Apriori Algorithm

The Apriori Algorithm is one of the most widely used data mining algorithms for discovering frequent itemsets and generating association rules from large datasets. It was introduced by Rakesh Agrawal and Ramakrishnan Srikant in 1994 and is mainly used in Market Basket Analysis.

The algorithm identifies relationships among items in transactional databases and helps organizations understand customer purchasing behavior. For example, in a supermarket, if customers who buy bread also frequently buy butter, the Apriori algorithm can discover this association.

Apriori is based on the principle that if an itemset is frequent, then all of its subsets must also be frequent. This principle is known as the Apriori Property.

The algorithm is widely applied in retail stores, e-commerce websites, recommendation systems, healthcare, banking, and web usage mining.

Association Rule Mining

Association Rule Mining is a technique used to discover relationships between variables in large datasets.

An association rule is generally represented as:

[X \rightarrow Y]

Where:

- X = Antecedent (If part)
- Y = Consequent (Then part)

Example

[Bread \rightarrow Butter]

This rule indicates that customers who purchase bread are likely to purchase butter as well.

The Apriori algorithm generates such rules by finding frequent itemsets.

Basic Terminology

1. Item

An item refers to a product or object in a transaction.

Example

- Milk
- Bread
- Butter

2. Itemset

A collection of one or more items is called an itemset.

Example

[{Milk, Bread}]

3. Transaction

A transaction is a set of items purchased together by a customer.

Example

Transaction ID	Items Purchased
T1	Milk, Bread, Butter
T2	Bread, Butter
T3	Milk, Bread
T4	Milk, Butter

4. Frequent Itemset

An itemset whose occurrence frequency is greater than or equal to the minimum support threshold is called a frequent itemset.

Apriori Property

The Apriori algorithm works based on the following principle:

Apriori Principle

"If an itemset is frequent, then all of its subsets must also be frequent."

Similarly,

"If an itemset is infrequent, then all of its supersets will also be infrequent."

This property helps reduce the search space and improves algorithm efficiency.

Example

If:

[{Milk, Bread, Butter}]

is frequent, then the following subsets must also be frequent:

[{Milk, Bread}]

[{Milk, Butter}]

[{Bread, Butter}]

Measures Used in Apriori

1. Support

Support measures how frequently an itemset appears in the dataset.

Formula

[Support(X)=\frac{\text{Number of transactions containing X}}{\text{Total number of transactions}}]

Example

If Milk appears in 4 out of 10 transactions:

[Support(Milk)=\frac{4}{10}=0.4]

or

[40%]

2. Confidence

Confidence measures the likelihood that item Y is purchased when item X is purchased.

Formula

[Confidence(X \rightarrow Y)\frac{Support(X \cup Y)}{Support(X)}]

Example

If 20 customers buy Bread and 15 of them also buy Butter:

[
Confidence(Bread \rightarrow Butter)\frac{15}{20}0.75]

or

[75%]

3. Lift

Lift measures the strength of an association rule.

Formula

[
Lift(X \rightarrow Y)\frac{Confidence(X \rightarrow Y)}{Support(Y)}]

Interpretation

- Lift > 1 : Positive association
- Lift = 1 : No association
- Lift < 1 : Negative association

Working of Apriori Algorithm

The Apriori algorithm follows an iterative process.

Step 1: Generate Candidate Itemsets

Create all possible itemsets of size 1.

Step 2: Calculate Support

Calculate support for each itemset.

Step 3: Prune Infrequent Itemsets

Remove itemsets whose support is less than the minimum support threshold.

Step 4: Generate Larger Itemsets

Combine frequent itemsets to form larger candidate itemsets.

Step 5: Repeat

Continue until no more frequent itemsets can be generated.

Step 6: Generate Association Rules

Use confidence and lift measures to create strong association rules.

Example of Apriori Algorithm

Consider the following transactions:

Transaction ID	Items
T1	Milk, Bread, Butter
T2	Bread, Butter
T3	Milk, Bread
T4	Milk, Butter
T5	Bread, Butter

Frequent 1-Itemsets

- Milk
- Bread
- Butter

Frequent 2-Itemsets

- Milk, Bread
- Milk, Butter

- Bread, Butter

Frequent 3-Itemsets

- Milk, Bread, Butter

After identifying frequent itemsets, association rules are generated.

Example Rule:

[Bread \rightarrow Butter]

This suggests customers buying bread are likely to buy butter.

Advantages of Apriori Algorithm

1. Easy to Understand

The algorithm is simple and straightforward.

2. Effective for Market Basket Analysis

Widely used for discovering purchasing patterns.

3. Generates Meaningful Rules

Provides valuable insights into customer behavior.

4. Reduces Search Space

Apriori property eliminates unnecessary candidate itemsets.

5. Useful for Decision Making

Helps businesses improve marketing and sales strategies.

Disadvantages of Apriori Algorithm

1. High Computational Cost

Generates a large number of candidate itemsets.

2. Multiple Database Scans

Requires repeated scanning of the database.

3. Poor Performance on Large Datasets

Execution time increases significantly with dataset size.

4. Memory Consumption

Large candidate itemsets require more memory.

Applications of Apriori Algorithm

Market Basket Analysis

Analyzes products frequently purchased together.

Recommendation Systems

Suggests products based on customer purchasing history.

Retail Business

Improves store layout and product placement.

E-Commerce

Generates personalized recommendations.

Web Usage Mining

Analyzes user navigation patterns.

Healthcare

Identifies relationships between diseases, symptoms, and treatments.

Banking and Finance

Detects patterns in customer transactions.

Comparison with Other Algorithms

Feature	Apriori	FP-Growth
Database Scans	Multiple	Fewer
Candidate Generation	Required	Not Required
Speed	Slower	Faster

Feature	Apriori	FP-Growth
Memory Usage	Higher	Lower
Implementation	Simple	More Complex