Introduction to Data Science

DSA1101

Semester 1, 2018/2019
Week 10

Association Rules

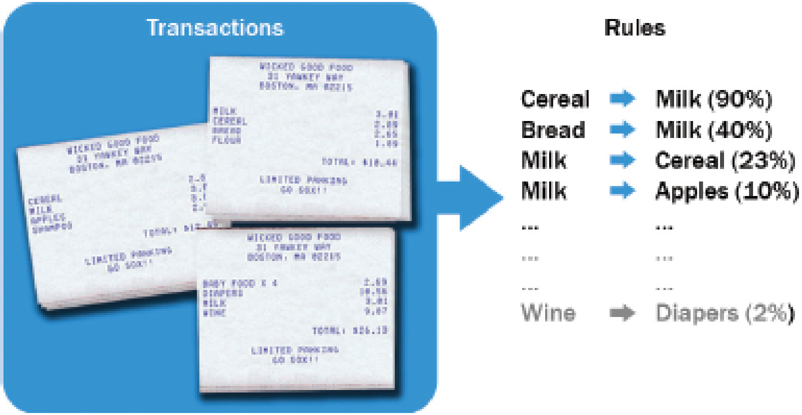
Association Rules

- This week, we will study another unsupervised learning method called association rules.
- This is a descriptive, not predictive, method often used to discover interesting relationships hidden in a large dataset.
- The disclosed relationships can be represented as rules or frequent itemsets.
- Association rules are commonly used for mining transactions in databases.

Association Rules

- For example, given a large collection of retail transactions, in which each transaction consists of one or more items, association rules go through the items being purchased to see what items are frequently bought together and to discover a list of rules that describe the purchasing behavior.
- The goal with association rules is to discover interesting relationships among the items.
- The relationships that are interesting depend both on the business context and the nature of the algorithm being used for the discovery.

Association Rules



Source: *Data Science & Big Data Analytics*

Association Rules

- For example, given a large collection of retail transactions, in which each transaction consists of one or more items, association rules go through the items being purchased to see what items are frequently bought together and to discover a list of rules that describe the purchasing behavior.
- The goal with association rules is to discover interesting relationships among the items.
- The relationships that are interesting depend both on the business context and the nature of the algorithm being used for the discovery.

Association Rules

- In the example of a retail store, association rules are used over transactions that consist of one or more items.
- In fact, because of their popularity in mining customer transactions, association rules are sometimes referred to as market basket analysis.
- Each transaction can be viewed as the shopping basket of a customer that contains one or more items.
- This is also known as an *itemset*.
- The term *itemset* refers to a collection of items or individual entities that contain some kind of relationship.

Association Rules

- An itemset containing k items is called a k-itemset.
- We will use the notation $\{\text{item 1}, \text{item 2}, ..., \text{item k}\}$ to denote a k-itemset.
- Computation of the association rules is typically based on itemsets.
- We will focus on the *Apriori* algorithm for generating association rules

Apriori Algorithm

- One major component of *Apriori* is support.
- Given an itemset L, the *support* of L is the percentage of transactions that contain L.
- For example, if 80% of all transactions contain itemset $\{bread\}$, then the support of $\{bread\}$ is 0.8. L can be 1-itemset
- Similarly, if 60% of all transactions contain itemset $\{bread, butter\}$, then the support of $\{bread, butter\}$ is 0.6.

  support of {bread} > support of {bread, butter}

  support of A > support of B, if A is a subset of B

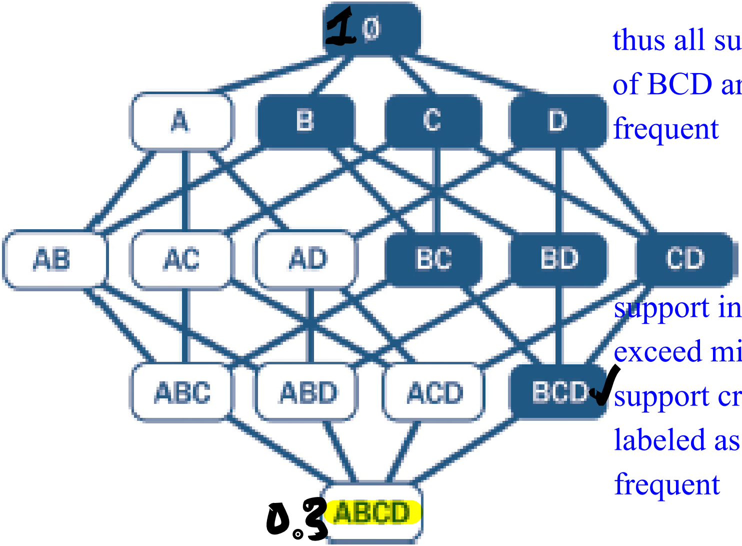
Apriori Algorithm

- A frequent itemset has items that appear together *often enough*.
- The term "often enough" is formally defined with a minimum support criterion.
- If the minimum support is set at 0.5, any itemset can be considered a frequent itemset if at least 50% of the transactions contain this itemset.
- For the previous example, both $\{bread\}$ and $\{bread, butter\}$ are considered frequent itemsets at the minimum support 0.5.
- If the minimum support is 0.7, only $\{bread\}$ is considered a frequent itemset.

Apriori Algorithm

- If an itemset is considered frequent, then any subset of the frequent itemset must also be frequent.
- This is referred to as the *Apriori property* (or downward closure property).
- For example, if 60% of the transactions contain $\{bread, jam\}$, then at least 60% of all the transactions will contain $\{bread\}$ or $\{jam\}$.
- The *Apriori property* provides the basis for the *Apriori* algorithm.

Association Rules



thus all subsets of BCD are frequent

support increases, exceed minimum support criteria, labeled as frequent

Source: *Data Science & Big Data Analytics*

Apriori Algorithm

- The Apriori algorithm takes a bottom-up iterative approach to uncovering the frequent itemsets by first determining all the possible items (or 1-itemsets, for example $\{bread\}$, $\{eggs\}$, $\{milk\}$,...) and then identifying which among them are frequent based on a minimum support threshold (or the minimum support criterion)
- For example, when the minimum support threshold is set at 0.5, the algorithm identifies and retains those itemsets that appear in at least 50% of all transactions and discards (or "prunes away") the itemsets that have a support less than 0.5 or appear in fewer than 50% of the transactions.

Apriori Algorithm

- In the next iteration of the Apriori algorithm, the identified frequent 1-itemsets are paired into 2-itemsets (for example, $\{bread, eggs\}$, $\{bread, milk\}$, $\{eggs, milk\}$,...) and again evaluated to identify the frequent 2-itemsets among them.
- This iterative process is repeated in the Apriori algorithm.
- At each iteration, the algorithm checks whether the support criterion can be met; if it can, the algorithm grows the itemset, repeating the process until it runs out of support or until the itemsets reach a predefined length.
  e.g {bread, milk}, {eggs, milk} meets the minimum
  -> {bread, eggs, milk}

Apriori Algorithm

- The growing and pruning process is repeated until no itemsets meet the minimum support threshold.
- Optionally, a threshold N can be set up to specify the maximum number of items the itemset can reach or the maximum number of iterations of the algorithm.
- Once completed, output of the *Apriori* algorithm is the collection of all the frequent k-itemsets.

Apriori Algorithm

- Finally, a collection of candidate rules is formed based on the frequent k-itemsets uncovered in the iterative process described earlier.
- For example, a frequent itemset $\{milk, eggs\}$ may suggest candidate rules $\{milk\} \rightarrow \{eggs\}$ and $\{eggs\} \rightarrow \{milk\}$.
- We can evaluate the appropriateness of these candidate rules using measures such as confidence, lift, and leverage.

Evaluation of Candidate Rules

- *Confidence* is defined as the measure of certainty or trustworthiness associated with each discovered rule.
- Mathematically, the confidence for candidate rule $X \rightarrow Y$ is the percent of transactions that contain both X and Y out of all the transactions that contain X:

$$Confidence(X \rightarrow Y) = \frac{Support(X \wedge Y)}{Support(X)} \leq 1$$

- For example, if $\{bread, eggs, milk\}$ has a support of 0.15 and $\{bread, eggs\}$ also has a support of 0.15, the confidence of rule $\{bread, eggs\} \rightarrow \{milk\}$ is 1, which means 100% of the time a customer buys bread and eggs, milk is bought as well.
- The rule is therefore correct for 100% of the transactions containing bread and eggs.

Evaluation of Candidate Rules

- A relationship may be thought of as interesting when the algorithm identifies the relationship with a measure of confidence greater than or equal to a predefined threshold.
- This predefined threshold is called the minimum confidence.
- A higher confidence indicates that the rule ($X \rightarrow Y$) is more interesting or more trustworthy, based on the sample dataset.

Evaluation of Candidate Rules

- Even though confidence can identify the interesting rules from all the candidate rules, it comes with a problem.
- Given rules in the form of $X \rightarrow Y$, confidence considers only the antecedent (X) and the cooccurrence of X and Y; it does not take the consequent of the rule (Y) into concern.
- Other measures such as lift and leverage are designed to address this issue.

Evaluation of Candidate Rules

- *Lift* measures how many times more often X and Y occur together than expected if they are statistically independent of each other.
- *Lift* is a measure of how X and Y are really related rather than coincidentally happening together:

$$Lift(X \rightarrow Y) = \frac{Support(X \wedge Y)}{Support(X) \times Support(Y)}$$

- Lift is 1 if X and Y are statistically independent of each other.
- In contrast, a lift of $X \rightarrow Y$ greater than 1 indicates that there is some usefulness to the rule.

Evaluation of Candidate Rules

- For example, assuming 1,000 transactions, with $\{milk, eggs\}$ appearing in 300 of them, $\{milk\}$ appearing in 500, and $\{eggs\}$ appearing in 400, then $Lift(milk \rightarrow eggs) = 0.3/(0.5 \times 0.4) = 1.5$.
- If $\{bread\}$ appears in 400 transactions and $\{milk, bread\}$ appears in 400, then $Lift(milk \rightarrow bread) = 0.4/(0.5 \times 0.4) = 2$.

Evaluation of Candidate Rules

- *Leverage* is a similar notion, but instead of using a ratio, leverage uses the difference:

$$Leverage(X \rightarrow Y) = Support(X \wedge Y) - Support(X) \times Support(Y)$$

- *Leverage* measures the difference in the probability of X and Y appearing together in the dataset compared to what would be expected if X and Y were statistically independent of each other.
- In theory, leverage is 0 when X and Y are statistically independent of each other.
- If X and Y have some kind of relationship, the leverage would be greater than zero.
- A larger leverage value indicates a stronger relationship between X and Y.

Evaluation of Candidate Rules

- For example, assuming 1,000 transactions, with $\{milk, eggs\}$ appearing in 300 of them, $\{milk\}$ appearing in 500, and $\{eggs\}$ appearing in 400, then $Leverage(milk \rightarrow eggs) = 0.3 - (0.5 \times 0.4) = 0.1$.
- If $\{bread\}$ appears in 400 transactions and $\{milk, bread\}$ appears in 400, then $Leverage(milk \rightarrow bread) = 0.4 - (0.5 \times 0.4) = 0.2$.

Applications of Association Rules

- The term market basket analysis refers to a specific implementation of association rules mining that many companies use for a variety of purposes, including these:
  1. Broad-scale approaches to better merchandising-what products should be included in or excluded from the inventory each month
  2. Cross-merchandising between products and high-margin or high-ticket items
  3. Physical or logical placement of product within related categories of products
  4. Promotional programs-multiple product purchase incentives managed through a loyalty card program

Applications of Association Rules

- Besides market basket analysis, association rules are commonly used for recommender systems and clickstream analysis

Applications of Association Rules

- Many online service providers such as Amazon and Netflix use recommender systems.
- Recommender systems can use association rules to discover related products or identify customers who have similar interests.
- For example, association rules may suggest that those customers who have bought product A have also bought product B.
- These findings provide opportunities for retailers to cross-sell their products.

Applications of Association Rules

- Clickstream analysis refers to the analytics on data related to web browsing and user clicks, which is stored on the client or the server side.
- Web usage log files generated on web servers contain huge amounts of information, and association rules can potentially give useful knowledge to web usage data analysts.
- For example, association rules may suggest that website visitors who land on page X click on links A, B, and C much more often than links D, E, and F.
- This observation provides valuable insight on how to better personalize and recommend the content to site visitors.

Association Rules: Example in R

- We will look at an example illustrating the application of the *Apriori algorithm* to grocery store transaction data.
- Using R and the 'arules' and 'arulesViz' packages, this example shows how to use the *Apriori algorithm* to generate frequent itemsets and rules and to evaluate and visualize the rules.
-

```
install.packages('arules')
install.packages('arulesViz')
library('arules')
library('arulesViz')
```

Association Rules: Example in R

- The example uses the Groceries dataset from the R arules package.
- The Groceries dataset is collected from 30 days of real-world point-of-sale transactions of a grocery store.
- The dataset contains 9,835 transactions, and the items are aggregated into 169 categories.
-

```
> data(Groceries)
> Groceries
transactions in sparse format with
 9835 transactions (rows) and
 169 items (columns)
```

Association Rules: Example in R

- The summary shows that the most frequent items in the dataset include items such as whole milk, other vegetables, rolls/buns, soda, and yogurt. These items are purchased more often than the others.
-

```
> summary(Groceries)
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
            2513             1903
      rolls/buns             soda
            1809             1715
          yogurt          (Other)
            1372            34055
```

Association Rules: Example in R

- The class of the dataset is transactions, as defined by the arules package. The transactions class contains three slots:
  1. transactionInfo: A data frame with vectors of the same length as the number of transactions
  2. itemInfo: A data frame to store item labels
  3. data: A binary incidence matrix that indicates which item labels appear in every transaction

Association Rules: Example in R

- `Groceries@itemInfo` display all 169 grocery labels as well as their categories.
-

```
> Groceries@itemInfo[1:10,]
               labels  level2           level1
1         frankfurter sausage meat and sausage
2             sausage sausage meat and sausage
3          liver loaf sausage meat and sausage
4                 ham sausage meat and sausage
5                meat sausage meat and sausage
6   finished products sausage meat and sausage
7     organic sausage sausage meat and sausage
8             chicken poultry meat and sausage
9              turkey poultry meat and sausage
10               pork    pork meat and sausage
```

Association Rules: Example in R

- `Groceries@data` indicates which item labels appear in every transaction.
- | indicates that the item appears in transaction, and · otherwise.
-

```
> Groceries@data[,100:110]
169 x 11 sparse Matrix of class "ngCMatrix"

  [1,] . . . | . . . . . . .
  [2,] . . | . | . . . . . .
  [3,] . . . . . . . . . . .
  [4,] . . . . . . . . . . .
  [5,] . . . . . . . . . . .
  [6,] . . . . . . . . . . .
  [7,] . . . . . . . . . . .
  [8,] . . . . . . . . . . .
  [9,] . . . . . . . . . . .
 [10,] . . | . . . . . . . .
 [11,] . . . . . . . | . . .
 [12,] . . . . . . . . . . .
```

Association Rules: Example in R

- The following code displays the 100^{th} to 105^{th} transactions of the Groceries dataset.
- $[100:110]$ can be changed to $[1:9835]$ to display all the transactions.
-

```
> apply(Groceries@data[,100:105], 2,
+ function(r) paste(Groceries@itemInfo[r,"labels"
    ], collapse=", ")
+ )
[1] "citrus fruit, tropical fruit"
[2] "soda, misc. beverages"
[3] "sausage, pork, grapes, whole milk, rolls/buns
    , pastry, soda, specialty bar, bathroom
    cleaner"
[4] "frankfurter, rolls/buns, bottled water"
[5] "sausage, whole milk, yogurt, coffee, fruit/
    vegetable juice, bottled beer, softener,
    napkins, photo/film, shopping bags"
[6] "soda"
```

Association Rules: Example in R

- The `apriori()` function from the ‘arule’ package implements the *Apriori algorithm* to create frequent itemsets.
- Note that, by default, the `apriori()` function executes all the iterations at once.
- However, to illustrate how the *Apriori algorithm* works, the code examples in this section manually set the parameters of the `apriori()` function to simulate each iteration of the algorithm.

Association Rules: Example in `R`

- Assume that the minimum support threshold is set to 0.02 based on management discretion.
- Because the dataset contains 9,853 transactions, an itemset should appear at least 198 times to be considered a frequent itemset.
- The first iteration of the *Apriori algorithm* computes the support of each product in the dataset and retains those products that satisfy the minimum support.
- The following code identifies 59 frequent 1-itemsets that satisfy the minimum support.
- The parameters of `apriori()` specify the minimum and maximum lengths of the itemsets, the minimum support threshold, and the target indicating the type of association mined.

Association Rules: Example in R

```
> itemsets <- apriori(Groceries, parameter=list(
    minlen=1, maxlen=1,
+ support=0.02, target="frequent itemsets"))
> summary(itemsets)
set of 59 itemsets

...
summary of quality measures:
    support            count
 Min.   :0.02105   Min.   : 207.0
 1st Qu.:0.03015   1st Qu.: 296.5
 Median :0.04809   Median : 473.0
 Mean   :0.06200   Mean   : 609.8
 3rd Qu.:0.07666   3rd Qu.: 754.0
 Max.   :0.25552   Max.   :2513.0
```

Association Rules: Example in R

- The following code uses the `inspect()` function to display the top 10 frequent 1-itemsets sorted by their support.
-

```
> inspect(head(sort(itemsets, by = "support"), 10)
    )
      items                support    count
[1]   {whole milk}         0.25551601 2513
[2]   {other vegetables}   0.19349263 1903
[3]   {rolls/buns}         0.18393493 1809
[4]   {soda}               0.17437722 1715
[5]   {yogurt}             0.13950178 1372
[6]   {bottled water}      0.11052364 1087
[7]   {root vegetables}    0.10899847 1072
[8]   {tropical fruit}     0.10493137 1032
[9]   {shopping bags}      0.09852567  969
[10]  {sausage}            0.09395018  924
```

Association Rules: Example in R

- In the next iteration, the list of frequent 1-itemsets is joined onto itself to form all possible candidate 2-itemsets.
- For example, 1-itemsets $\{wholemilk\}$ and $\{soda\}$ would be joined to become a 2-itemset $\{wholemilk, soda\}$.
- The algorithm computes the support of each candidate 2-itemset and retains those that satisfy the minimum support.
- The output that follows shows that 61 frequent 2-itemsets have been identified.
-

```
> itemsets <- apriori(Groceries, parameter=list(
    minlen=2, maxlen=2,
+ support=0.02, target="frequent itemsets"))
> summary(itemsets)
set of 61 itemsets
...
```

Association Rules: Example in R

- The top 10 most frequent 2-itemsets are displayed next, sorted by their support.
- Notice that whole milk appears six times in the top 10 2-itemsets ranked by support.
- As seen earlier, whole milk has the highest support among all the 1-itemsets.
- These top 10 2-itemsets with the highest support may not be interesting; this highlights the limitations of using support alone.
- The output that follows shows that 61 frequent 2-itemsets have been identified.

Association Rules: Example in R

```
> inspect(head(sort(itemsets, by ="support"),10))
     items                       support    count
[1]  {other vegetables,
      whole milk}              0.07483477   736
[2]  {whole milk,
      rolls/buns}              0.05663447   557
[3]  {whole milk,
      yogurt}                  0.05602440   551
[4]  {root vegetables,
      whole milk}              0.04890696   481
[5]  {root vegetables,
      other vegetables} 0.04738180   466
[6]  {other vegetables,
      yogurt}                  0.04341637   427
[7]  {other vegetables,
      rolls/buns}              0.04260295   419
[8]  {tropical fruit,
      whole milk}              0.04229792   416
[9]  {whole milk,
      soda}                    0.04006101   394
[10] {rolls/buns
```

Association Rules: Example in R

- Next, the list of frequent 2-itemsets is joined onto itself to form candidate 3-itemsets.
- For example $\{othervegetables, wholemilk\}$ and $\{wholemilk, rolls/buns\}$ would be joined as $\{othervegetables, wholemilk, rolls/buns\}$.
- The algorithm retains those itemsets that satisfy the minimum support.
- The following output shows that only two frequent 3-itemsets have been identified.
-

```
>itemsets <- apriori(Groceries, parameter=list(
    minlen=3, maxlen=3,
+ support=0.02, target="frequent itemsets"))
> summary(itemsets)
set of 2 itemsets
...
```

Association Rules: Example in R

```
> inspect(sort(itemsets, by ="support"))
    items                      support count
[1] {root vegetables,
     other vegetables,
     whole milk}            0.02318251   228
[2] {other vegetables,
     whole milk,
     yogurt}                0.02226741   219
```

Association Rules: Example in R

- In the next iteration, there is only one candidate 4-itemset $\{rootvegetables, othervegetables, wholemilk, yogurt\}$, and its support is below 0.02.
- No frequent 4-itemsets have been found, and the algorithm converges.

```
> itemsets <- apriori(Groceries, parameter=list(
    minlen=4, maxlen=4,
+ support=0.02, target="frequent itemsets"))
> summary(itemsets)
set of 0 itemsets
```

Association Rules: Example in `R`

- The previous steps simulate the *Apriori algorithm* at each iteration.
- For the Groceries dataset, the iterations run out of support when k = 4.
- Therefore, the frequent itemsets contain 59 frequent 1-itemsets, 61 frequent 2-itemsets, and 2 frequent 3-itemsets.
- When the `maxlen` parameter is not set, the algorithm continues each iteration until it runs out of support or until k reaches the default `maxlen`=10.

Association Rules: Example in R

```
> inspect(sort(itemsets, by ="support"))> itemsets <- apriori(Groceries, parameter=list(minlen =1, support=0.02,
+ target="frequent itemsets"))
> summary(itemsets)
set of 122 itemsets

most frequent items:
      whole milk other vegetables                yogurt
              28               20                    11
      rolls/buns             soda               (Other)
              10               10                   108

element (itemset/transaction) length distribution: sizes
 1  2  3
59 61  2
...
```

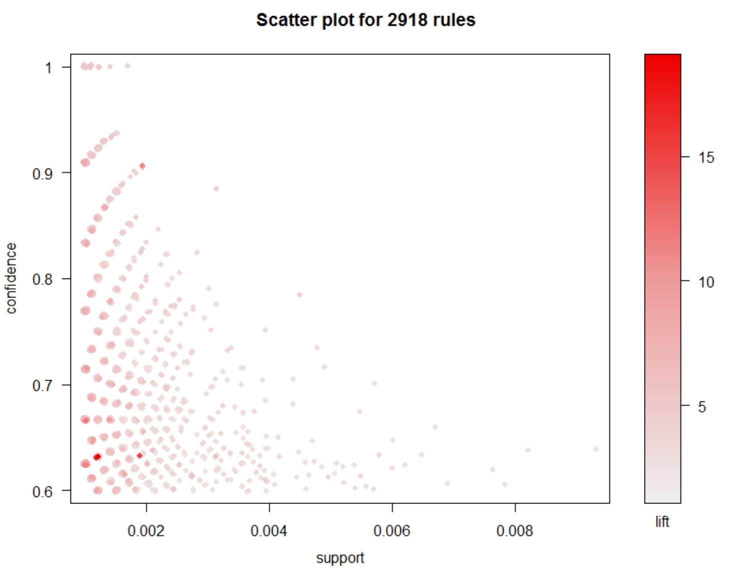
Association Rules: Example in R

- The apriori() function can also be used to generate rules.
- Assume that the minimum support threshold is now set to a lower value 0.001, and the minimum confidence threshold is set to 0.6.
- A lower minimum support threshold allows more rules to show up.
- The following code creates 2,918 rules from all the transactions in the Groceries dataset that satisfy both the minimum support and the minimum confidence.

```
rules <- apriori(Groceries, parameter=list(support
    =0.001,
confidence=0.6, target = "rules"))
```

- The command `plot(rules)` display the scatterplot of the 2,918 rules, where the horizontal axis is the support, the vertical axis is the confidence, and the shading is the lift.
- The scatterplot shows that, of the 2,918 rules generated from the Groceries dataset, the highest lift occurs at a low support and a low confidence.

Association Rules: Example in R



Association Rules: Example in R

- The inspect() function can display content of the rules generated previously.
- The following code shows the top three rules sorted by the lift. Rule $\{Instantfoodproducts, soda\} \rightarrow \{hamburgermeat\}$ has the highest lift of ≈ 19.
-

```
> inspect(head(sort(rules, by="lift"), 3))
    lhs                                         rhs
                                   support  confidence
        lift  count
[1] {Instant food products,
     soda}                            => {hamburger meat}
         0.001220132  0.6315789 18.99565     12
[2] {soda,
     popcorn}                         => {salty snack}
         0.001220132  0.6315789 16.69779     12
[3] {ham,
     processed cheese}                => {white bread}
         0.001931876  0.6333333 15.04549     19
```

Association Rules: Example in R

- We can also visualize the top three rules with the highest lift using the following code.
-

```
plot(highLiftRules, method="graph", control=list(
    alpha=1))
```

Association Rules: Example in R