Association Rule Mining

=> It discovers the probability of occurrence of items in a collection. Helps in discovering some interesting relationships in large data sets.

=> A data set contains data objects and each data object contains a set of attributes.

An attribute is called as dimension or feature or variable which represents the characteristic feature of the data object.

eg: Height, qualification, colour etc.

Association Rule Mining :-

It finds the interesting associations and relationships among large sets of data items. This rules shows how frequently a itemset occurs in a transaction.

A → Laptop, Antivirus → common Laptop
B → Laptop, mouse

for eg: Market Basket Data.

Trans ID	Item
1 →	{ Milk, Bread, Rice, Book }
2 →	{ Bread, Jam, Book, Pen }
3 →	{ Jam, Milk, Bread, Rice, Eggs }
4 →	{ Rice, Eggs, Jam, Book }
5 →	{ Eggs, Jam, Milk, Bread, Jam }
6 →	{ Eggs, Rice, Bread, Jam }

let us consider one transaction like:

{Milk, Bread, Rice, Book} {Milk} → {Bread}
{Bread, Jam, Book, Pen} {Book} → {Pen},
{Bread} → {Jam}

Some similar associations

{Milk, Oil, Liquid} → {Scrubber}.

Interesting pattern

{Eggs} → {Mouse}

Itemset :- {Milk, Bread, Jam, Rice, Eggs, Book, Pen}

Frequent Item sets:-

Two Itemsets :- {M, B}, {B, Jam}, {Rice, Eggs} {Book, Pen}

Three Itemset :- {M, B, Jam} {Rice, Eggs, Bread}, {Book, Pen, Eggs}

Four Itemsets :- {Milk, Bread, Rice, Eggs} etc.

Sup, Conf, rule Mining → 5 steps.

Support :- It is a measure of how frequently a set of items occur in total number of transactions.

eg:- {Milk, bread} → {X, Y} {X → Milk), (Y → Bread)

Therefore the frequency of occurrence of X and Y together in total No of transactions is support.

eg {Milk, Bread, Jam} → {X, Y}, {X → Milk}, {Y → Bread, Jam}

⇒ Here the frequency of occurrence of {Bread, Jam} with {Milk} in whole transaction is support. $\text{Support}(s) = \frac{\sigma(X \cup Y)}{N} =$

Confidence :- It is a measure of how often item in Y appear in transactions that contains X.

{ Milk, Bread } → { X, Y }

(X → Milk), (Y : Bread)

Therefore the frequency of occurance of x and y in all the transactions where x exists.

Confidence (C) = $\frac{\sigma(X \cup Y)}{\sigma X}$ =

Association Rule Mining :- Given a set of transactions T, goal of association rule mining is to find all rules having

support ≥ min sup threshold

confidence ≥ min conf threshold.

Eg :- Suppose, minsup = 0.3

minconf = 0.6

Consider 2-Itemsets {Rice, Eggs} → {X, Y}

Then

Support (s) : $\frac{\sigma(X \cup Y)}{N}$

$$\text{Support}(s) = \frac{3}{6} = 0.5$$

and Confidence (C) = $\frac{\sigma(X \cup Y)}{\sigma X}$

$$= \frac{3}{4} = 0.75$$

Association Rule Mining [Threshold value]

Here support :- $\sigma \geq$ minsup (0.3)

Confidence = 0.75 ≥ minconf (0.6)

Therefore, we can mine.

{Rice, Eggs} as a Rule.

Eg 2:- Suppose minsup = 0.3

minconf = 0.6

Consider, {Milk, Bread, Jam} → {X, Y}

X = {Milk} Y = {Bread, Jam}

Then Support (S) = $\frac{\sigma(X \cup Y)}{N}$

$= \frac{2}{6} = 0.333$

and Confidence (C) = $\frac{\sigma(X \cup Y)}{\sigma X}$

$= \frac{2}{3} = 0.667$

Association Rule Mining

here, Support: 0.333 ≥ minSup (0.3)

Confidence: 0.667 ≥ minConf (0.6)

Therefore, we can mine {Milk, Bread, Jam} as a Rule.

11 Apriori Algorithm :-

Transaction list :-

1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

1

⇒ Here are a dozen sales transactions

⇒ The objective is to use this transaction data to find affinities between products, that is, which products sell together often.

⇒ The support level will be set at 30% percent, the confidence level will be set at 60% percent.

⇒

1-Itemset	frequency
Milk	9
Bread	10
Butter	10
Egg	3
Ketchup	5
Cookies	5

→

frequent-1-itemset	frequency
Milk	9
Bread	10
Butter	10
Cookies	5

$\frac{4}{12} = 33\%$. Here Egg, Ketchup is not satisfied here min-support count

[4 above transactions only work here]

2-Itemset	frequency
Milk, Bread	7
Milk, Butter	7
Milk, Cookies	3
Bread, Butter	9
Butter, Cookies	3
Bread, Cookies	4

→

frequent 2-Itemset	frequency
Milk, Bread	7
Milk, Butter	7
Bread, Butter	9
Bread, Cookies	4

It should appear more than '4' Item then only it will satisfying the min-support count.

here {Milk, Cookies} {Butter, Cookies} → 3 [min-supp — 30%]

{Milk, Bread, Butter, Cookies}

3-Item set	frequency
Milk, Bread, Butter	6
Milk, Bread, Cookies	1
Bread, Butter, Cookies	3
Milk, Butter, Cookies	2

→

frequent 3-Itemset	frequency
Milk, Bread, Butter	6

one only.

Association Rule Mining - Subset Creation

⇒ frequent 1-Itemset set = I => {Milk, Bread, Butter}

⇒ Non-Empty subset are

- {Milk}, {Bread}, {Butter}, {Milk, Bread}, {Milk, Butter}, {Bread, Butter}

⇒ How to form Association Rule .. ?

- for every non-empty subset S of I, the association rule is,

* $S \rightarrow (I - S)$

* if support (I) / support (S) >= min_confidence.

Non-Empty subset are :-

- {Milk}, {Bread}, {Butter}, {Milk, Bread}, {Milk, Butter}, {Bread, Butter}

- min-support = 30%. and confidence = 60%

Rule 1:- {Milk} → {Bread, Butter} {S = 50%, C = 66.67%}

- support = 6/12 = 50%.

- Confidence = support {Milk, Bread, Butter} / support (Milk)

$= \frac{6/12}{9/12} = \frac{6}{9} = 66.67 \geq 60\%$.

- valid.

Rule 2 :- {Bread} → {Milk, Butter} {S = 50%, C = 60%.

- Support = 6/12 = 50%.

- Confidence = support {Milk, Bread, Butter} / support (Bread).

$= \frac{6}{10} = 60\% \geq 60\%$.

→ valid

Rule 1: {Butter} → {Milk, Bread} {S = 50%, C = 60%}

- support = 6/12 = 50%.
- confidence: support (Milk, Bread, Butter) / support (Butter)
  = 6/10 = 60% >= 60
- valid

Rule 2: {Milk, Bread} → {Butter} {S = 50%, C = 85.7%}

- support = 6/12 = 50%.
- Confidence = support (Milk, Bread, Butter) / support (Milk, Bread)
  = 6/7 = 85.7% ≥ 60%.
- valid.

Rule 3: {Milk, Butter} → {Bread} {S = 50%, C = 85.7%}

- support = 6/12 = 50%.
- Confidence = Supp {Milk, Bread, Butter) / support (Milk, Butter)
  = 6/7 = 85.7% >= 60%.
- valid

Rule 6: {Bread, Butter} → {Milk} {S = 50%, C = 66.67%}

- support = 6/12 = 50%.
- confidence = support {Milk, Bread, Butter) / support (Bread, Butter)
  = 6/9 = 66.67% >= 60.

⇒ Frequent pattern (FP) Growth Algorithm

Transaction ID	Item
T_1	{E, K, M, N, O, Y}
T_2	{D, E, K, N, O, Y}
T_3	{A, E, K, M}
T_4	{C, K, M, O, Y}
T_5	{C, E, I, K, O, O}

set of transaction

group item are bough together

Item	frequency
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

dataset of transaction

⇒ The above-given data is a hypothetical dataset of transaction with each letter representing an item.

⇒ Let the minimum support be 3.

⇒ A frequent pattern set $(L)_0$ is built which will contain all the elements whose frequency is greater than or equal to the minimum support.

As minimum support be 3:-

Then elements are stored in descending order of their respective frequencies.

After insertion of the relevant items, the set L looks like this :-

$$L = \{K:5, E:4, M:3, O:3, Y:3\}$$ Descending order

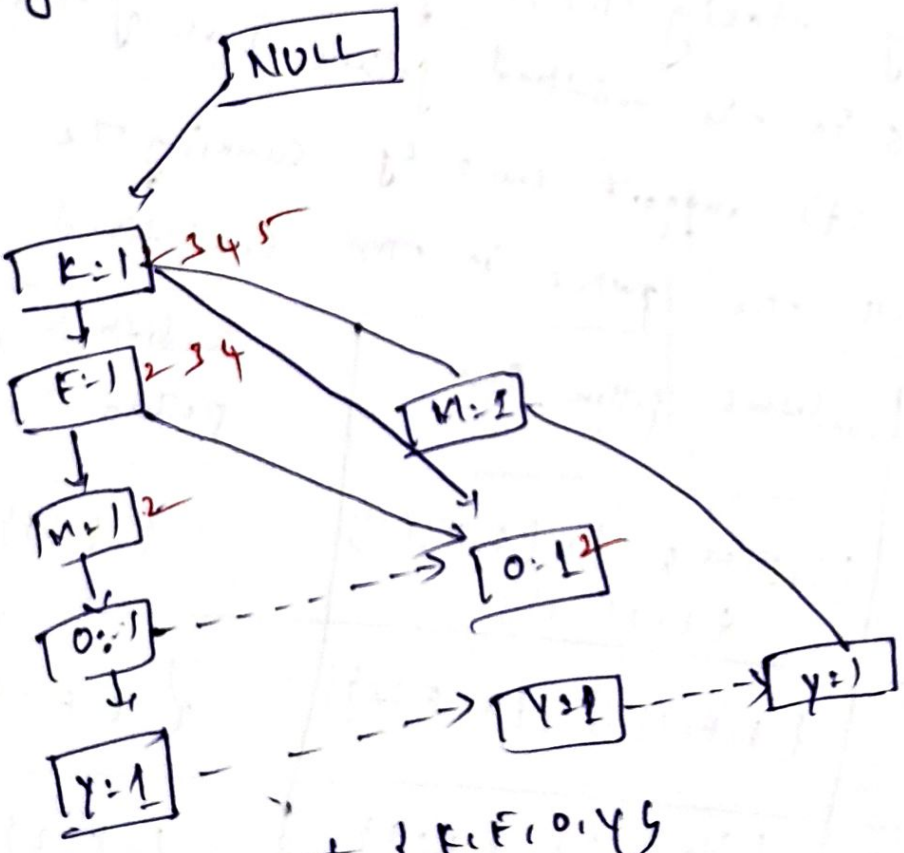
Now, for each transaction, the respective Order-Item set is built

Frequent pattern set $L = \{K:5, E:4, M:3, O:3, Y:3\}$

Transaction ID	Items	Ordered - Item set
T_1	{E, K, M, N, O, Y}	{K, E, M, O, Y}
T_2	{D, E, K, N, O, Y}	{K, E, O, Y}
T_3	{A, E, K, M}	{K, E, M}
T_4	{C, K, M, O, Y}	{K, M, Y}
T_5	{C, E, I, K, O, O}	{K, E, O}

⇒ Now, all the Ordered - Itemset are Inserted into a Trie Data structure.

a) Inserting the set {K, E, M, O, Y}



b) Inserting the set {K, E, O, Y}

c) {K E M}

d) {K, E, Y}

e) {K, E, O}

Now for each item, the Conditional pattern Base is computed which is path labels of all the paths which lead to node of the given item in the frequent-pattern tree.

Item	Conditional pattern Base
Y	{{K, F, M, O : 1}, {K, F, O : 1}, {K, M : 1}}
O	{{K, F, M : 1}, {K, F : 2}}
M	{K, F : 2}, {K : 1}}
F	{K : 4}
K	

Now for each item the Conditional frequent pattern Tree is built: It is done by taking the set of element which is common in all the paths in the conditional pattern Base of that item and calculating it's support count by summing the support count of all the paths in the Conditional pattern Base.

Item	Conditional pattern Base	Conditional frequent pattern Tree
Y	{K, F, M, O : 1}, {K, F, O : 1} {K, M : 1}}	{K : 3}
O	{K, F, M : 1}, {K, F : 2}}	{K, F : 3}
M	{K, F : 2} {K : 1}	{K : 3}
F	{K : 4}	{K : 4}
K		

From the conditional frequent pattern tree, the frequent pattern rules are generated by pairing the items of the conditional frequent pattern Tree set to the corresponding to the item as given in the below table.

Item	frequent pattern generated
Y	$\{K, Y : 3\}$
O	$\{\{K, O : 3\}, \{E, O : 3\}, \{E, K, O : 3\}\}$
M	$\{\{K, M : 3\}\}$
F	$\{\{E, F : 3\}\}$
K	

And Association rules.