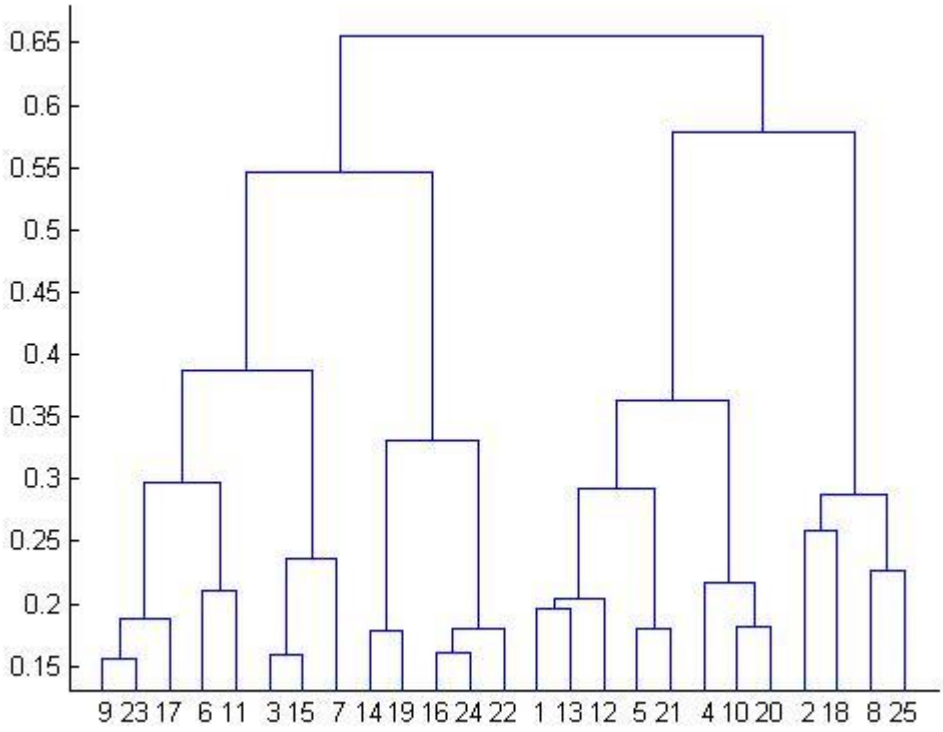
MACHINE LEARNING – WORKSHEET (CLUSTERING)

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.
Note- Answers are marked in green color.

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

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

a. 1 and 2
b. 2 and 3
c. 2 and 4
d. 1, 2 and 4
e. 1, 2, 3 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

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

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

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

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

8. Clustering is a-
   a. Supervised learning
   b. Unsupervised learning
   c. Reinforcement learning
   d. None

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

10. Which version of the clustering algorithm is most sensitive to outliers?
   a. K-means clustering algorithm
   b. K-modes clustering algorithm

c. K-medians clustering algorithm
d. None

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

12. For clustering, we do not require-
a. Labeled data
b. Unlabeled data
c. Numerical data
d. Categorical data
e.

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.

13. How is cluster analysis calculated?

Ans: Determining the **optimal number of clusters** in a data set is a fundamental issue in partitioning clustering, such as k-means clustering, which requires the user to specify the number of clusters k to be generated.
Unfortunately, there is no definitive answer to this question. The optimal number of clusters is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning.
A simple and popular solution consists of inspecting the dendrogram produced using hierarchical clustering to see if it suggests a particular number of clusters. Unfortunately, this approach is also subjective.
These methods include direct methods and statistical testing methods:

1. Direct methods: consists of optimizing a criterion, such as the within cluster sums of squares or the average silhouette. The corresponding methods are named *elbow* and *silhouette* methods, respectively.
2. Statistical testing methods: consists of comparing evidence against null hypothesis. An example is the *gap statistic*.

14. How is cluster quality measured?

Ans: To measure a cluster's fitness within a clustering, we can compute the average silhouette coefficient value of all objects in the cluster. To measure the quality of a clustering, we can use the average silhouette coefficient value of all objects in the data set. The silhouette coefficient and other intrinsic measures can also be used in the elbow method to heuristically derive the number of clusters in a data set by replacing the sum of within-cluster variances.
We have a few methods to choose from for measuring the quality of a clustering. In general, these methods can be categorized into two groups according to whether ground truth is available.
Here, *ground truth* is the ideal clustering that is often built using human experts.
If ground truth is available, it can be used by **extrinsic methods**, which compare the clustering against the group truth and measure. If the ground truth is unavailable, we can use **intrinsic methods**, which evaluate the goodness of a clustering by considering how well the clusters are separated.

Ground truth can be considered as supervision in the form of "cluster labels." Hence, extrinsic methods are also known as *supervised methods*, while intrinsic methods are *unsupervised methods*.

15. What is cluster analysis and its types?

Ans: **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including pattern recognition, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances between cluster members, dense areas of the data space, intervals or particular statistical distributions.

Types of cluster analysis

- Automatic Clustering Algorithms
- Balanced clustering
- Clustering high-dimensional data
- Conceptual clustering
- Consensus clustering
- Constrained clustering
- Community detection
- Data stream clustering
- HCS clustering
- Sequence clustering
- Spectral clustering