UNIT V

1. Define Clustering?

Clustering is a process of grouping the physical or conceptual data object into clusters.

2. What do you mean by Cluster Analysis?

A cluster analysis is the process of analyzing the various clusters to organize the different objects into meaningful and descriptive objects.

3. What are the fields in which clustering techniques are used?

• Clustering is used in biology to develop new plants and animal taxonomies.

• Clustering is used in business to enable marketers to develop new distinct groups of their customers and characterize the customer group on basis of purchasing.

• Clustering is used in the identification of groups of automobiles Insurance policy customer.

• Clustering is used in the identification of groups of house in a city on the basis of house type, their cost and geographical location.• Clustering is used to classify the document on the web for information discovery.

4.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

5.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.

6. What are interval scaled variables?

Interval scaled variables are continuous measurements of linear scale. For Example , height and weight, weather temperature or coordinates for any cluster. These measurements can be calculated using Euclidean distance or Minkowski distance.

7. Define Binary variables? And what are the two types of binary variables?

Binary variables are understood by two states 0 and 1, when state is 0, variable is absent and when state is 1, variable is present. There are two types of binary variables, symmetric and asymmetric binary variables. Symmetric variables are those variables that have same state values and weights. Asymmetric variables are those variables that have not same state values and weights.

8. Define nominal, ordinal and ratio scaled variables?

A nominal variable is a generalization of the binary variable. Nominal variable has more than two states, For example, a nominal variable, color consists of four states, red, green, yellow, or black. In Nominal variables the total number of states is N and it is denoted by letters, symbols or integers. An ordinal variable also has more than two states but all these states are ordered in a meaningful sequence. A ratio scaled variable makes positive measurements on a non-linear scale, such as exponential scale, using the formula AeBt or Ae-Bt Where A and B are constants.

9. What do you mean by partitioning method?

In partitioning method a partitioning algorithm arranges all the objects intovarious partitions, where the total number of partitions is less than the total number of objects. Here each partition represents a cluster. The two types of partitioning method are k-means and k-medoids.

10. Define CLARA and CLARANS?

Clustering in LARge Applications is called as CLARA. The efficiency of CLARA depends upon the size of the representative data set. CLARA does not work properly if any representative data set from the selected representative data sets does not find best k-medoids. To recover this drawback a new algorithm, Clustering Large Applications based upon RANdomized search

(CLARANS) is introduced. The CLARANS works like CLARA, the only difference between CLARA and CLARANS is the clustering process that is done after selecting the representative data sets.

11. What is Hierarchical method?

Hierarchical method groups all the objects into a tree of clusters that are arranged in a hierarchical order. This method works on bottom-up or top-down approaches.

12. Differentiate Agglomerative and Divisive Hierarchical Clustering?

Agglomerative Hierarchical clustering method works on the bottom-up approach.In Agglomerative hierarchical method, each object creates its own clusters. The single Clusters are merged to make larger clusters and the process of merging continues until all the singular clusters are merged into one big cluster that consists of all the objects.

Divisive Hierarchical clustering method works on the top-down approach. In this method all the objects are arranged within a big singular cluster and the large cluster is continuously divided into smaller clusters until each cluster has a single object.

13. What is CURE?

Clustering Using Representatives is called as CURE. The clustering algorithms generally work on spherical and similar size clusters. CURE overcomes the problem of spherical and similar size cluster and is more robust with respect to outliers.

14. Define Chameleon method?

Chameleon is another hierarchical clustering method that uses dynamic modeling. Chameleon is introduced to recover the drawbacks of CURE method. In this method two clusters are merged, if the interconnectivity between two clusters is greater than the interconnectivity between the objects within a cluster.

15. Define Density based method?

Density based method deals with arbitrary shaped clusters. In density-based method, clusters are formed on the basis of the region where the density of the objects is high.

16. What is a DBSCAN?

Density Based Spatial Clustering of Application Noise is called as DBSCAN. DBSCAN is a density based clustering method that converts the high-density objects regions into clusters with arbitrary shapes and sizes. DBSCAN defines the cluster as a maximal set of density connected points.

17. What do you mean by Grid Based Method?

In this method objects are represented by the multi resolution grid data structure. All the objects are quantized into a finite number of cells and the collection of cells build the grid structure of objects. The clustering operations are performed on that grid structure. This method is widely used because its processing time is very fast and that is independent of number of objects.

18. What is a STING?

Statistical Information Grid is called as STING; it is a grid based multi resolution clustering method. In STING method, all the objects are contained into rectangular cells, these cells are kept into various levels of resolutions and these levels are arranged in a hierarchical structure.

19. Define Wave Cluster?

It is a grid based multi resolution clustering method. In this method all the objects are represented by a multidimensional grid structure and a wavelet transformation is applied for finding the dense region. Each grid cell contains the information of the group of objects that map into a cell. A wavelet transformation is a process of signaling that produces the signal of various frequency sub bands.

20. What is Model based method?

For optimizing a fit between a given data set and a mathematical model based methods are used. This method uses an assumption that the data are distributed by probability distributions. There are two basic approaches in this method that are

- Statistical Approach
- Neural Network Approach.

21. 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.

22. What are the reasons for not using the linear regression model to estimate the output data?

There are many reasons for that, One is that the data do not fit a linear model, It is possible however that the data generally do actually represent a linear model, but thelinear model generated is poor because noise or outliers exist in the data. Noise is erroneous data and outliers are data values that are exceptions to the usual and expected data.

23. What are the two approaches used by regression to perform classification?

Regression can be used to perform classification using the following approaches

- Division: The data are divided into regions based on class.
- Prediction: Formulas are generated to predict the output class value.

24. What do u mean by logistic regression?

Instead of fitting a data into a straight line logistic regression uses a logistic curve.
The formula for the univariate logistic curve is

P= e (C0+C1X1)
1+e (C0+C1X1)

The logistic curve gives a value between 0 and 1 so it can be interpreted as the probability of class membership.

25. What is Time Series Analysis?

A time series is a set of attribute values over a period of time. Time Series Analysis may be viewed as finding patterns in the data and predicting future values.

26. What are the various detected patterns?

Detected patterns may include:

- Trends: It may be viewed as systematic non-repetitive changes to the values over time.
- Cycles: The observed behavior is cyclic.
- Seasonal: The detected patterns may be based on time of year or month or day.
- Outliers: To assist in pattern detection , techniques may be needed to remove or reduce the impact of outliers.

27. What is Smoothing?

Smoothing is an approach that is used to remove the nonsystematic behaviors found in time series. It usually takes the form of finding moving averages of attribute values. It is used to filter out noise and outliers.