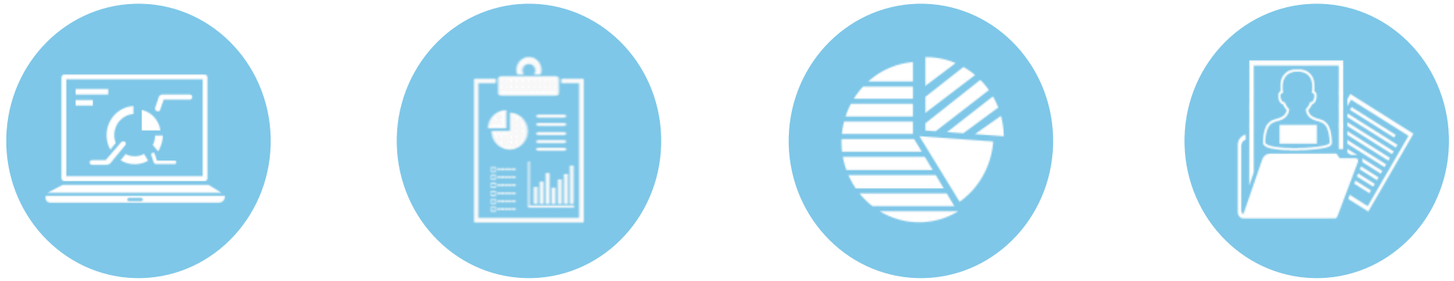
Data Science with R

Lesson 10—Association



Learning Objectives

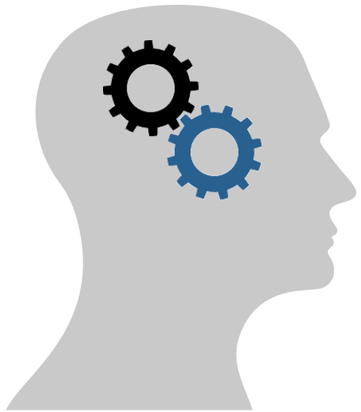
- Explain association rule
- Discuss the Apriori algorithm and the steps to apply it

Association

Topic 1—Association Rule

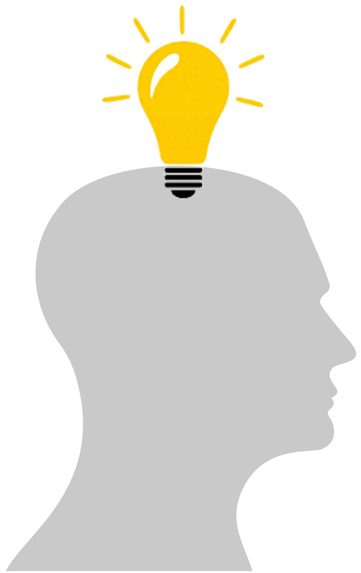
The Classic Anecdote of Beer and Diaper

In one of the researches conducted by a supermarket in the US, it was found that young men who visited the store on Fridays to buy diapers had a tendency to grab a bottle of beer too.



How did the supermarket arrive at this conclusion?

The Classic Anecdote of Beer and Diaper



The store collected data using the barcode scanners during the payment and stored the data in a database. A single record lists all the items purchased by a customer that was later analyzed to understand the trend.

The technique used is called “Market Basket Analysis” better known as “Association Rule.”

Association Rule

An Association rule is a classical data mining technique that finds **interesting patterns or relations** in a dataset.

The relation between the order of an item and the frequency of its occurrence is known as Interesting Relation.

Association Rule

MATHEMATICAL REPRESENTATION

An association rule is a pattern that states when X occurs, Y occurs with a certain probability.

$$X \Rightarrow Y$$

Where, $X, Y \subset I$, and $X \cap Y = \varnothing$

Association rule is not suitable for numeric data and assumes all data elements to be categorical.

Association Rule

CASE STUDY: WALMART

Problem Statement

Data Study

Observations and Conclusions

Thousands of customers visit Walmart every day and transactions with distinct combinations of products are recorded.

The Regional Sales Director wants to conduct a study of the transactions data to plan a business strategy based on the customer behavior.

Association Rule

CASE STUDY: WALMART

Problem statement

Data Study

Observations and Conclusions

To begin with, divide the transactions into two categories: Groceries and Apparels.

	ID	Product 1	Product 2	Number of transactions in a month
Groceries	T1	Milk	Bread	523,457
	T2	Milk	-	2,461
Apparels	T3	Jeans	Shoes	198
	T4	Jeans	-	29,846

Association Rule

CASE STUDY: WALMART

Problem statement

Data Study

Observations and Conclusions

After studying the data, following observations were noted in the groceries business unit:

- T2 Milk is bought as a single product in 2,461 transactions in the month of May 2018.
- T1 Milk is bought along with bread in 523,457 transactions in the month of May 2018.

Association Rule

CASE STUDY: WALMART

Problem statement

Data Study

Observations and Conclusions

Conclusions from the study for the Groceries Business Unit:

- When customers buy milk, they also buy bread along with it.
- Total transactions where milk is present = 523,457 + 2461 = 525,918
- Total transactions where milk and bread are present = 523,457
- The association or probability of customers buying bread given the fact that they buy milk is = 523,457/525,918 = 99%

Association Rule

CASE STUDY: WALMART

Problem statement

Data Study

Observations and Conclusions

After studying the data, following observations were noted in the apparels business unit:

- T4 jeans is bought as a single product in 29,846 transactions in the month of May 2018.
- T3 jeans is bought along with formal shoes in 198 transactions in the month of May 2018.

Association Rule

CASE STUDY: WALMART

Problem statement

Data Study

Observations and Conclusions

Conclusions from the study for the Apparels Business Unit:

- There is no association between jeans and formal shoes.
- Total transactions in which jeans were present = 29,846 + 198 = 30,044
- Total transactions in which jeans and formal shoes were present = 198
- The association or probability of customers buying formal shoes given the fact that they buy jeans is = 198/30,44 = 1%

Measures of Association Rule

The association rules for a set of observations can be large and can vary at a few instances.

It is crucial to find the rules that are useful to the users and can be measured **subjectively and objectively.**

Measures of Association Rule

SUBJECTIVE MEASURES

Subjective measures are more oriented toward the user. Unexpectedness and actionability are the two parameters of subjective measures.

- Unexpectedness states that rules are only useful if they are previously unknown to the user or contradict the user's knowledge.
- Actionability states that rules are only useful if they can be acted upon with some advantage.

Subjective measures are sometimes difficult to determine and varies on a case to case basis. Due to this, objective measures are preferred.

https://www.researchgate.net/publication/233754781_Support_vs_Confidence_in_Association_Rule_Algorithms

Measures of Association Rule

OBJECTIVE MEASURES

They involve the following statistical analysis of the data:

Support

- Represents the frequency of an item in a dataset.
- It holds true with support **sup** in T, if sup% of transactions contain X ∪ Y.

  $sup = Pr(X \cup Y)$

Confidence

- The confidence for the rule {X} ➞ {Y} is defined as support({X, Y})/support({Y}).
- It holds true in T with confidence **conf** if conf% of transactions that contain X also contains Y.

  $conf = Pr(Y \mid X)$

Measures of Association Rule

OBJECTIVE MEASURES: EXAMPLE

Transaction number	Items
0	soy milk, lettuce
1	lettuce, diapers, wine, chard
2	soy milk, diapers, wine, orange juice
3	lettuce, soy milk, diapers, wine
4	lettuce, soy milk, diapers, orange juice

In the “Items” table, the support of {soy milk} is 4/5 and of {soymilk, diapers} is 3/5.

In the “Items” table, the confidence for diapers → wine is 3/5/4/5 = 3/4 = 0.75.

Measures of Association Rule

OBJECTIVE MEASURES: LIMITATIONS

While support and confidence can help you quantify the success of association analysis for thousands of sale items, the process of finding them can be really slow as the item list grows.

In such cases, **Apriori algorithm** is used.

Association

Topic 2—Apriori Algorithm

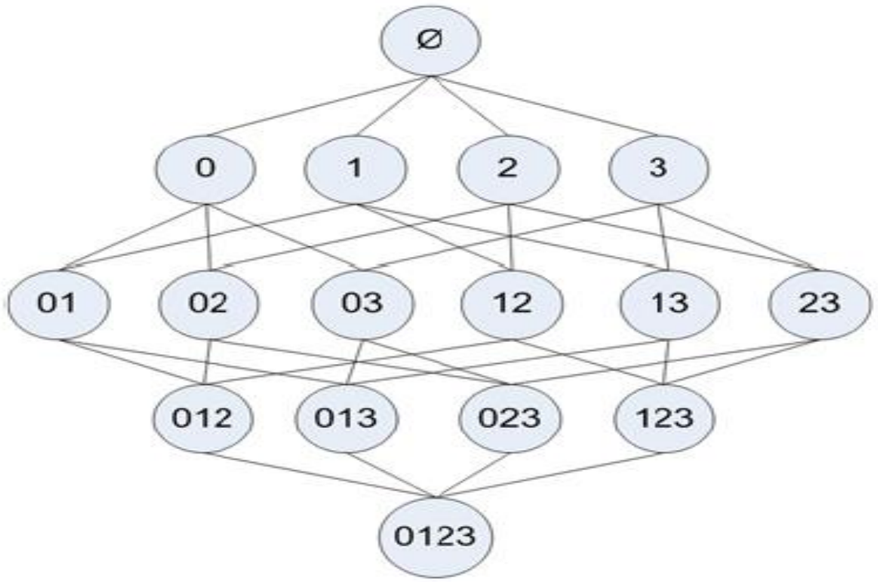
Apriori Algorithm

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

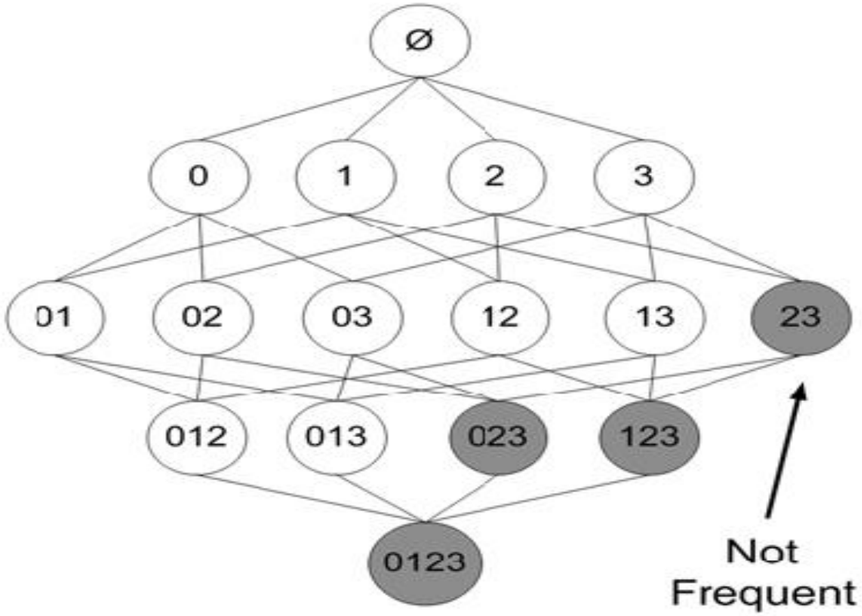
Apriori Algorithm

USES

It helps in reducing the number of possible interesting item sets by identifying the non frequent ones.



All possible item sets from the set {1, 2, 3}



(Assumes all subsets of a frequent item set are frequent) – Identifies the non frequent item sets

Applying Apriori Algorithm

To apply Apriori algorithm, there are two steps:

1. Mine all frequent item sets
2. Generate association rules from frequent item sets

Applying Apriori Algorithm

Mine all frequent item sets

Generate Association rules from frequent item sets

A frequent item set is any subset of frequent item set and has sup ≥ minsup.

Example: For the following dataset:

t1:	Beef, Chicken, Milk
t2:	Beef, Cheese
t3:	Cheese, Boots
t4:	Beef, Chicken, Cheese
t5:	Beef, Chicken, Clothes, Cheese, Milk
t6:	Chicken, Clothes, Milk
t7:	Chicken, Milk, Clothes

Assume: minsup = 30% and minconf = 80%

Frequent item set: {Chicken, Clothes, Milk} [sup = 3/7]

Applying Apriori Algorithm

STEPS

Mine all frequent item sets

Generate Association rules from frequent item sets

1. Find all 1-item frequent item sets; then all 2-item frequent item sets, and so on
2. In each iteration k, consider item sets that contain some k-1 frequent item sets **(Candidate Itemset Generation)**
3. Find frequent item sets of size 1: F1

Applying Apriori Algorithm

CANDIDATE ITEMSET GENERATION

Mine all frequent item sets

Generate Association rules from frequent item sets

The candidate itemset generation takes F_{k-1} and returns candidates as the superset of the set of all frequent k item sets using the candidate-gen function.

It includes the following two steps:

1. **Join:** Generate all possible candidate item sets C_k of length k
2. **Prune:** Remove the candidates in C_k that cannot be frequent

Applying Apriori Algorithm

CANDIDATE ITEMSET GENERATION: ALGORITHM

Mine all frequent item sets

Generate Association rules from frequent item sets

```
Function candidate-gen(F_{k-1})
  C_k ← ∅;
  forall f_1, f_2 ∈ F_{k-1}
      with f_1 = {i_1, … , i_{k-2}, i_{k-1}}
      and f_2 = {i_1, … , i_{k-2}, i'_{k-1}}
      and i_{k-1} < i'_{k-1} do
    c ← {i_1, …, i_{k-1}, i'_{k-1}};            // join f_1 and f_2
    C_k ← C_k ∪ {c};
    for each (k-1)-subset s of c do
      if (s ∉ F_{k-1}) then
        delete c from C_k;          // prune
    end
  end
  return C_k;
```

Applying Apriori Algorithm

EXAMPLE

Mine all frequent item sets

Generate Association rules from frequent item sets

Consider the following dataset T with minsup = 0.5:

TID	Items
T100	1, 3, 4
T200	2, 3, 5
T300	1, 2, 3, 5
T400	2, 5

Calculating the frequent itemsets

itemset:count

1. scan T ➔ C_1: {1}:2, {2}:3, {3}:3, {4}:1, {5}:3

   ➔ F_1: {1}:2, {2}:3, {3}:3, {5}:3

   ➔ C_2: {1,2}, {1,3}, {1,5}, {2,3}, {2,5}, {3,5}

2. scan T ➔ C_2: {1,2}:1, {1,3}:2, {1,5}:1, {2,3}:2, {2,5}:3, {3,5}:2

   ➔ F_2: **{1,3}**:2, **{2,3}**:2, **{2,5}**:3, **{3,5}**:2

   ➔ C_3: {2, 3,5}

3. scan T ➔ C_3: **{2, 3, 5}**:2 ➔ F_3: **{2, 3, 5}**

Here the items are sorted in a Lexicographic order (refers to the increasing numerical order). For example: The permutations of Lexicographic order of [1, 2, 3], are 123, 132, 213, 231, 312, and 321.

Applying Apriori Algorithm

Mine all frequent item sets

Generate Association rules from frequent item sets

For each frequent item set X and proper non empty subset A of X, assume B = X – A.

$A \rightarrow B$ is an association rule if:

$Confidence(A \rightarrow B) \geq minconf$
$support(A \rightarrow B) = support(A \cup B) = support(X)$
$confidence(A \rightarrow B) = support(A \cup B) / support(A)$

Applying Apriori Algorithm

EXAMPLE 1

Mine all frequent item sets

Generate Association rules from frequent item sets

t1:	Beef, Chicken, Milk
t2:	Beef, Cheese
t3:	Cheese, Boots
t4:	Beef, Chicken, Cheese
t5:	Beef, Chicken, Clothes, Cheese, Milk
t6:	Chicken, Clothes, Milk
t7:	Chicken, Milk, Clothes

For the dataset given above;

Association rules from the item set:

Clothes → Milk, Chicken [sup = 3/7, conf = 3/3]

... ...

Clothes, Chicken → Milk [sup = 3/7, conf = 3/3]

Applying Apriori Algorithm

EXAMPLE 2

Mine all frequent item sets

Generate Association rules from frequent item sets

Problem statement

Solution

If {2,3,4} is frequent with sup = 50% and proper nonempty subsets: {2,3}, {2,4}, {3,4}, {2}, {3}, {4}, with sup = 50%, 50%, 75%, 75%, 75%, 75%, respectively, find the association rule.

Applying Apriori Algorithm

EXAMPLE 2

Mine all frequent item sets

Generate Association rules from frequent item sets

Problem statement

Solution

Association rules:

2,3 → 4, confidence = 100%
2,4 → 3, confidence = 100%
3,4 → 2, confidence = 67%
2 → 3,4, confidence = 67%
3 → 2,4, confidence = 67%
4 → 2,3, confidence = 67%
Support of all rules = 50%

Key Takeaways

- Association rule mining finds interesting patterns in a dataset.
- The interesting relationships can have two parameters: frequent item sets and association rules.
- An association rule is a pattern that states when X occurs, Y occurs with a certain probability.
- The measures of the strength of association rules are support and confidence.
- Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases.
- The Apriori algorithm includes two steps: mining all frequent item sets and generating rules from frequent item sets.