Data Analytics with R

simplilearn

DATA AND
ARTIFICIAL INTELLIGENCE

Association Mining

Business Scenario

- Mike works at a restaurant and wants to understand customer preferences and choices to create combo meals to improve the customer experience.

Approach: To explore the different combinations, Mike needs to know about the association between itemsets and learn how to derive interesting patterns from the available data.

Learning Objectives

By the end of this lesson, you will be able to:

- Perform association mining
- Discuss Apriori algorithm
- Calculate the measures of association
- Explore analyzing and visualizing association rules
- Elaborate real-life applications of association mining

DATA AND
ARTIFICIAL INTELLIGENCE

Introduction to Association Mining

Association Rule Mining

Association rule mining helps to explore relationships between items and sets of items.

It also helps to enumerate interesting interactions of items that may include:

Apriori Algorithm

Apriori

Apriori is an algorithm for frequent itemset mining and association rule learning over transactional databases.

The Apriori algorithm calculates the rules that express probabilistic relationships between items in frequent itemsets.

Transaction Data

Apriori algorithm processes transactional data.

Transaction data has a one-to-many relationship between the transaction ID and the items of each transaction.

Every transaction includes a set of items, such as the items in a market basket.

Collection of items for a transaction is an attribute of that transaction.

Transaction Data

Transaction data may be stored and used in the traditional format with non-unique transaction IDs as rows and items as a single transaction.

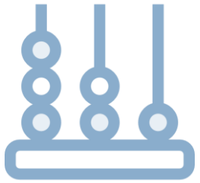
Rest_data

	Bill_Number	Item_Desc	Quantity
1	G0470111	OMELET BREAKFAST	1
2	G0470111	LEMONADE	1
3	G0470114	OMELET BREAKFAST	2
4	G0470114	CHICKEN BURGER	1
5	G0470114	MASALA CHAI	3
6	G0470116	LEMONADE	1
7	G0470116	OMELET BREAKFAST	1
8	G0470116	CHICKEN BURGER	1
9	G0470123	OMELET BREAKFAST	1
10	G0470123	CHICKEN BURGER	1
11	G0470128	OMELET BREAKFAST	1
12	G0470128	MINERAL WATER	1
13	G0470129	PASTA ARRABIATA	1
14	G0470129	OMELET BREAKFAST	1
15	G0470129	LEMONADE	2
16	G0470130	ROAST CHICKEN	1
17	G0470130	LEMONADE	2

Consider the example of transactional data from a restaurant.
The bill number is the transaction ID column and the description is the column of items. That is, the attributes of the corresponding transaction.

Transaction Data

Transaction data may also be stored in standard predictive model format with unique transaction IDs as rows and unique items of the data as a column.



Such data usually contains the count values of items for that transaction.

Rest_data

	MASALA CHAI	CHICKEN BURGER	LEMONADE	MINERAL WATER	OMELET BREAKFAST	PASTA ARRABIATA	ROAST CHICKEN
G0470111	0	0	1	0	1	0	0
G0470114	3	1	0	0	2	0	0
G0470116	0	1	1	0	1	0	0
G0470123	0	1	0	0	1	0	0
G0470128	0	0	0	1	1	1	0
G0470129	0	0	2	0	1	0	0
G0470130	0	0	2	0	0	0	1

Transaction Data

For Apriori processing in R, transaction data is stored as a sparse matrix in R, where the missing items in the transactions are represented as 0 and items present in the transaction are represented as 1.

Rest_data

	MASALA CHAI	CHICKEN BURGER	LEMONADE	MINERAL WATER	OMELET BREAKFAST	PASTA ARRABIATA	ROAST CHICKEN
G0470111	0	0	1	0	1	0	0
G0470114	3	1	0	0	2	0	0
G0470116	0	1	1	0	1	0	0
G0470123	0	1	0	0	1	0	0
G0470128	0	0	0	1	1	1	0
G0470129	0	0	2	0	1	0	0
G0470130	0	0	2	0	0	0	1

Apriori Basic Concepts

Association Rules

They are explicit mentions of relationships in the data that may be expressed by the “if-then” format.

Example:

In the association rule,

$X => Y$

represents the association as: if X, then Y

It has two parts:

- Antecedent, or the left side, represents the “if” part of the rule.
- Consequent, or the right side, represents the “then” part of the rule.

Measures of Association: Support

Support represents the proportion of items or itemset of interest among the entire data.

It is also defined as the probability of an item to be present, selected, or bought in any transaction.

$$\text{Support } \{X\} = \frac{\#\text{ of transactions with X}}{\#\text{ of total transactions}}$$

Measures of Association: Example for Support

Transaction Data

Bill_Number	Item1	Item2	Item3
G0470111	OMELET BREAKFAST	LEMONADE	
G0470114	OMELET BREAKFAST	CHICKEN BURGER	MASALA CHAI
G0470116	OMELET BREAKFAST	CHICKEN BURGER	LEMONADE
G0470123	OMELET BREAKFAST	CHICKEN BURGER	
G0470128	OMELET BREAKFAST	MINERAL WATER	
G0470129	PASTA ARRABIATA	OMELET BREAKFAST	LEMONADE
G0470130	ROAST CHICKEN	LEMONADE	

$$\text{Support}\{X\} = \frac{\#\text{ of transactions with X}}{\#\text{ of total transactions}}$$

$$\text{Support}\{\text{Omelet Breakfast}\} = \frac{6}{7} = 0.86$$

Measures of Association: Confidence

It defines the accuracy of a rule.

It is the proportion of transactions where consequent (Y) is true, given that antecedent (X) is true.

$$Confidence\ \{X \Rightarrow Y\} = \frac{\text{Support}\{X\ \boldsymbol{and}\ Y\}}{\text{Support}\{X\}}$$

Measures of Association: Example for Confidence

Transaction Data

Bill_Number	Item1	Item2	Item3
G0470111	OMELET BREAKFAST	LEMONADE	
G0470114	OMELET BREAKFAST	CHICKEN BURGER	MASALA CHAI
G0470116	OMELET BREAKFAST	CHICKEN BURGER	LEMONADE
G0470123	OMELET BREAKFAST	CHICKEN BURGER	
G0470128	OMELET BREAKFAST	MINERAL WATER	
G0470129	PASTA ARRABIATA	OMELET BREAKFAST	LEMONADE
G0470130	ROAST CHICKEN	LEMONADE	

$$Confidence\ \{X \Rightarrow Y\} = \frac{\text{Support}\{X\ \boldsymbol{and}\ Y\}}{\text{Support}\{X\}}$$

$$Confidence\ \{Chicken\ Burger \Rightarrow Omelette\ Breakfast\} = \frac{3/7}{3/7} = 1$$

Measures of Association: Lift

It is a measure of the times that the consequent (Y) is likely to occur when antecedent (X) is true compared to how often consequent (Y) occurs on its own.

It may be expressed as the confidence of the rule divided by the baseline confidence of the consequent (Y) alone.

$$Lift\ \{X \Rightarrow Y\} = \frac{Support\ \{X\ \&\ Y\}}{Support\{X\}\ *\ Support\{Y\}}$$

Or

$$\text{Lift}\ \{\text{X} \Rightarrow \text{Y}\} = \frac{\text{Confidence}\ \{\text{X} \Rightarrow \text{Y}\}}{\text{Support}\ \{\text{Y}\}}$$

Measures of Association: Example of Lift

Transaction Data

Bill_Number	Item1	Item2	Item3
G0470111	OMELET BREAKFAST	LEMONADE	
G0470114	OMELET BREAKFAST	CHICKEN BURGER	MASALA CHAI
G0470116	OMELET BREAKFAST	CHICKEN BURGER	LEMONADE
G0470123	OMELET BREAKFAST	CHICKEN BURGER	
G0470128	OMELET BREAKFAST	MINERAL WATER	
G0470129	PASTA ARRABIATA	OMELET BREAKFAST	LEMONADE
G0470130	ROAST CHICKEN	LEMONADE	

$$Lift\{X \Rightarrow Y\} = \frac{Support\{X \,\&\, Y\}}{Support\{X\} * Support\{Y\}}$$

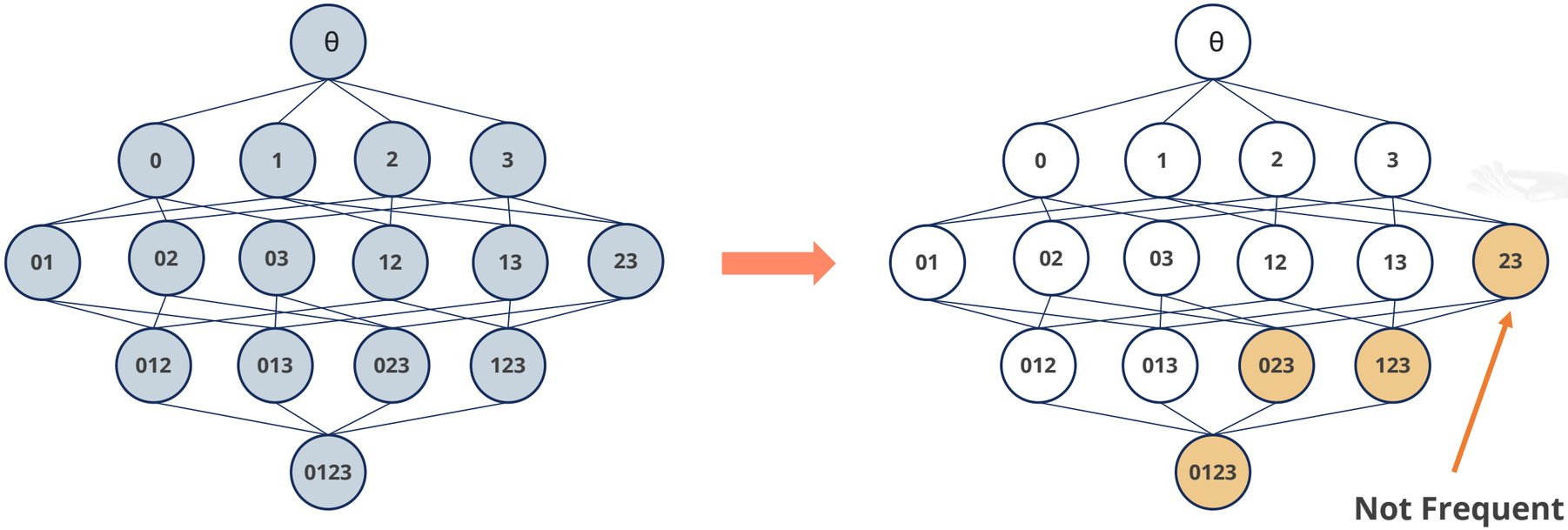
$$Lift\{Chicken\ Burger \Rightarrow Omelette\ Breakfast\} = \frac{3/7}{3/7 * 6/7} = 1.16$$

Working of Apriori Algorithm

Apriori Algorithm

Apriori algorithm helps to create frequent itemset and association rules in an efficient way.

The algorithm computes the measures of association of the rules generated.

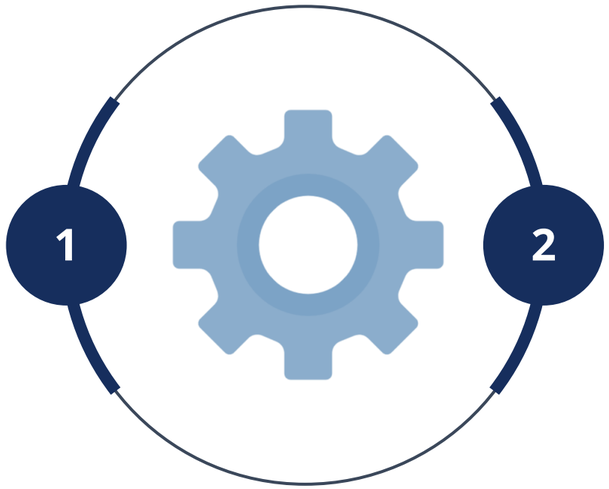


Generating Itemsets with Support-Based Pruning

Support-based pruning is the most straightforward way of computing frequent itemsets.

Steps to generate itemsets are:

1. Consider all possible itemsets from the data.



2. Discard the combinations with support value lower than a threshold.

Generating Itemsets with Support-Based Pruning

Considering a minimal support of 0.2, any itemset containing roast chicken or pasta can be considered as noninteresting.

Bill Number	Item 1	Item 2	Item 3
G0470111	LEMONADE	OMELET BREAKFAST	
G0470114	MASALA CHAI	CHICKEN BURGER	OMELET BREAKFAST
G0470116	CHICKEN BURGER	LEMONADE	OMELET BREAKFAST
G0470123	CHICKEN BURGER	OMELET BREAKFAST	
G0470128	OMELET BREAKFAST	MINERAL WATER	
G0470129	LEMONADE	OMELET BREAKFAST	PASTA ARRABIATA
G0470130	LEMONADE	ROAST CHICKEN	

However, the itemset {chicken burger and omelet} is more frequent at a minimal support of 0.6. It can be said that all items in the itemsets are also frequent.

Apriori uses these strategies to generate itemsets.

Generating Itemsets with Support-Based Pruning

Apriori generates rules by computing the possible association rules that have one item as a consequent.

By merging such rules with high confidence, it builds rules with two items as consequent and so on.

Generating Rules By Confidence-Based Pruning

Rest_data

Bill Number	Item Desc		
G0470111	LEMONADE	OMELET BREAKFAST	
G0470114	MASALA CHAI	CHICKEN BURGER	OMELET BREAKFAST
G0470116	CHICKEN BURGER	LEMONADE	OMELET BREAKFAST
G0470123	CHICKEN BURGER	OMELET BREAKFAST	
G0470128	OMELET BREAKFAST	MINERAL WATER	
G0470129	LEMONADE	OMELET BREAKFAST	PASTA ARRABIATA
G0470130	LEMONADE	ROAST CHICKEN	

Consider these rules:

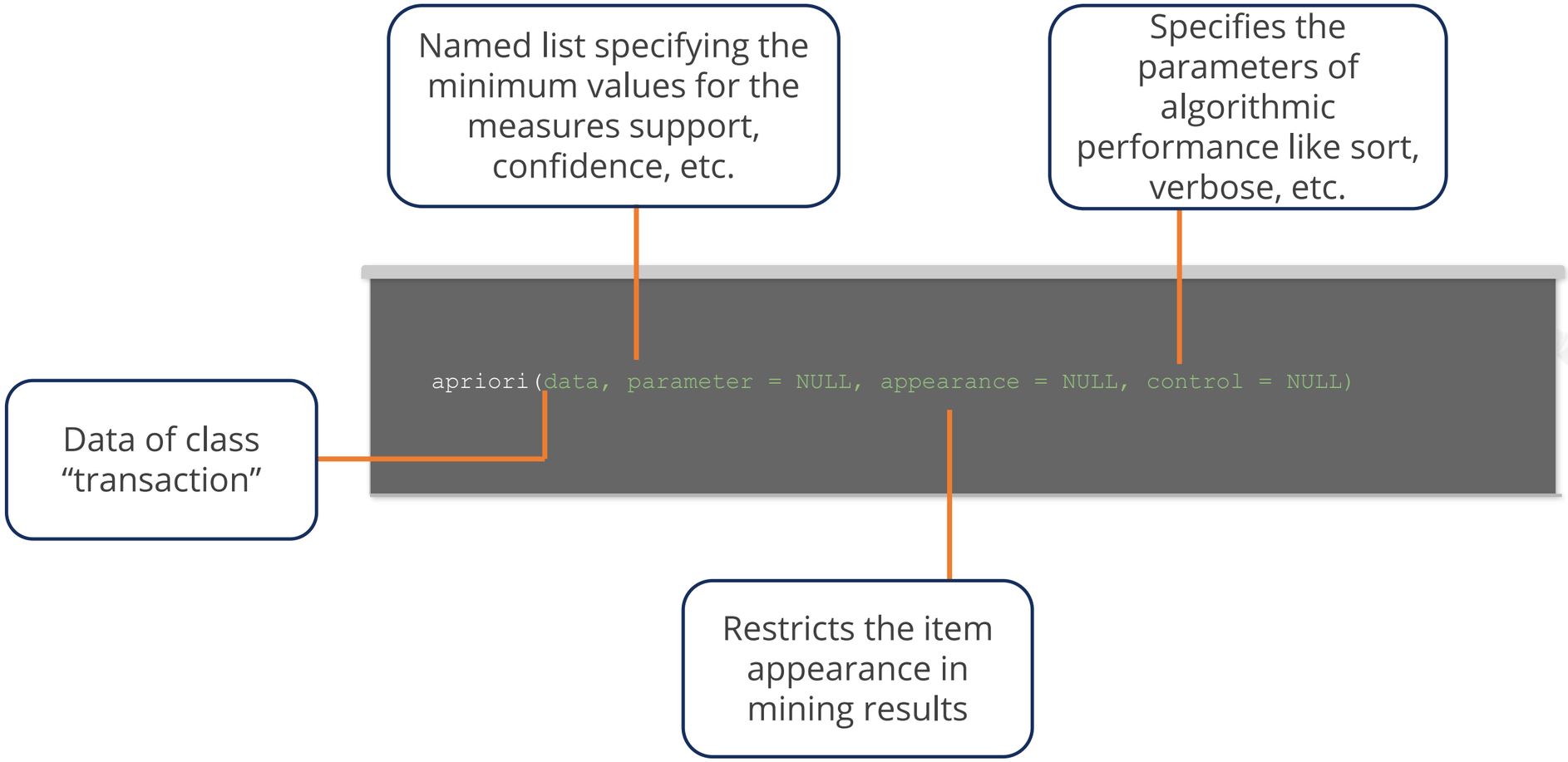
- {chicken burger, masala chai} => {omelet breakfast}
- {chicken burger, omelet breakfast} => {masala chai}

If the above rules are known to have high confidence, then this rule:

{chicken burger} => {masala chai, omelet breakfast} also has high confidence.

Apriori in R

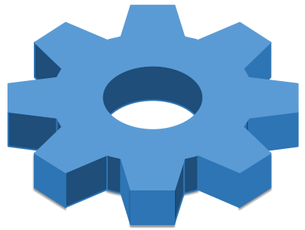
Apriori can be implemented in R using apriori() function of package rules.



Applications of Association Rules

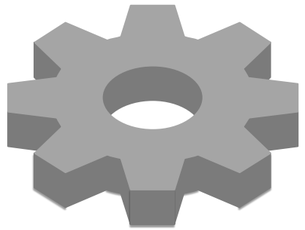
Applications of Association Mining

Association mining is employed for:

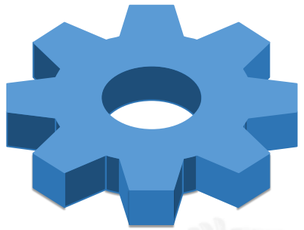


Market Basket Analysis

Medical Diagnosis



Tele-Communications



Banks and Insurance

Association Mining

Duration: 10 minutes

Problem Scenario: Consider the groceries data available in the “arules” package of R.

Explore the data and identify the most frequent item and itemset from the data.

Also, visualize the possible relationships between the items.

ASSISTED PRACTICE

Note: Please download the solution document from the **Course Resources** section and follow the steps given in the document

Key Takeaways

- The purpose of association rules mining is to enumerate interesting interactions of items.
- The interesting relationships can have two parameters: frequent itemsets and association rules.
- The measures of the strength of association rules are support, confidence, and lift.
- Apriori is an algorithm for frequent itemset mining and association rule learning over transactional databases.
- The Apriori algorithm includes two approaches: mining all frequent itemsets and generating rules from frequent itemsets.

Knowledge Check

Knowledge Check

1

What holds true for Apriori algorithm?

A. It mines all frequent patterns through pruning rules with lesser support.

B. It mines all frequent patterns through pruning rules with higher support.

C. Both a and b

D. None of these

Knowledge Check

1

What holds true for Apriori algorithm?

A. It mines all frequent patterns through pruning rules with lesser support.

B. It mines all frequent patterns through pruning rules with higher support.

C. Both a and b

D. None of these

The correct answer is **A**

It mines all frequent patterns through pruning rules with lesser support.

Knowledge Check

2

What does support{X} represent?

A. Total number of transactions containing X

B. Total number of transactions not containing X

C. Number of transactions containing X divided by the total number of transactions

D. Number of transactions not containing X divided by the total number of transactions

Knowledge Check

2

What does support{X} represent?

A. Total number of transactions containing X

B. Total number of transactions not containing X

C. Number of transactions containing X divided by the total number of transactions

D. Number of transactions not containing X divided by the total number of transactions

The correct answer is **C**

Support{X} represents the number of transactions containing X divided by the total number of transactions.

Knowledge Check

3

What is the principle on which Apriori algorithm works?

A. If an item is less frequent, then any itemset containing the item is also less frequent.

B. When two rules with one consequent have high confidence, then the rule created by merging them will also have high confidence.

C. Both A and B

D. None of these

Knowledge Check

3

What is the principle on which Apriori algorithm works?

A. If an item is less frequent, then any itemset containing the item is also less frequent.

B. When two rules with one consequent have high confidence, then the rule created by merging them will also have high confidence.

C. Both A and B

D. None of these

The correct answer is **C**

If an item is less frequent, then any itemset containing the item is also less frequent. When two rules with one consequent have high confidence, the rule created by merging them will also have high confidence.