Unit -3

Association Rule Mining :→

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some intersting relations or associations among the variables of datasets. The association rule learning is one of the very important concept of machine learning, and it is employed in market based Analysis, Web usage mining, continuous production etc.

For example, if a customer buys bread he most likely can also buy butter, eggs, or milk. So these products are stored within a shelf or mostly nearby.

Association rule learning can be divided into three type of algorithms :—

1) Apriori
2. Eclat
3 F-P Growth Algorithm

Apriori Algorithm

It uses
to generate association rules & it
designed to work on datasets or datab
that contain transactions.

This algo. uses a BFS & Hash tree to calcu
the itemset associations efficiently.

Frequent itemset :— FI are those item
whose support is greater than threshol
value.

Eg :— A = {1,2, 3,4,5} B = {2,3,7}

2,3 are FI values.

Step-I Determine support of itemsets &
Select min support & confidence.

Step-II Take all supports in transaction w
higher support value than min or selected s
value.

Step-III Find all rules of these sets that h
higher confidence value than threshold
min confidence

Step-IV Sort the rules as decreasing ord
of lift.

Apriori Algorithm:-

It uses frequent itemsets to generate association rules & it is designed to work on datasets or databases that contain transactions.

This algo. uses a BFS & Hash tree to calculate the itemset associations efficiently.

Frequent itemset:- FI are those items whose support is greater than threshold value.

Eg:- $A = \{1,2,3,4,5\}$ $B = \{2,3,7\}$

2,3 are FI values.

Step-I: Determine support of itemsets & Select min support & confidence.

Step-II Take all supports in transaction with higher support value than min or selected support value.

Step-III Find all rules of these sets that have higher confidence value than threshold or min confidence

Step-IV Sort the rules as decreasing order of lift.

Apriori Algo.

Eg:-

Tid.	Itemsets
T_1	A, B
T_2	B, D
T_3	B, C
T_4	A, B, D
T_5	A, C
T_6	B, C
T_7	A, C
T_8	A, B, C, E
T_9	A, B, C

Given :

min support = 2.

min confidence = 50%.

C → Customer

FI → Frequent itemset.

Step ① Calculating C_1 and F_1

C_1

Itemset	Support count
A	6
B	7
C	6
D	2
E	1

→ F_1

Itemset	SC
A	6
B	7
C	6
D	2

Step ② Calculating C_2 and F_2.

C_2

Itemset	SC
(A, B)	4
(A, C)	4
(A, D)	1
(B, C)	4
(B, D)	2
(C, D)	0

→ F_2

Itemset	SC
(A, B)	4
(B, C)	4
(A, C)	4
(B, D)	2

Step ③ Candidate Generation C_3 & F_3

IS	SC
(A, B, C)	2
(B, C, D)	0
(A, C, D)	0
(A, B, D)	1

→ F_3

Itemset	SC
(A, B C)	2

Step(4) Finding Association rules for subsets:—

→ We will calculate the confidence using

$$\frac{Sup(A \cap B)}{A} \rightarrow \quad \text{Sup} = \frac{A \rightarrow B}{A}$$

→ After calculating we will exclude rules that have less than confidence than min threshold (50%)

Rules	Support	Confidence.
$A \cap B \rightarrow C$	2	$Sup((A \cap B) \cap C) / Sup(A \cap B) = 2/4 = 50\%$
$B \cap C \rightarrow A$	2	$Sup((B \cap C) \cap A) / Sup(B \cap C) = 2/4 = 50\%$
$A \cap C \rightarrow B$	2	$Sup((A \cap C) \cap B) / Sup(A \cap C) = 2/4 = 50\%$
$C \rightarrow A \cap B$	2	$Sup((C) \cap (A \cap B)) / Sup(C) = 2/5 = 40\%$
$A \rightarrow B \cap C$	2	$Sup(A) (B \cap C) / Sup(A) = 2/6 = 33.33\%$
$B \rightarrow A \cap C$	2	$Sup((B)(A \cap C) / Sup(B) = 2/7 = 28.\%$

Frequent Pattern Growth Algo :→ It is improved version of Apriori Algo. It represents the database in the form of tree structure. that is known as frequent pattern or tree.

Tid	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, U, Y
T_5	C, E, I, K, O, O

min Support = 3.

②

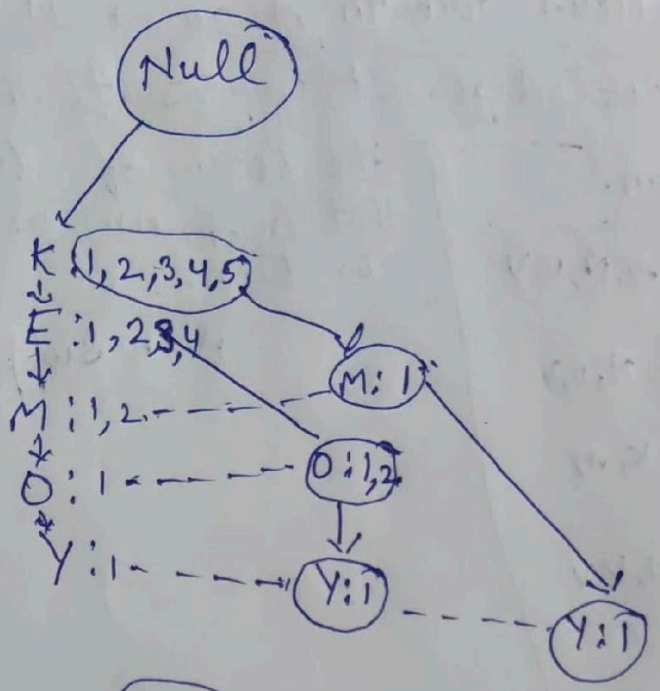
Item	Frequency
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
U	1
O	3
Y	3

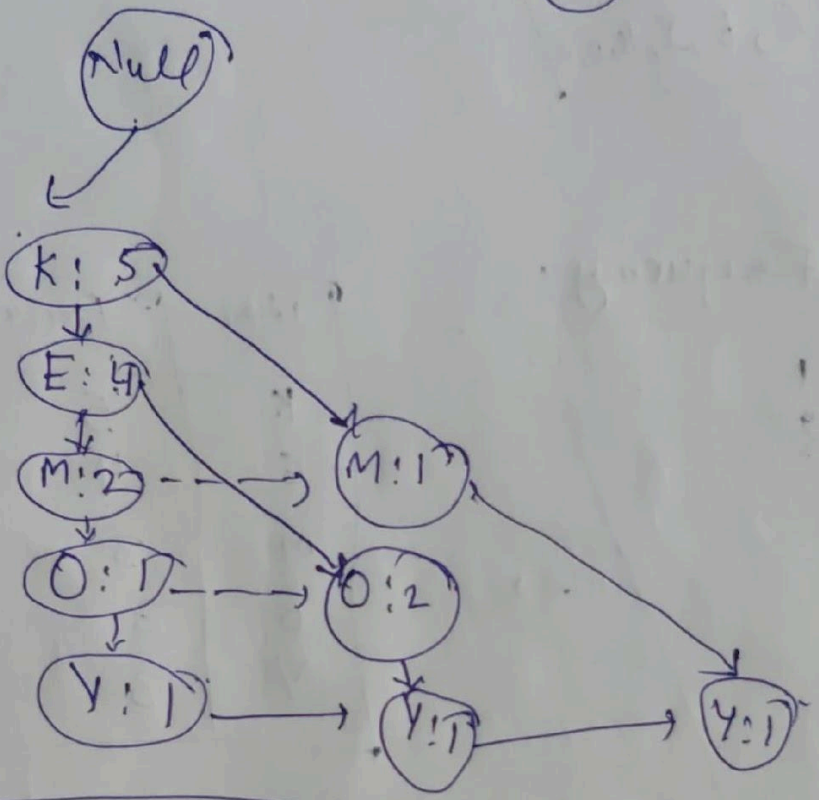
→

Item	Frequency
K	5
E	4
M	3
O	3
Y	3

③

Tid	Items	ordered - Itemset
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, U, Y	K, M, Y
T_5	C, E, I, K, O, O	K, E, O





Item.	Conditional Pattern Base	Conditional Freq.
Y.	{K,E,M,O : 1}, {K,E,O : 1}, {K,M,Y : 1}	{K : 3}
O	{K,E,M : 1}, {K,E : 2}	{K,E : 3}
M.	{K,E : 2}, {K : 1}	{K : 3}
E	{K : 4}	{K : 4}

Items.	Frequent Pattern Generated
Y	{<K,Y : 3>}
O	{<K,O : 3>}, {<E,O : 3>}, {<K,E,O : 3>}
M	{<K,M : 3>}
E	{<K,E : 3>}
K	———

$K \rightarrow Y$, $Y \rightarrow K$

$K \rightarrow O$, $E \rightarrow O$, $O \rightarrow K$, $O \rightarrow E$, $E \rightarrow K$, $K \rightarrow E$

$K \rightarrow M$, $M \rightarrow K$.

$K \rightarrow E$, $E \rightarrow K$.

* Eclat Algorithm

Eclat Algorithm stands for Equivalence class. Clustering and bottom-up Lattice Traversal It is one of popular methods of Association rule mining Algorithm.

Transaction id	Bread	Butter	milk	Coke	Jam
T_1	1	1	0	0	1
T_2	0	1	0	1	0
T_3	0	1	1	0	0
T_4	1	1	0	1	0
T_5	1	0	1	0	0
T_6	0	1	1	0	0
T_7	1	0	1	0	0
T_8	1	1	1	0	1
T_9	1	1	1	0	1

K=1, Minimum Support = 2

Item	Tidset
Bread	$T_1, T_4, T_5, T_7, T_8, T_9$
Butter	$T_1, T_2, T_3, T_4, T_6, T_8, T_9$
Milk	$T_3, T_5, T_6, T_7, T_8, T_9$
Cake	T_2, T_4
Jam	T_1, T_8

K=2, MS=2.

item	Tidset
(B, Bu)	T_1, T_4, T_8, T_9
(B, M)	T_5, T_7, T_8, T_9
(B, C)	T_4
(B, J)	T_1, T_8
(Bu, M)	T_3, T_6, T_8, T_9
(Bu, C)	T_2, T_4
(Bu, J)	T_1, T_8
(M, C)	∅
(M, J)	T_8
(C, J)	∅

K=3, MS=2.

item	Tidset
B, BU, M	T_8, T_9
B, BU, J	T_8, T_1
B, BU, C	T_4
B, M, J	T_8

K=4.

item	Tidset
B, BU, M, J	T_8

Advantages:

1. **Memory Requirements** :→ ECLAT Algo. uses a Depth-first Search Approach, it uses less memory than Apriori.

- **Speed** :→ The ELCAT Algo. is typically faster than Apriori Algo.

- **Number of Computations**:— The ECLAT Algo. does not involves the repeated scanning of the data to compute the individual support values.

Decision Tree :-

Decision tree algorithm falls under the category of supervised learning. The can be used to solve both regression and classification problems. It is a tree that helps us in decision making. purpose. the decision tree creates classification or regression models as a tree structure. Decision tree uses the tree representation to solve the problem decision tree contains 3 types of nodes.

① root node :- It is the top most node in the tree. data which is inside the node is known as attribute.

② Internal node :→ Each internal node denotes a test on attribute. nodes which are in between root node and are called as internal nodes.

3. Leaf node :- We call last nodes as leaf nodes. They represents output is class label.

Advantages : 1) It does not require any domain knowledge

② classification steps of decision tree are simple and fast.

③. Missing values in data does not effect output.

④. A decision tree model is automatic and does not require a standardization of data.

key factors :- Building a decision tree is all about discovery attributes that return the highest data gain

Entropy :→

Entropy refers to a common way to measure impurity. In the decision tree. measure the impurity in dataset.

Information Gain :-

Information Gain refers to the decline in entropy after the dataset is split. It is also called as entropy reduction.

Example :-

Day	Weather	Temperature	Humidity	Wind	Play
1	Sunny	Hot	High	weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Cloudy	mild	High	Strong	Yes
4	Rainy	mild	High	Strong	No
5	Sunny	mild	Normal	Strong	Yes
6	Rainy	Cool	Normal	Strong	No
7	Rainy	mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	mild	High	Strong	No