

# Association Rule

# Agenda

---

---

1. Apriori Algorithm

---

2. Market Basket Analysis

---

# Association Rule mining

Association rule mining is a method for discovering interesting relations between variables in the dataset.

Pattern that states when an event occurs, one more event occurs with a certain probability in parallel.

Customers who purchase a laptops have 55% likelihood of also purchasing a mouse for their laptops.



# Association Rule mining – Parameters

**Support:** Gives fraction of transactions which contains the item X and Y.

**Lift:** Lift indicates the strength of the rule over the random co-occurrence of X and Y

**Confidence:** Gives how often the items X and Y occurs together, given no. of times X occurs.



$$\begin{array}{l} \text{Rule: } X \Rightarrow Y \\ \begin{array}{l} \nearrow \text{Support} = \frac{\text{freq}(X, Y)}{N} \\ \rightarrow \text{Confidence} = \frac{\text{freq}(X, Y)}{\text{freq}(X)} \\ \searrow \text{Lift} = \frac{\text{Support}}{\text{Supp}(X) \times \text{Supp}(Y)} \end{array} \end{array}$$



# Association Rule mining – An Example

Lets take an example

Suppose we have five transactions i.e. T1, T2, T3, T4 and T5

T1: L, M, N

T2: L, N, P

T3: M, N, P

T4: L, M, Q

T5: M, N, Q

Here

L, M, N, P, Q are the items in the store,  $I = \{L, M, N, P, Q\}$

T1, T2, T3, T4, T5 are the transactions,  $T = \{T1, T2, T3, T4, T5\}$



# Association Rule mining – An Example

Suppose, you made some association rules using our dataset

$L \rightarrow P$

$N \rightarrow L$

$L \rightarrow N$

$M \ \& \ N \rightarrow P$

Now we can find support, confidence and lift for these rules using the formula explained earlier:

Rule	Support	Confidence	Lift
$L \rightarrow P$	1/5	1/3	1/6
$N \rightarrow L$	2/5	2/4	2/12
$L \rightarrow N$	2/5	2/3	2/12
$M \ \& \ N \rightarrow P$	1/5	1/3	1/6

# Rule-based machine learning

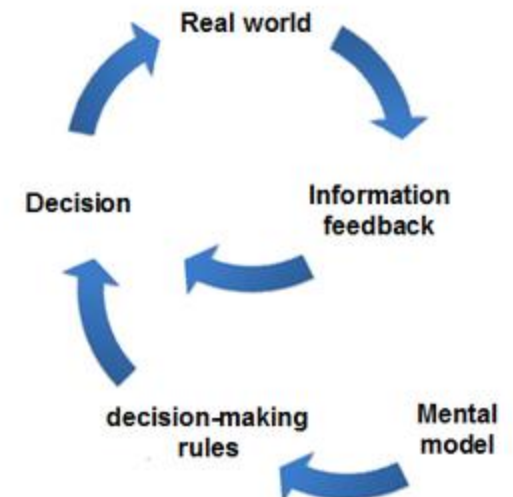
❖ **Rule-based machine learning (RBML)** is a model to circumscribe any machine learning method that identifies, learns and evolves the rules to store, manipulate and apply.

Algorithm to support RBML: “**Association Rule Mining**”

❖ **Association Rule Mining** is about discovering relations between variables in large datasets. Conditional Statements like IF: Then Statement is primarily used in facilitating this relationship findings.

❖ **Applications:**

1. Data Analysis for direct marketing
2. Catalog Design
3. Point of Sale (POS Product placement)





# Association Rules

Most Widely used algorithms for association rules are the following:

1. Apriori Algorithm
2. Market Basket Analysis

**Apriori Algorithm:** Uses a breadth-first search strategy to count the support of item-sets and uses a candidate generation function which exploits the downward closure property of support.

- ❖ It uses knowledge from previous iteration phase to produce frequent item sets.
- ❖ occurrence frequency for each candidate itemset is counted
- ❖ Those candidate itemsets that have higher frequency than minimum support threshold are qualified to be frequent itemsets
- ❖ A subset of a frequent itemset must also be a frequent itemset.

i.e., if  $\{AB\}$  is a frequent itemset, both  $\{A\}$  and  $\{B\}$  should be a frequent itemset.

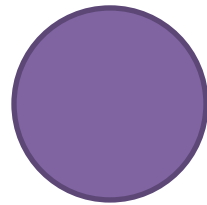
*Iteratively find frequent itemsets with cardinality from 1 to k (k-itemset)*

- ❖ Use the frequent itemsets to generate association rules

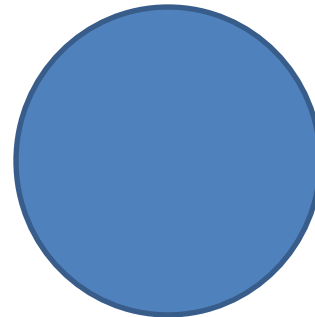
# Apriori Algorithm

Uses frequent itemsets to generate association rules,

"A subset of a frequent itemset must also be a frequent itemset"



Frequent Itemset



Frequent Itemset

# The Apriori Algorithm : Pseudo code

**Join Step:**  $C_k$  is generated by joining  $L_{k-1}$  with itself

**Prune Step:** Any  $(k-1)$ -itemset that is not frequent cannot be a subset of a frequent  $k$ -itemset.

**Pseudo-code:**

$C_k$ : Candidate itemset of size  $k$

$L_k$  : frequent itemset of size  $k$

$L_1 = \{\text{frequent items}\};$

**for**( $k = 1; L_k \neq \emptyset; k++$ ) **do begin**

$C_{k+1}$  = candidates generated from  $L_k$ ;

**for each** transaction  $t$  in database **do**

    increment the count of all candidates in  $C_{k+1}$  that are contained in  $t$

$L_{k+1}$  = candidates in  $C_{k+1}$  with  $\text{min\_support}$

**end**

**return**  $\bigcup_k L_k$ ;


# The Apriori Algorithm: Example

- Consider a database DB, consisting of 12 transactions.
- Suppose min. support count required is 3 (i.e.  $\text{min\_sup} = 3/12 = 25\%$ )
- Let minimum confidence required is 70%.
- Step 1 is to find out the frequent itemset using **Apriori algorithm**.
- Step 2 is to generate **association rules** using min. support & min. confidence.

Batch. No.	List of Items
PSD507	L1, L5, L3
PSD507	L2,L1,L8,L5
PSD507	L6,L5,L4,L3
PSD507	L3,L2,L1,L4,L6
PSD507	L4,L7,L8,L1,L6
PSD507	L3,L6,L5,L4
PSD507	L2,L4,L6
PSD507	L3,L4,L7
PSD507	L1,L7,L2,L5,L8
PSD507	L2,L7,L8,L6
PSD507	L6, L4,L2,L3
PSD507	L1,L4,L2,L6


## Step 1: Generating 1-itemset Frequent Pattern

Scan DB for count of each candidate



Itemset	Sup.count
{L1}	6
{L2}	7
{L3}	6
{L4}	7
{L5}	5
{L6}	7
{L7}	5
{L8}	4

Compare min sup count with min sup. count



Itemset	Sup.count
{L1}	6
{L2}	7
{L3}	6
{L4}	7
{L5}	5
{L6}	7
{L7}	5
{L8}	3

- The set of frequent 1-itemsets, L1 , consists of the candidate 1- itemsets satisfying minimum support.
- In the first iteration of the algorithm, each item is a member of the set of candidate.

## Step 2: Generating 2-itemset Frequent Pattern

Generate C2  
from L1

Itemset	Itemset
{L1,L2,L3}	{L2,L3,L4}
{L1,L2,L4}	{L5,L3,L4}
{L1,L2,L5}	{L6,L3,L4}
{L1,L2,L6}	{L7,L3,L4}
{L1,L2,L7}	{L8,L3,L4}
{L1,L2,L8}	{L5,L4,L6}
{L1,L3,L4}	{L5,L4,L6}
{L1,L3,L5}	{L5,L7,L6}
{L1,L3,L6}	{L5,L8,L6}
{L1,L3,L7}	{L4,L8,L6}
{L1,L3,L8}	{L4,L5,L3}
{L1,L4,L5}	{L4,L3,L6}
{L1,L4,L6}	{L5,L3,L6}
{L1,L4,L7}	{L2,L6,L4}
{L1,L4,L8}	{L1,L8,L5}
{L1,L5,L6}	{L2,L8,L5}
{L1,L5,L7}	{L2,L1,L5}
{L1,L5,L8}	
{L1,L6,L7}	
{L1,L6,L8}	
{L1,L7,L8}	

Scan DB for  
count of each C

Itemset	Sub.count
{L1,L2,L3}	1
{L1,L2,L4}	1
{L1,L2,L5}	2
{L1,L2,L6}	1
{L1,L2,L7}	1
{L1,L2,L8}	3
{L1,L3,L4}	0
{L1,L3,L5}	1
{L1,L3,L6}	1
{L1,L3,L7}	0
{L1,L3,L8}	0
{L1,L4,L5}	0
{L1,L4,L6}	1
{L1,L4,L7}	0
{L1,L4,L8}	0
{L1,L5,L6}	0
{L1,L5,L7}	1
{L1,L5,L8}	2
{L1,L6,L7}	0
{L1,L6,L8}	0
{L1,L7,L8}	3

Itemset	Sub.count
{L2,L3,L4}	2
{L5,L3,L4}	1
{L6,L3,L4}	1
{L7,L3,L4}	1
{L8,L3,L4}	0
{L5,L4,L6}	2
{L5,L7,L6}	0
{L5,L8,L6}	0
{L4,L8,L6}	1
{L4,L5,L3}	2
{L4,L3,L6}	3
{L5,L3,L6}	2
{L2,L6,L4}	1
{L1,L8,L5}	2
{L2,L8,L5}	2
{L2,L1,L5}	2

## Step 2: Generating 2-itemset Frequent Pattern

Itemset	Sub.count
{L1,L2,L3}	1
{L1,L2,L4}	1
{L1,L2,L5}	2
{L1,L2,L6}	1
{L1,L2,L7}	1
{L1,L2,L8}	3
{L1,L3,L4}	0
{L1,L3,L5}	1
{L1,L3,L6}	1
{L1,L3,L7}	0
{L1,L3,L8}	0
{L1,L4,L5}	0
{L1,L4,L6}	1
{L1,L4,L7}	0
{L1,L4,L8}	0
{L1,L5,L6}	0
{L1,L5,L7}	1
{L1,L5,L8}	0
{L1,L6,L7}	0
{L1,L6,L8}	0
{L1,L7,L8}	3

Itemset	Sub.count
{L2,L3,L4}	2
{L5,L3,L4}	1
{L6,L3,L4}	1
{L7,L3,L4}	1
{L8,L3,L4}	0
{L5,L4,L6}	2
{L5,L7,L6}	0
{L5,L8,L6}	0
{L4,L8,L6}	1
{L4,L5,L3}	2
{L4,L3,L6}	3
{L5,L3,L6}	2
{L2,L6,L4}	1
{L1,L8,L5}	2
{L2,L8,L5}	2
{L2,L1,L5}	2



Compare min  
sup count with  
min sup. count

Itemset	Sub.Count
{L1,L2,L5}	2
{L1,L2,L8}	3
{L1,L7,L8}	3
{L2,L3,L4}	2
{L5,L4,L6}	2
{L4,L5,L3}	2
{L4,L3,L6}	3
{L5,L3,L6}	2
{L1,L8,L5}	2
{L2,L8,L5}	2
{L2,L1,L5}	2

## Step 2: Generating 3-itemset Frequent Pattern

- ✓ To discover the set of frequent 3-itemsets,  $L_2$ , the algorithm uses  $L_1 \text{ Join } L_1$  to generate a candidate set of 3-itemsets,  $C_2$ .
- ✓ The batch nos. in DB are scanned and the support count for each candidate itemset in  $C_2$  is accumulated.
- ✓ The set of frequent 3-itemsets,  $L_2$ , is determined.



## Step 3: Generating 4-itemset Frequent Pattern

Scan DB for  
count of each C



Itemset
{L6,L5,L4,L3}
{L2,L1,L4,L6}
{L2,L1,L8,L5}

**C3**

Scan DB for  
count of each C



Itemset	Sub.count
{L6,L5,L4,L3}	2
{L2,L1,L4,L6}	2
{L2,L1,L8,L5}	2

**C3**

Compare min  
sup count with  
min sup. count



Itemset	Sub.count
{L6,L5,L4,L3}	2
{L2,L1,L4,L6}	2
{L2,L1,L8,L5}	2

**L3**

➤ The generation of the set of candidate 4-itemsets, C3 , involves use of the apriori Property.

❑ In order to find C3, we compute L2 Join L2.

❑ Next step is to prune, i.e. reducing the size of C3. Prune step helps to avoid heavy computation due to large Ck.

## Step 3: Generating 4-itemset Frequent Pattern

- Apriori principle also states that, that all subsets of a frequent itemset must also be frequent.
- For example , lets take  $\{L6, L5, L4, L3\}$ , The 3-item subsets of it are  $\{L6, L5, L4\}$ ,  $\{L6, L4, L3\}$ ,  $\{L6, L3, L5\}$ ,  $\{L5, L4, L3\}$ .

Since all 3-item subsets of  $\{L6, L5, L4, L3\}$  are members of  $L2$ , We will keep  $\{L6, L5, L4, L3\}$  in  $C3$ .

- Lets take another example of  $\{L7, L6, L8, L1\}$ , which shows how the pruning is performed. The 3-item subsets  $\{L7, L6, L8\}$ ,  $\{L7, L8, L1\}$ ,  $\{L1, L8, L6\}$ ,  $\{L1, L7, L6\}$ . But,  $\{L7, L6, L1\}$  is not a member of  $L2$  and hence it is not frequent itemset and hence violating apriori principle. So, We will have to remove  $\{L7, L6, L8, L1\}$  from  $C3$ .

- Therefore,  $C3 = \{\{L6, L5, L4, L3\}, \{L2, L1, L4, L6\}, \{L2, L1, L8, L5\}\}$  after checking for all members of result of Join operation for Pruning.
- Now, the batch no.s in DB are scanned in order to determine  $L3$ , consisting of those candidates 4-itemsets in  $C3$  having minimum support.

## Step 4: Generating 5-itemset Frequent Pattern

- The algorithm uses L3 Join L3 to generate a candidate set of 5-itemsets, C4. Although the join results in:

Itemset
{L6,L5,L4,L3,L1}
{L2,L1,L4,L6,L8}

**C4**

these itemsets are pruned since their subsets are not frequent.

- Thus,  $C4 = \phi$ , and algorithm terminates, having found all of the frequent items. This completes our Apriori Algorithm.

### ➤ Application of Apriori Rule:

These frequent itemsets will be used to generate strong association rules ( where strong association rules satisfy both minimum support & minimum confidence)

## Apriori's Efficiency: Tips

- Hash-based itemset counting: A k-itemset whose corresponding hashing bucket count is below the threshold cannot be frequent.
- Transaction reduction: A transaction that does not contain any frequent k-itemset is useless in subsequent scans.
- Partitioning: Any itemset that is potentially frequent in DB must be frequent in at least one of the partitions of DB.
- Sampling: mining on a subset of given data, lower support threshold + a method to determine the completeness.
- Dynamic itemset counting: add new candidate itemsets only when all of their subsets are estimated to be frequent.

# Market Basket Analysis

Market Basket Analysis is a mathematical modeling technique based upon the assumption that buying certain group of items would lead to buy another associated group of items.

## Applications:

- Analyze the customer purchasing behavior and helps in increasing the sales and maintain inventory by focusing on the point of sale transaction data.
- Customized emails with add-on sales.
- Catalogue design
- Identifying items in trend

## Example:

### ❑ Frequent itemset:

{Diaper, Beer}

### ❑ Association rule:

{Diaper} → {Beer}

Id	Items
1	{Milk, Bread}
2	{Bread, <b>Diaper, Beer</b> , Eggs}
3	{Bread, Milk, <b>Diaper, Beer</b> }
4	{ <b>Beer</b> , Milk, <b>Diaper</b> , Cola}

# Market Basket Analysis

**Transaction** is a set of items (Itemset).

**Confidence** : It is the measure of uncertainty or trust worthiness associated with each discovered pattern.

**Support** : It is the measure of how often the collection of items in an association occur together as percentage of all transactions

**Frequent itemset** : If an itemset satisfies minimum support, then it is a frequent itemset.

**Strong Association rules**: Rules that satisfy both a minimum support threshold and a minimum confidence threshold.

# Association Rule Application on “Baskets”

- Items purchased on a credit card, for example, Laptop, Optical drive can give idea of the next product that customer may buy.
- Optional services purchased by telecommunications customers (call waiting, call forwarding, DSL, speed call, and so on) help to determine how to bundle these services together to maximize revenue.
- Banking products used by retail customers (money market accounts, certificate of deposit, investment services, car loans, and so on) identify customers likely to want other products.
- Unusual combinations of insurance claims can be a sign of fraud and can spark further investigation.
- Medical patient histories can give indications of likely complications based on certain combinations of treatments.

# Apriori Code - Dataset

Index	0	1	2	3	4
0	shrimp	almonds	avocado	vegetables mix	green grapes
1	burgers	meatballs	eggs	nan	nan
2	chutney	nan	nan	nan	nan
3	turkey	avocado	nan	nan	nan
4	mineral water	milk	energy bar	whole wheat rice	green tea
5	low fat yogurt	nan	nan	nan	nan
6	whole wheat pasta	french fries	nan	nan	nan
7	soup	light cream	shallot	nan	nan
8	frozen vegetables	spaghetti	green tea	nan	nan
9	french fries	nan	nan	nan	nan
10	eggs	pet food	nan	nan	nan
11	cookies	nan	nan	nan	nan
12	turkey	burgers	mineral water	eggs	cooking oil
13	spaghetti	champagne	cookies	nan	nan
14	mineral water	salmon	nan	nan	nan
15	mineral water	nan	nan	nan	nan
16	shrimp	chocolate	chicken	honey	oil
17	turkey	eggs	nan	nan	nan
18	turkey	fresh tuna	tomatoes	spaghetti	mineral water



# Apriori Code

```
# Apriori

# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

import os
os.chdir('Give path')
# Data Preprocessing
dataset = pd.read_csv('Market_Basket_Optimisation.csv', header = None)
transactions = []
for i in range(0, 7501):
    transactions.append([str(dataset.values[i,j]) for j in range(0, 20)])

# Training Apriori on the dataset
from apyori import apriori
rules = apriori(transactions, min_support = 0.003, min_confidence = 0.2, min_lift = 3, min_length = 2)

# Visualising the results
results = list(rules)
```

```
RelationRecord(items=frozenset({'chicken', 'light cream'}), support=0.004532728969470737,
ordered_statistics=[OrderedStatistic(items_base=frozenset({'light cream'}),
items_add=frozenset({'chicken'}), confidence=0.29059829059829057, lift=4.84395061728395)])
```

From above output, we can clearly say that Chicken and light cream is having good association as the lift is more than 4



**THANK YOU**

[www.cognixia.com](http://www.cognixia.com)