UNIT-4

ASSOCIATION RULES

Association Rules(**if this then that)**

- Association Rule Mining is used when you want to find an association between different objects in a set, find frequent patterns in a transaction database, relational databases or any other information repository.

- The applications of Association Rule Mining are found in Marketing, Basket Data Analysis (or Market Basket Analysis) in retailing, clustering and classification.

- It can tell you what items do customers frequently buy together by generating a set of rules called **Association Rules**.

ASSOCIATION RULES

- If we think of the universe as the set of items available at the store, then each item has a Boolean variable representing the presence or absence of that item.

- Each basket can then be represented by a Boolean vector of values assigned to these variables.

- The Boolean vectors can be analyzed for buying patterns that reflect items that are frequently *associated or purchased together*.

- *These patterns can be represented in the* form of association rules

- For example, the information that customers who purchase computers also tend to buy antivirus software

Marketing Strategies

Clients can use those rules for numerous marketing strategies:

- Changing the store layout according to trends
- Customer behavior analysis
- Catalogue design
- Cross marketing on online stores
- What are the trending items customers buy
- Customized emails with add-on sales

- Association rules are given in the form as below:

  A=>B[Support, Confidence]
- The part before => is referred to as *if (Antecedent)* and the part after => is referred to as *then (Consequent)*.

Where A and B are sets of items in the transaction data. *A and B are disjoint sets*.

ASSOCIATION RULES

- ***Computer=>antivirus software [support = 20%; confidence = 60%]***

Above rule says:

- 20% transaction show Anti-virus software is bought with purchase of a Computer

- 60% of customers who purchase Anti-virus software is bought with purchase of a Computer

- Typically, association rules are considered interesting if they satisfy both a **minimum support threshold** and a **minimum confidence threshold**.

- Such thresholds can be set by users or domain experts.

- Support and Confidence measure how interesting the rule is.
- It is set by the minimum support and minimum confidence thresholds.
- These thresholds set by client help to compare the rule strength according to your own or client's will.
- The closer to threshold the more the rule is of use to the client.

Basic Concepts of Association Rule Mining

- **Itemset:** Collection of one or more items. K-item-set means a set of k items.

- **Support Count:** Frequency of occurrence of an item-set

- **Support (s):** Fraction of transactions that contain the item-set 'X'

  Support(X)=frequency(X)/Support(X)

-

What is itemset?

- An itemset is a set of one or more items.

Transaction ID	Items bought
1	Tea, Cake, Cold Drink
2	Tea, Coffee, Cold Drink
3	Eggs, Tea, Cold Drink
4	Cake, Milk, Eggs
5	Cake, Coffee, Cold Drink, Milk, Eggs

Example:

- Transaction 1 showing an itemset containing items like Tea, Cake, Cold Drink.

- Transaction 2 showing an itemset containing items like Tea, Coffee, Cold Drink.

- Transaction 3 showing an itemset containing items like Eggs, Tea, Cold Drink.

- Transaction 4 showing an itemset containing items like Cake, Milk, Eggs.

- Transaction 5 showing an itemset containing items like Cake, Coffee, Cold Drink, Milk, Eggs.

What is K-itemset?

- When K=1, then K-Itemset is itemset 1.
- When K=2, then K-Itemset is itemset 2.
- When K=3, then K-Itemset is itemset 3.
- When K=4, then K-Itemset is itemset 4.
- When K=5, then K-Itemset is itemset 5.

What is a frequent itemset?

- An itemset is *frequent* if its support is no less than "minimum support threshold".
- Minimum support is always supposed according to the choice.
- You can select any minimum support to decide that the itemset is frequent or not.

How to calculate support and confidence

- For a Rule A=>B, Support is given by:

  Support(A=>B)=frequency(A,B)/N

- $support(A=>B) = P(AUB)$
- $confidence(A=>B) = P(B|A)$

$$confidence(A \Rightarrow B) = P(B|A) = \frac{support(A \cup B)}{support(A)} = \frac{support\_count(A \cup B)}{support\_count(A)}.$$

Support or Absolute support

- Absolute Support of Tea: 3
- Absolute Support of Cake : 3
- Absolute Support of Cold Drink: 4
- Absolute Support of Milk: 2
- Absolute Support of Eggs: 3
- Support that if a person buy Tea, also buy Cake : 1 / 5 = 0.2 = 20%
- Support that if a person buy Tea, also buy Cold Drink : 3 / 5 = 0.6 = 60%
- The support that if a person buys Eggs, also buy Cold Drink: 2 / 5 = 0.4 = 40%

and similarly, we can calculate support for all itemsets.

Relative support

- The relative number of transactions which contains an itemset relative to the total transactions.

Formula:

- Total number of transactions containing an itemset X / Total number of transactions

- Relative Support of Tea: 3 / 5 = 0.6

- Relative Support of Cake : 3 / 5 = 0.6

- Relative Support of Cold Drink : 4 / 5 = 0.8

- Relative Support of Milk : 2 / 5 = 0.4

- Relative Support of Eggs: 3 / 5 = 0.6

Confidence

- Confidence is the probability that if a person buys an item A, then he will also buy an item B.
- Confidence that if a person buy Tea, also buy Cake : 1 / 3 = 0.2 = 20%
  - Why 1? because Tea and Cake occur together only in 1 transaction
  - Why 3? because there are three transactions in which Tea is occurring.
- Confidence that if a person buy Cake, also buy Tea : 1 / 3 = 0.2 = 20%.
  - Why 1? because Tea and Cake are occurring together only in 1 transaction
  - Why 3? because there are three transactions in which Tea is occurring.
- Confidence that if a person buy Milk, also buy Tea : 0 / 2 = 0 = 0%
  - Why 0? because Milk and Tea are not occurring together in any transaction
  - Why 2? because there are 2 transactions in which Milk is occurring
- and similarly, we can calculate confidence for all itemsets.