UNIT IV

1. What is the purpose of Apriori Algorithm?

Apriori algorithm is an influential algorithm for mining frequent item sets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties.

2. Define anti-monotone property?

If a set cannot pass a test, all of its supersets will fail the same test as well.

3. How to generate association rules from frequent item sets?

Association rules can be generated as follows

For each frequent item set1, generate all non empty subsets of 1.

For every non empty subsets s of 1, output the rule "S=>(1-s)"if

Support count(1)=min_conf,

Support_count(s)

where min_conf is the minimum confidence threshold.

4. Give few techniques to improve the efficiency of Apriori algorithm?

- Hash based technique
- Transaction Reduction
- Portioning
- Sampling
- Dynamic item counting

5. What are the things suffering the performance of Apriori candidate generation technique?

- Need to generate a huge number of candidate sets
- Need to repeatedly scan the scan the database and check a large set of candidates by pattern matching

6. Describe the method of generating frequent item sets without candidate generation?

Frequent-pattern growth(or FP Growth) adopts divide-and-conquer strategy.

Steps:

- Compress the database representing frequent items into a frequent pattern tree or FP tree,

- Divide the compressed database into a set of conditional database,
- Mine each conditional database separately.

7. Mention few approaches to mining Multilevel Association Rules?

• Uniform minimum support for all levels(or uniform support)

• Using reduced minimum support at lower levels(or reduced support)

• Level-by-level independent

• Level-cross filtering by single item

• Level-cross filtering by k-item set

8. What are multidimensional association rules?

Association rules that involve two or more dimensions or predicates

- **Inter dimension association rule:** Multidimensional association rule with no repeated predicate or dimension.
- **Hybrid-dimension association rule:** Multidimensional association rule with multiple occurrences of some predicates or dimensions.

9. Define constraint-Based Association Mining?

Mining is performed under the guidance of various kinds of constraints provided by the user. The constraints include the following

• Knowledge type constraints

• Data constraints

• Dimension/level constraints

• Interestingness constraints

• Rule constraints.

10. Define the concept of classification?

Two step process

• A model is built describing a predefined set of data classes or concepts.

• The model is constructed by analyzing database tuples described by attributes.The model is used for classification.

11 What is Decision tree?

A decision tree is a flow chart like tree structures, where each internal node denotes a test on an attribute, each branch represents an outcome of the test,and leaf nodes represent classes or class distributions. The top most in a tree is the root node.

12. What is Attribute Selection Measure?

The information Gain measure is used to select the test attribute at each node in the decision tree. Such a measure is referred to as an attribute selection measure or a measure of the goodness of split.

13. Describe Tree pruning methods.

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outlier. Tree pruning methods address this problem of over fitting the data.

Approaches:

- Pre pruning
- Post pruning

14. Define Pre Pruning

A tree is pruned by halting its construction early. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset samples.

15. Define Post Pruning.

Post pruning removes branches from a "Fully grown" tree. A tree node is pruned by removing its branches.

Eg: Cost Complexity Algorithm

16. What is meant by Pattern?

Pattern represents the knowledge.

17. Define the concept of prediction.

Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled sample or to assess the value or value ranges of an attribute that a given sample is likely to have.

18 What is the use of Regression?

Regression can be used to solve the classification problems but it can also be used for applications such as forecasting. Regression can be performed using many different types of techniques; in actually regression takes a set of data and fits the data to a formula.

19 What are the requirements of cluster analysis?

The basic requirements of cluster analysis are

- Dealing with different types of attributes.
- Dealing with noisy data.
- Constraints on clustering.
- Dealing with arbitrary shapes.
- High dimensionality
- Ordering of input data
- Interpretability and usability
- Determining input parameter and
- Scalability

20.What are the different types of data used for cluster analysis?

The different types of data used for cluster analysis are interval scaled, binary, nominal, ordinal and ratio scaled data.