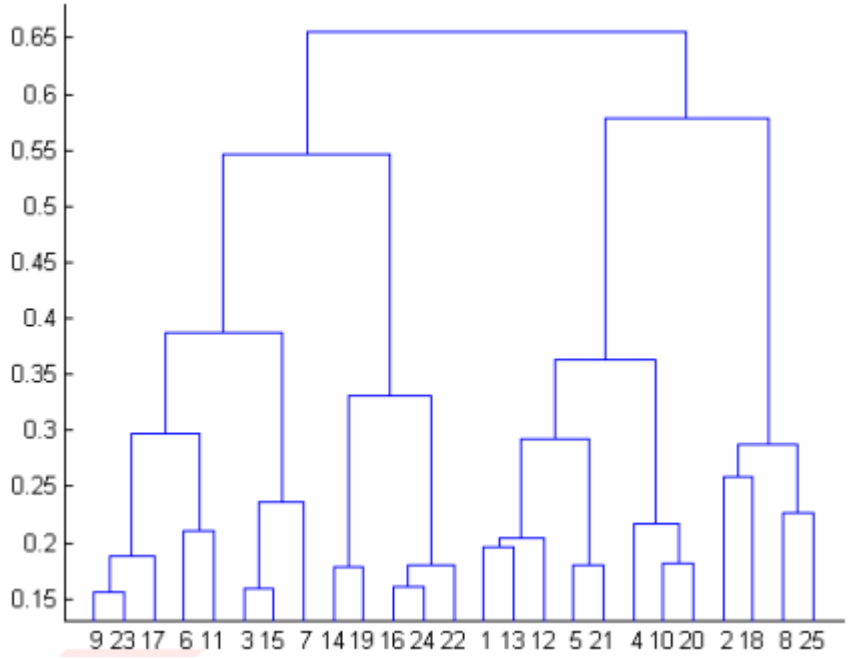
Assignment - 1

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



a) 2

b) 4

c) 6

d) 8

Correct answer (b) 4

2. In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers

2. Data points with different densities

3. Data points with round shapes

4. Data points with non-convex shapes

Options:

a) 1 and 2

b) 2 and 3

c) 2 and 4

d) 1, 2 and 4

Correct answer (d) 1,2 and 4

3. The most important part of is selecting the variables on which clustering is based.

a) interpreting and profiling clusters

b) selecting a clustering procedure

c) assessing the validity of clustering

d) formulating the clustering problem

Correct answer (d) formulating the clustering problem

4. The most commonly used measure of similarity is the or its square.

a) Euclidean distance

b) city-block distance

c) Chebyshev's distance

d) Manhattan distance

Correct answer (a) Euclidean distance

5. _____ is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

a) Non-hierarchical clustering

b) Divisive clustering

c) Agglomerative clustering

d) K-means clustering

Correct answer (b) Divisive clustering

6. Which of the following is required by K-means clustering?

a) Defined distance metric

b) Number of clusters

c) Initial guess as to cluster centroids

d) All answers are correct

Correct answer (d) All answers are correct

7. The goal of clustering is to_

a) Divide the data points into groups

b) Classify the data point into different classes

c) Predict the output values of input data points

d) All of the above

Correct answer (a) Divide the data points into groups

8. Clustering is a

a) Supervised learning

b) Unsupervised learning

c) Reinforcement learning

d) None

Correct answer (b) Unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

a) K- Means clustering

b) Hierarchical clustering

c) Diverse clustering

d) All of the above

Correct answer (a) K-means clustering

10. Which version of the clustering algorithm is most sensitive to outliers?

a) Data points with outliers

b) Data points with different densities

c) Data points with non-convex shapes

d) All of the above

Correct answer (a) K-means clustering algorithm

11. Which of the following is a bad characteristic of a dataset for clustering analysis

a) Data points with outliers

b) Data points with different densities

c) Data points with non-convex shapes

d) All of the above

Correct answer (d) All of the above

12. For clustering, we do not require

a) Labeled data

b) Unlabeled data

c) Numerical data

d) Categorical data

Correct answer (a) Labeled data

13. How is cluster analysis calculated?

Answer

Cluster analysis is a type of data analysis that involves grouping similar objects or data points together based on their characteristics or attributes. The goal of cluster analysis is to identify meaningful patterns and relationships within a dataset.

The process of cluster analysis involves several steps:

1. Data preparation: The first step is to collect and prepare the data for analysis. This involves selecting the variables or features that will be used to group the data and cleaning or transforming the data as necessary.

2. Similarity measurement: The next step is to measure the similarity or distance between each pair of data points. This can be done using various methods depending on the type of data and the research question.

3. Clustering algorithm: Once the similarity matrix is created, a clustering algorithm is used to group the data points into clusters. There are many different clustering algorithms, each with its own strengths and weaknesses. Some of the most commonly used algorithms include K-means clustering, hierarchical clustering, and DBSCAN.

4. Evaluation: Finally, the clusters are evaluated to determine their quality and whether they make sense

in the context of the research question. This may involve visualizing the clusters, comparing them to external criteria, or using statistical measures to assess their validity.

Overall, cluster analysis is a powerful tool for uncovering patterns and relationships in complex datasets, and it can be used in a wide range of fields, including biology, psychology, marketing, and finance.

14. How is cluster quality measured?

Answer

Cluster quality can be measured using several metrics that help evaluate the effectiveness of the clustering algorithm and the quality of the resulting clusters. Here are some common metrics used to measure cluster quality:

Cohesion: This measures how closely related the objects within each cluster are. Cohesion is calculated as the average distance between all pairs of objects within a cluster. The lower the cohesion value, the more tightly the objects are clustered together.

Separation: This measures how distinct the clusters are from each other. Separation is calculated as the average distance between all pairs of objects from different clusters. The higher the separation value, the more distinct the clusters are.

Silhouette score: This is a measure of how well each object fits within its assigned cluster compared to other nearby clusters. The silhouette score ranges from -1 to 1, with higher values indicating better clustering. A score of 0 means that the object is equally close to two or more clusters.

Cluster validity index: This is a measure that combines cohesion and separation to evaluate the overall quality of the clustering solution. Some commonly used cluster validity indices include the Dunn index, the Davies-Bouldin index, and the Calinski-Harabasz index.

Visual inspection: Finally, cluster quality can also be evaluated by visually inspecting the clusters to see if they make sense and align with the researcher's expectations or domain knowledge.

Overall, by using these metrics, researchers can evaluate the quality of the clustering solution and make informed decisions about which algorithm and parameter settings to use for future analyses.

15. What is cluster analysis and its types?

Answer

Cluster analysis is a statistical technique used to group similar objects or data points together based on their characteristics or attributes. It is a type of unsupervised learning, which means that it is used to discover patterns in data without being explicitly told what to look for.

There are several types of cluster analysis:

Hierarchical clustering: This involves grouping objects into a tree-like structure, with similar objects grouped together at each level of the tree. Hierarchical clustering can be either agglomerative (starting

with each object in its own cluster and progressively merging clusters) or divisive (starting with all objects in a single cluster and recursively splitting them into smaller clusters).

K-means clustering: This involves dividing objects into a predetermined number of clusters, with each object assigned to the cluster with the nearest mean value. K-means clustering is an iterative process, with the mean value of each cluster being recalculated after each iteration.

Density-based clustering: This involves identifying areas of high object density and grouping objects within those areas together. Density-based clustering is particularly useful for identifying clusters of irregular shapes or sizes.

Model-based clustering: This involves using statistical models to identify the best-fitting number and shape of clusters for a given dataset. Model-based clustering is particularly useful for datasets with complex patterns or overlapping clusters.

Overall, cluster analysis is a powerful tool for discovering patterns and relationships in complex datasets, and it can be used in a wide range of fields, including biology, psychology, marketing, and finance.