APRIORI ALGORITHM

1. Introduction

The Apriori Algorithm is a popular data mining algorithm used to discover frequent itemsets and generate association rules from transactional datasets. It is mainly used in Association Rule Mining and Market Basket Analysis.

Example:

Bread, Milk → Butter

This represents a possible association between products purchased by customers.

2. Objectives of Apriori Algorithm

- Find frequent itemsets from transaction data.
- Identify relationships between items.
- Generate useful association rules.
- Support decision-making in applications such as recommendation and market basket analysis.

3. Important Terms

Term	Meaning
Transaction	A collection of items occurring or purchased together.
Itemset	A set of one or more items.
Frequent Itemset	An itemset whose support is at least the minimum support threshold.
Minimum Support	The minimum frequency required for an itemset to be considered frequent.
Association Rule	A relationship represented as X → Y, where X is antecedent and Y is consequent.

4. Apriori Principle

The most important concept of Apriori is the Apriori Principle: If an itemset is frequent, all of its non-empty subsets must also be frequent. Similarly, if an itemset is not frequent, any larger itemset containing it cannot be frequent.

Example:

{Bread, Milk} is not frequent

Therefore, {Bread, Milk, Butter} cannot be frequent.

This principle helps Apriori reduce the number of candidate itemsets by pruning unnecessary candidates.

5. Support

Support measures how frequently an itemset appears in the transaction database.

$$\textbf{Support(X) = Number of transactions containing X / Total number of transactions}$$

Example: If {Bread, Milk} occurs in 3 out of 5 transactions:

$$\textbf{Support(Bread, Milk) = 3/5 = 60\%}$$

6. Confidence

Confidence measures how reliably the occurrence of X is associated with Y.

$$\textbf{Confidence}(X \rightarrow Y) = \textbf{Support}(X \cup Y) / \textbf{Support}(X)$$

Example: If Bread occurs in 4 transactions and Bread + Milk occur together in 3:

$$\textbf{Confidence(Bread} \rightarrow \textbf{Milk) = 3/4 = 75\%}$$

7. Steps of Apriori Algorithm

1. Set a minimum support threshold.
2. Generate candidate 1-itemsets (C1).
3. Calculate the support of each candidate itemset.
4. Remove itemsets below minimum support to obtain frequent 1-itemsets (L1).
5. Generate candidate 2-itemsets (C2) from L1.
6. Calculate support and prune infrequent 2-itemsets to obtain L2.
7. Generate larger candidate itemsets from the previous frequent itemsets.
8. Repeat support calculation and pruning until no more frequent itemsets can be generated.
9. Generate association rules from the frequent itemsets.
10. Calculate confidence and select rules satisfying the minimum confidence threshold.

8. Apriori Algorithm Flow

Transaction Database

↓

Set Minimum Support

↓

Generate C1

↓

Calculate Support

↓

Prune Infrequent Itemsets

↓

Generate C2, C3, ...

↓

Repeat Support Calculation and Pruning

↓

Frequent Itemsets

↓

Generate Association Rules

↓

Calculate Confidence

↓

Final Rules

9. Example of Apriori Algorithm

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

Assume minimum support = 40%.

Since there are 5 transactions:

$$\mathbf{40\% \times 5 = 2}$$

Therefore, an itemset must occur in at least 2 transactions.

9.1 Candidate 1-Itemsets (C1)

Item	Count	Support
Bread	4	80%
Milk	4	80%
Butter	4	80%

All satisfy the minimum support, so they form L1.

9.2 Candidate 2-Itemsets (C2)

Itemset	Count	Support
Bread, Milk	3	60%
Bread, Butter	3	60%
Milk, Butter	3	60%

All satisfy minimum support, so they form L2.

9.3 Candidate 3-Itemset (C3)

Candidate:

{Bread, Milk, Butter}

It occurs in T3 and T5, so:

Support = 2/5 = 40%

Therefore, it satisfies the minimum support.

10. Generating Association Rules

From the frequent itemset {Bread, Milk, Butter}, possible rules include:

- Bread, Milk → Butter
- Bread, Butter → Milk
- Milk, Butter → Bread

For the rule Bread, Milk → Butter:

Confidence = Support(Bread, Milk, Butter) / Support(Bread, Milk)

= (2/5) / (3/5) = 2/3 = 66.67%

If the minimum confidence is 60%, this rule is accepted.

11. Pseudocode

Apriori(Database, Minimum Support)

1. Find candidate 1-itemsets C1
2. Calculate support of each itemset
3. Remove itemsets below minimum support
4. Store remaining itemsets as L1

5. For k = 2 until no frequent itemsets:
    a. Generate candidate itemsets Ck from L(k-1)
    b. Remove candidates having infrequent subsets
    c. Calculate support of candidates
    d. Remove candidates below minimum support
    e. Store remaining itemsets as Lk

6. Combine all frequent itemsets
7. Generate association rules
8. Calculate confidence of each rule
9. Select rules satisfying minimum confidence
10. Return frequent itemsets and rules

12. Advantages of Apriori

- Simple and easy to understand.
- Easy to implement.
- Uses the Apriori principle for pruning.
- Useful for discovering patterns and relationships.
- Has applications in retail, marketing, healthcare and recommendation systems.

13. Limitations of Apriori

- Can be computationally expensive for large datasets.
- Requires repeated scanning of the transaction database.
- May generate a large number of candidate itemsets.
- Choosing suitable minimum support and confidence thresholds can be difficult.
- For some large datasets, methods such as FP-Growth can be more efficient.

14. Applications

- Market Basket Analysis – finding products frequently purchased together.
- Recommendation Systems – recommending products based on previous purchases.
- Retail – supporting product placement and promotions.
- Healthcare – discovering relationships among symptoms, diseases and treatments.
- Web Usage Mining – finding frequently occurring browsing patterns.
- Fraud Analysis – identifying unusual combinations of transaction attributes.

Conclusion

Apriori is a systematic and widely used approach for association rule mining. Its strength comes from pruning infrequent itemsets using the Apriori Principle. For a 16-mark answer, include the definition, Apriori Principle, support and confidence, algorithm steps, flowchart, numerical example, pseudocode, advantages, limitations and applications.