to identify relationships between items purchased, i.e. Market Basket Analysis.

Procedural Steps

1. Determine how often items are purchased
2. Determine how often items are purchased in conjunction with other items
3. Identify which purchased items are indicative of others being purchased

Data Technologies

1. Filtering/Subsets
2. Creating Tables
3. Data Summaries

Consider the following subset of data from a collection of transactions from a grocery store.

|  |  |
| --- | --- |
| Transaction ID | Items Purchased |
| 1 | {Bread, Milk} |
| 2 | {Eggs, Ham} |
| 3 | {Bread, Fruit, Milk} |
| 4 | {Beer, Bread, Butter, Fruit, Soda} |
| 5 | {Bread, Fruit, Milk, Soda} |

Association rules are developed under the following guiding principles.

|  |  |  |
| --- | --- | --- |
| 1. | Items should be purchased somewhat often | **Support** |
| 2. | Reliability, i.e. the degree to which one set of items predicts the purchase of another set of items | **Confidence** |

Consider the following association rule – the purchase of Bread indicates the purchase of Milk.

|  |  |
| --- | --- |
| Rule #1 |  |

Compute the support and confidence for this rule.

Questions

1. What is the interpretation of the Support(Bread AND Milk)?

1. What is the interpretation of Confidence of this rule? Discuss.  
   Note: Confidence is simply a conditional probability, i.e P(Milk | Bread).

Consider a second association rule for the purchase of Milk.

|  |  |
| --- | --- |
| Rule #2 |  |

Compute the support and confidence for this rule.

Question

1. Why might Rule #1 be considered “better” than Rule #2 when interest lies in the purchase of Milk?

Consider a third association rule for the purchase of Milk.

|  |  |
| --- | --- |
| Rule #3 |  |

Compute the support and confidence for this rule.

**Lift** is another measure often considered when evaluating rules of association.

For our example, realize that the support for Milk is fairly large. i.e, Milk was purchased in 60% of the transactions. This provides a baseline value for confidence. That is, rules that exceed this value indicate gains when considering the association provided by the rule. When the lift of a rule is near 1, then the rule provides little information to understanding the purchase of the item.

* implies positive association between items
* implies no association between items
* implies negative association between items

|  |  |  |  |
| --- | --- | --- | --- |
| **Rule** | **Support** | **Confidence** | **Lift** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Some Comments

* Association rules with no support have zero confidence. E.g. Beer is never purchased with Milk, so the rule should not be considered.
* The confidence of a rule should not be considered independent of it’s support. For example, the rule has Confidence = 1. That is, 100% of the time eggs were purchased, so was Ham. However, this rule has very low support as Eggs and Ham were only purchased once.
* Association rules are not invariant. For example, the confidence for the rule is different than the confidence of the rule .

Common Data Structure

|  |  |  |
| --- | --- | --- |
| List |  | Binary Representation (Matrix) |

Next, let us consider the complete grocery dataset. Open the Grocery dataset in R. This dataset contains 9835 transactions and 169 unique items.



The grocery dataset consists of the binary representation of this market basket data.



Using simple logic statements on the columns of Whole Milk and Butter allows one to easy compute the support and confidence for the rule .

|  |  |  |  |
| --- | --- | --- | --- |
| |  |  | | --- | --- | | Rule |  | | |
|  |  |
|  | |

Evaluating Several Rules

The procedure provided above lack efficiencies and does not scale well when several rules need to be evaluated. For example, there are 169 items in this analysis, a more efficient process would be to write a loop to cycle through all columns automatically. This is the purpose of the code provided below.

|  |  |
| --- | --- |
| Rule #1 |  |
| Rule #2 |  |
| : | : |
| Rule #169 |  |

> #Prepare a data.frame to dump outcomes

> output<-data.frame(Item=rep(NA, 169), Numberi=rep(NA,169),Numberi\_Wholemilk=rep(NA,169),Confidence=rep(NA,169),Lift=rep(NA,169))

>

>

> #Get support for whole milk

> support.wholemilk <- sum(Groceries[,25]==1) / length(Groceries[,25])

>

> #Loop through all columns, skip column 25 as this contains whole milk

> for(i in 1:169){

+

+ output[i,1]<- colnames(Groceries)[i]

+

+ if(i != 25){

+ support\_i\_wholemilk = sum(Groceries[,i]==1 & Groceries[,25]==1) / length(Groceries[,i])

+ support\_i = sum(Groceries[,i]==1) / length(Groceries[,i])

+ confidence = support\_i\_wholemilk / support\_i

+ lift = confidence / support.wholemilk

+

+ output[i,2] = sum(Groceries[,i]==1)

+ output[i,3] = sum(Groceries[,i]==1 & Groceries[,25]==1)

+ output[i,4] = confidence

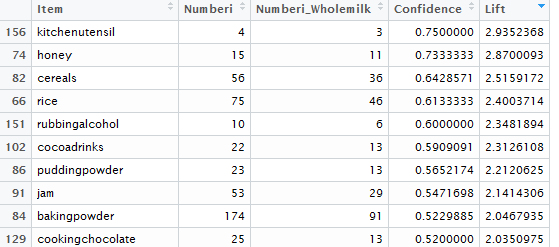
+ output[i,5] = lift

+ }

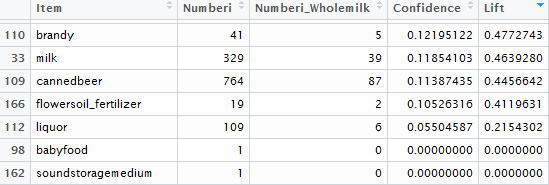
+ }

The output can be displayed via the command View(output). Sorting of this data.frame can be done by clicking the arrow next to lift.

The following items have high lift, i.e. much greater than 1, – indicating that there is a high probability these items were purchased with Whole Milk.



The following items have a lift value much smaller than 1 – indicating that there is a low probability these items were purchased with Whole Milk.



The following can be used to sort data.frame explicitly in R.

> #Sort the outptu data.frame by lift, largest to smallest

> output[order(-1\*output[,5]),]

Questions

1. The Lift for is about 2.5 which is fairly high. Thus, given that the transaction includes cereal, there is 2.5 fold increase in the likelihood of whole milk being purchased.
   1. Compute Support(Cereal AND Milk).
   2. This value is fairly low. Why does a low support value negate the usefulness of a rule?
2. The Lift value for the rule is lowest on this list. What can be said about the purchase of Liquor AND Whole Milk?

Association Rules in R

The **arules** package in R can be used to expand upon what we’ve done above. The following code can be used to recreate the table above. The maxlen=2 specification in the apriori() function restricts rules to single items. The subset() function reduces the rules – here to include only rules for which whole milk is on the right-hand side.

> #Arules code

> library(arules)

>

> #All variaables must be factors, forcing factor level on each column

> for(i in 1:169){

+ Groceries[,i] <- as.factor(Groceries[,i])

+ }

>

> #Forcing Groceries to a transaction object

> gr.trans = as(Groceries,"transactions")

>

> #Rule development vai apriori function

> gr.rules = apriori(gr.trans, parameter = list(supp = 0.0, conf = 0.0, maxlen=2))

>

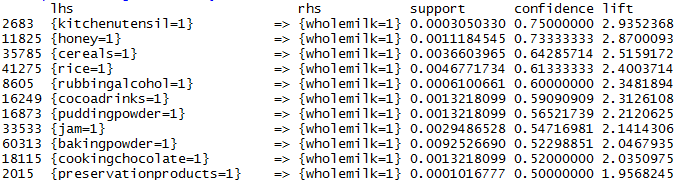
> #Rules that have whole milk on right side

> gr.subset = subset(gr.rules, subset = rhs %in% "wholemilk=1")

>

> #Print rules to screen - sorted by lift, view top 10

> inspect(sort(gr.subset,by="lift")[1:10])



The following parameter specification in the apiori() function will limit rules with support larger than 0.01 and confidence larger than 0.25.

> #Rule development vai apriori function

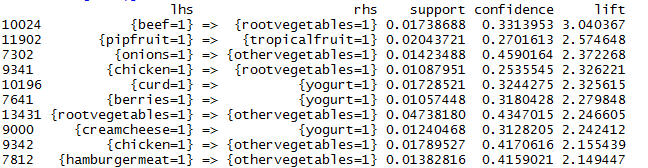
> gr.rules = apriori(gr.trans, parameter = list(supp = 0.01, conf = 0.25, maxlen=2))

> #Print rules to screen - sorted by lift

> outcomes<-inspect(sort(gr.subset,by="lift"))

> #Looking at top 10 outcomes

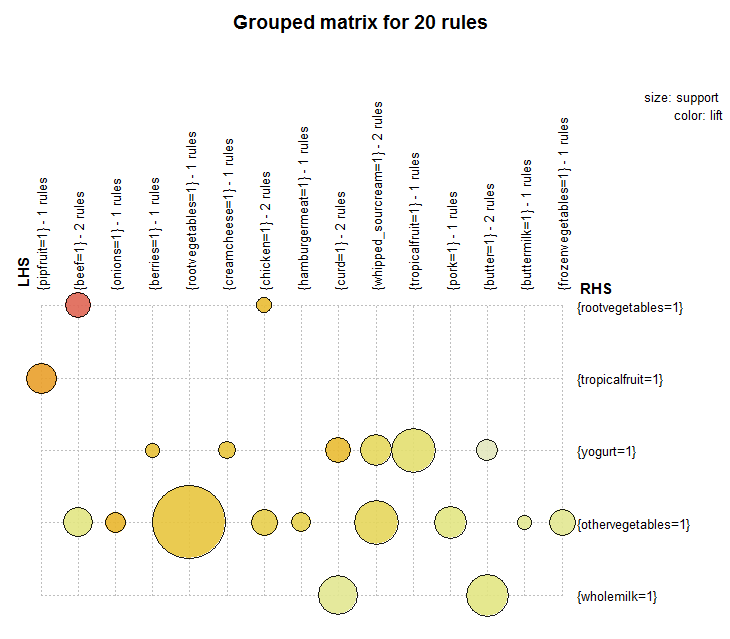
> outcomes[1:10,]



The arulesViz package contains functions for plotting output from the arules. The following was used to make the plot below.

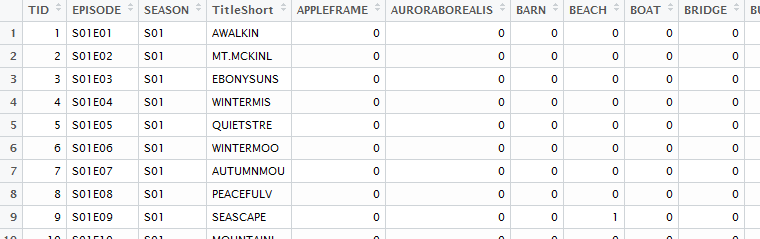
|  |
| --- |
| #Rule development via apriori function  gr.rules = apriori(gr.trans, parameter = list(supp = 0.01, conf = 0.25, maxlen=2))  #Print rules to screen - sorted by lift  outcomes<-head(sort(gr.rules,by="lift"),20)  library(arulesViz)  plot(outcomes, method="grouped") |

The bubble size is proportional to the support and color is proportion to the lift (darker is high lift).



**Task**

|  |  |
| --- | --- |
| For this task, we will consider the BobRoss dataset. Bob Ross was a painter than was often on public television. The purpose of this analysis will be to determine which features were commonly used in conjunction with each other. | http://a.scpr.org/i/16c112c4e72c74fdafc615917a92e63c/43786-full.jpg |

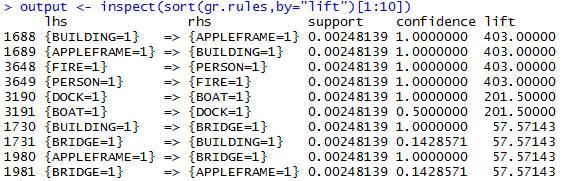


Run the following code

|  |
| --- |
| #BobRoss A Rules analysis  #Create a subset to inclue only the feature columns  BobRoss2 <- BobRoss[,5:71]  #All variables must be factors, forcing factor level on each column  for(i in 1:67){  BobRoss2[,i] <- as.factor(BobRoss2[,i])  }  #Forcing Groceries to a transaction object  gr.trans = as(BobRoss2,"transactions")  #Rule development vai apriori function  gr.rules = apriori(gr.trans, parameter = list(supp = 0.0, conf = 0.0, maxlen=2))  #Rules that have whole milk on right side  #gr.subset = subset(gr.rules, subset = rhs %in% "wholemilk=1")  #Print rules to screen - sorted by lift  output <- inspect(sort(gr.rules,by="lift"))  #Top 10 via lift  output[1:10,] |

Questions:

* 1. The top 10 rules returned from the above code are provided below. There rules have a very high lift value; however, the support is very low. Why are these rules not very useful when the support is so low? Why does this imply about the occurrence of these features?



* 1. Run the following code what restrict support to be above 0.05 and confidence to be above 0.20. Beach and Waves appear on the right-hand side often. What features tend to go along with Beach and Waves?

> #Rule development vai apriori function

> gr.rules = apriori(gr.trans, parameter = list(supp = 0.05, conf = 0.20, maxlen=3))

> #Print rules to screen, sorted by lift

> output <- inspect(sort(gr.rules,by="lift")[1:10])

* 1. Using the arulesViz package, create a plot similar to the one below for the rules you have investigate above.

