

Association rule



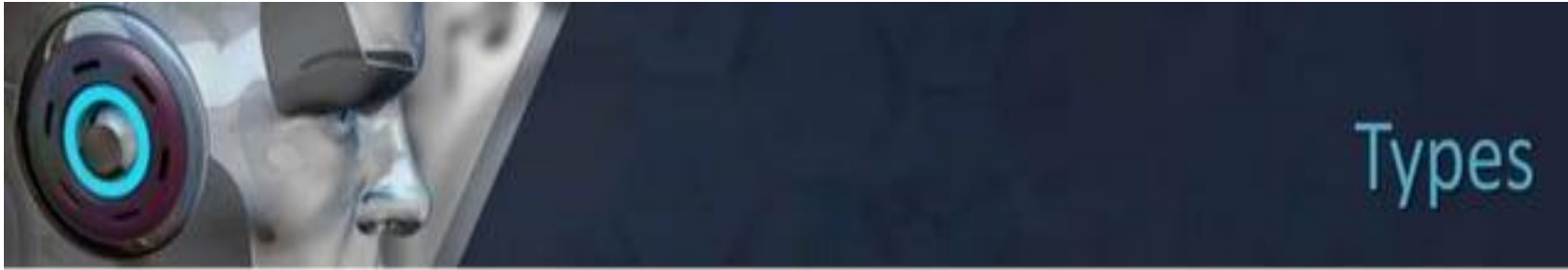
Definition

- 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.



Explanation

- The association rule learning is one of the very important concepts of machine learning, and it is employed in **Market Basket analysis**, **Web usage mining**, **continuous production**, etc. Here market basket analysis is a technique used by the various big retailer to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.
- 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 types of algorithms:
 1. **Apriori**
 2. **Eclat**
 3. **F-P Growth Algorithm**



Apriori Algorithm

- This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset efficiently.
- It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.



Eclat Algorithm

- Eclat algorithm stands for **Equivalence Class Transformation**. This algorithm uses a depth-first search technique to find frequent itemsets in a transaction database. It performs faster execution than Apriori Algorithm.



F-P Growth Algorithm

- The F-P growth algorithm stands for **Frequent Pattern**, and it is the improved version of the Apriori Algorithm. It represents the database in the form of a tree structure that is known as a frequent pattern or tree. The purpose of this frequent tree is to extract the most frequent patterns.



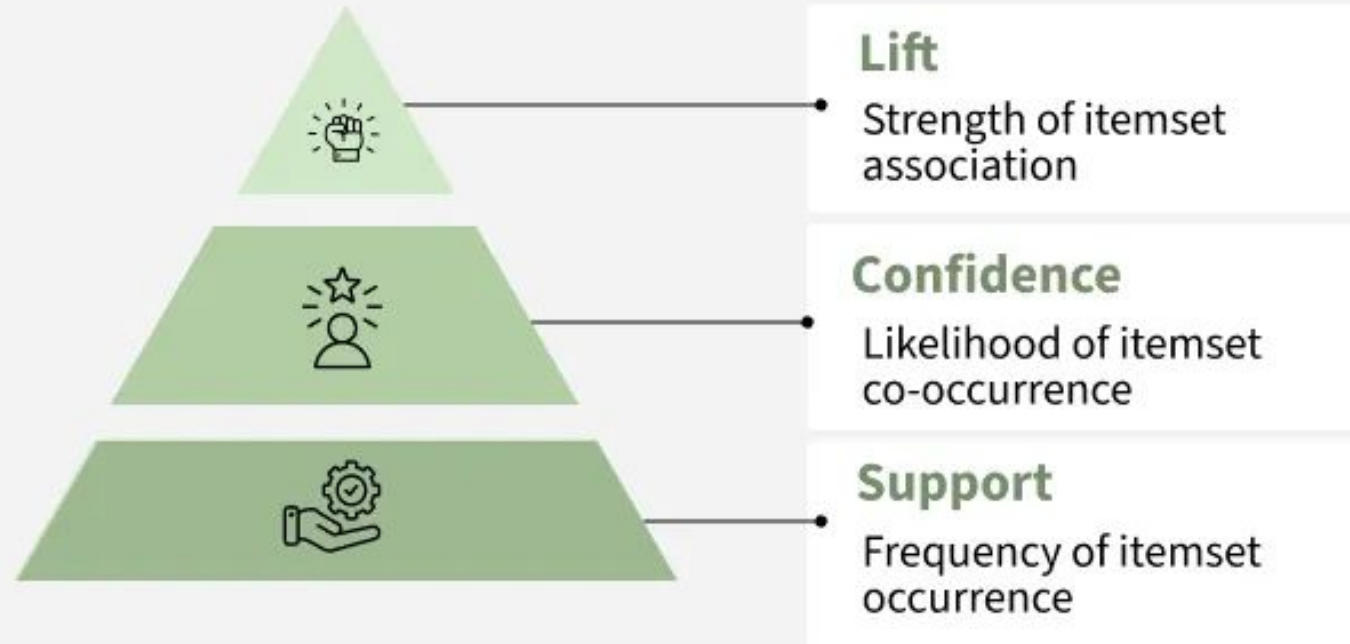
- Association rule learning works on the concept of If and Else Statement, such as if A then B.

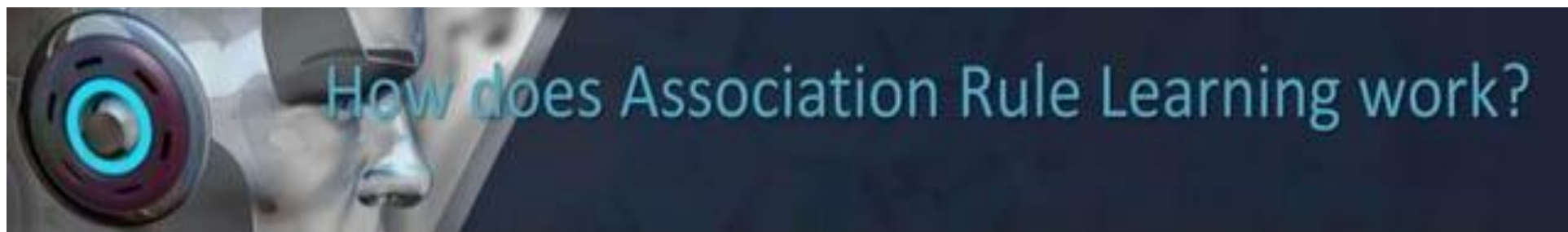




- Here the If element is called **antecedent**, and then statement is called as **Consequent**. These types of relationships where we can find out some association or relation between two items is known *as single cardinality*. It is all about creating rules, and if the number of items increases, then cardinality also increases accordingly. So, to measure the associations between thousands of data items, there are several metrics. These metrics are given below:
 - **Support**
 - **Confidence**
 - **Lift**

Apriori Algorithm Metrics Hierarchy

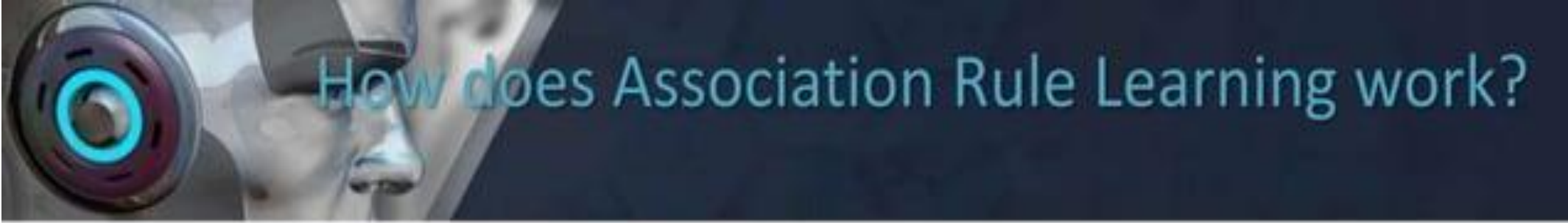




Support

Support is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the itemset X. If there are X datasets, then for transactions T, it can be written as:

$$\text{Supp}(X) = \frac{\text{Freq}(X)}{T}$$



How does Association Rule Learning work?

Confidence

Confidence indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.

$$\text{Confidence} = \frac{\text{Freq}(X, Y)}{\text{Freq}(X)}$$



How does Association Rule Learning work?

Lift

It is the strength of any rule, which can be defined as below formula:

$$\text{Lift} = \frac{\text{Supp}(X,Y)}{\text{Supp}(X) \times \text{Supp}(Y)}$$

It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has three possible values:

If **Lift= 1**: The probability of occurrence of antecedent and consequent is independent of each other.

Lift>1: It determines the degree to which the two itemsets are dependent to each other.

Lift<1: It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

Example

TID	Items Bought
T1	Milk, Bread, Butter
T2	Bread, Butter
T3	Milk, Bread
T4	Milk, Butter
T5	Milk, Bread, Butter

Total transactions = 5

Candidate Rules

Let's evaluate the rule:

Milk & Bread \Rightarrow Butter

Step 1: Support

Support = proportion of all transactions that contain Milk, Bread, and Butter together

Transactions containing {Milk, Bread, Butter} = T1 and T5 \rightarrow 2 transactions

Total transactions = 5

Support = $2/5 = 0.4$

Step 2: Confidence

Confidence = proportion of transactions containing Milk and Bread that also contain Butter

Transactions containing {Milk, Bread} = T1, T3, T5 \rightarrow 3 transactions

Out of these, those also containing Butter = T1 and T5 \rightarrow 2 transactions

Confidence = $2/3 = 0.67$

Step 3: Lift

Lift compares the observed confidence with the expected confidence if the items were independent.

Lift = Confidence (Milk, Bread \Rightarrow Butter) / Support(Butter)

Support(Butter) = transactions containing Butter = T1, T2, T4, T5 $\rightarrow 4/5 = 0.8$

Lift = $0.67/0.8 = 0.84$



Applications of Association Rule Learning

- It has various applications in machine learning and data mining. Below are some popular applications of association rule learning:
- **Market Basket Analysis:** It is one of the popular examples and applications of association rule mining. This technique is commonly used by big retailers to determine the association between items.
- **Medical Diagnosis:** With the help of association rules, patients can be cured easily, as it helps in identifying the probability of illness for a particular disease.
- **Protein Sequence:** The association rules help in determining the synthesis of artificial Proteins.
- It is also used for the **Catalog Design** and **Loss-leader Analysis** and many more other applications.

Apriori Algorithm



Definition

- The Apriori algorithm is an algorithm which uses frequent itemsets to generate association rules, and it is designed to work on the databases that contain transactions.



Explanation

- With the help of these association rule, it determines how strongly or how weakly two objects are connected.
- This algorithm uses a **breadth-first search** and **Hash Tree** to calculate the itemset associations efficiently. It is the iterative process for finding the frequent itemsets from the large dataset.
- This algorithm was given by the **R. Agrawal** and **Srikant** in the year **1994**.
- It is mainly used for *market basket analysis* and helps to find those products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.



What is Frequent Itemset?

- Frequent itemsets are those items whose support is greater than the threshold value or user-specified minimum support. It means if A & B are the frequent itemsets together, then individually A and B should also be the frequent itemset.
- Suppose there are the two transactions: A = {1,2,3,4,5}, and B = {2,3,7}, in these two transactions, 2 and 3 are the frequent itemsets.



Steps for Apriori Algorithm

- Below are the steps for the apriori algorithm:
- **Step-1:** Determine the support of itemsets in the transactional database, and select the minimum support and confidence.
- **Step-2:** Take all supports in the transaction with higher support value than the minimum or selected support value.
- **Step-3:** Find all the rules of these subsets that have higher confidence value than the threshold or minimum confidence.
- **Step-4:** Sort the rules as the decreasing order of lift.

