Assignment No. 5 Apriori

**Name: Sumit Avinash Solanke**

**Roll No: TYITB121**

**Assignment No. 5**

**AIM:** Assignment of Aproori algorithm on OnlineRetail.csv

**PREREQUISITE:** Python programming

**THEORY:**

**Introduction to Association Rule Learning**

**Association rule learning** is a rule-based machine learning method for discovering interesting relations between variables in large datasets. It is commonly used in **market basket analysis**, where the goal is to find associations between items bought together.

**Objective**

The objective of this assignment is to apply the Apriori algorithm on the *OnlineRetail.csv* dataset to discover meaningful association rules and frequent itemsets from transaction data. The aim is to uncover hidden patterns in customer purchasing behavior, which can be used for decision-making in areas such as product bundling, recommendation systems, and inventory management. This assignment focuses on understanding market basket analysis and the core concepts of support, confidence, and lift in association rule mining.

**Apriori Algorithm – Overview**

The **Apriori algorithm** is a classic algorithm in data mining for learning **association rules**. It is named “Apriori” because it uses a prior knowledge of frequent itemset properties. The algorithm works by identifying the frequent individual items in the dataset and extending them to larger and larger itemsets as long as those itemsets appear sufficiently often in the dataset.

It is a **bottom-up approach**, where:

1. It starts from single items (1-itemsets),
2. Builds larger itemsets (k-itemsets),
3. And prunes those that don’t satisfy the **minimum support threshold**.

**Key Concepts**

**1. Itemset**

An **itemset** is simply a collection of one or more items. For example, in a supermarket: {Milk, Bread} is an itemset.

**2. Support**

The **support** of an itemset is the proportion of transactions in the dataset that contain the itemset.  
**Support(A)** = (Number of transactions containing A) / (Total transactions)

**3. Confidence**

The **confidence** of a rule A → B indicates how often B appears in transactions that contain A.  
**Confidence(A→B)** = Support(A ∪ B) / Support(A)

**4. Lift**

The **lift** of a rule compares the observed support with what would be expected if A and B were independent.  
**Lift(A→B)** = Confidence(A→B) / Support(B)

* Lift > 1: Positive correlation
* Lift = 1: No correlation
* Lift < 1: Negative correlation

**Working of Apriori Algorithm**

1. **Generate 1-itemsets** from the transaction database that meet the minimum support.
2. **Generate candidate 2-itemsets** using the frequent 1-itemsets.
3. **Prune** itemsets that have infrequent subsets (Apriori Property).
4. **Repeat** until no more frequent itemsets can be generated.
5. **Generate association rules** from the frequent itemsets that meet the confidence threshold.

**Professional Example: Market Basket Analysis**

Suppose we analyze a grocery store’s transaction log:

| **Transaction ID** | **Items** |
| --- | --- |
| T1 | Milk, Bread, Butter |
| T2 | Bread, Butter |
| T3 | Milk, Bread |
| T4 | Milk, Bread, Butter, Eggs |
| T5 | Bread, Butter |

* **Step 1:** Count itemsets like {Milk}, {Bread}, etc.
* **Step 2:** Build combinations like {Milk, Bread}, {Bread, Butter}, etc.
* **Step 3:** Apply support and confidence thresholds to extract rules like:  
  **“If Bread, then Butter”** with a confidence of 80%.

**Advantages:**

* Easy to implement and interpret
* Effective for datasets with large transactions and discrete values
* Generates all possible frequent itemsets and rules

**Limitations:**

* May generate **too many rules**, some of which may be redundant
* High **computational cost** with large datasets
* Requires **setting thresholds** (support/confidence) carefully

**Dataset Overview:**

This dataset contains transactional data for a UK-based online retail store that sells gifts and other products to customers worldwide. It includes information on invoices, products, quantities, prices, and customer IDs, collected between December 2010 and December 2011.

**Dataset Characteristics:**

| Attribute | Value |
| --- | --- |
| Total Records | 541,909 |
| Features | 8 columns |
| Time Span | 1 year (2010–2011) |
| Missing Values | Yes (CustomerID has missing entries) |
| Country Scope | Primarily UK, with others too |

**Column Descriptions:**

| Column Name | Data Type | Description |
| --- | --- | --- |
| InvoiceNo | String | Unique invoice number. If it starts with 'C', it indicates a cancellation |
| StockCode | String | Product/item code |
| Description | String | Product description |
| Quantity | Integer | Number of units purchased (can be negative for returns) |
| InvoiceDate | Datetime | Date and time of the invoice |
| UnitPrice | Float | Price per unit of product (in GBP) |
| CustomerID | Float | Unique customer identifier (some are missing) |
| Country | String | Country of the customer |

Sample Records:

| InvoiceNo | StockCode | Description | Quantity | InvoiceDate | UnitPrice | CustomerID | Country |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 536365 | 85123A | WHITE HANGING HEART... | 6 | 01/12/2010 08:26 | 2.55 | 17850.0 | United Kingdom |
| 536365 | 71053 | WHITE METAL LANTERN | 6 | 01/12/2010 08:26 | 3.39 | 17850.0 | United Kingdom |
| 536366 | 84406B | CREAM CUPID HEARTS... | 8 | 01/12/2010 08:28 | 2.75 | 13047.0 | United Kingdom |

**CONCLUSION:**

The **Apriori algorithm** remains one of the foundational algorithms in data mining, widely used in practical scenarios like market basket analysis, recommendation systems, and inventory planning. It helps uncover hidden patterns and associations in large datasets. Understanding Apriori provides a strong base for learning more advanced machine learning concepts like frequent pattern trees (FP-Growth) and sequential pattern mining.