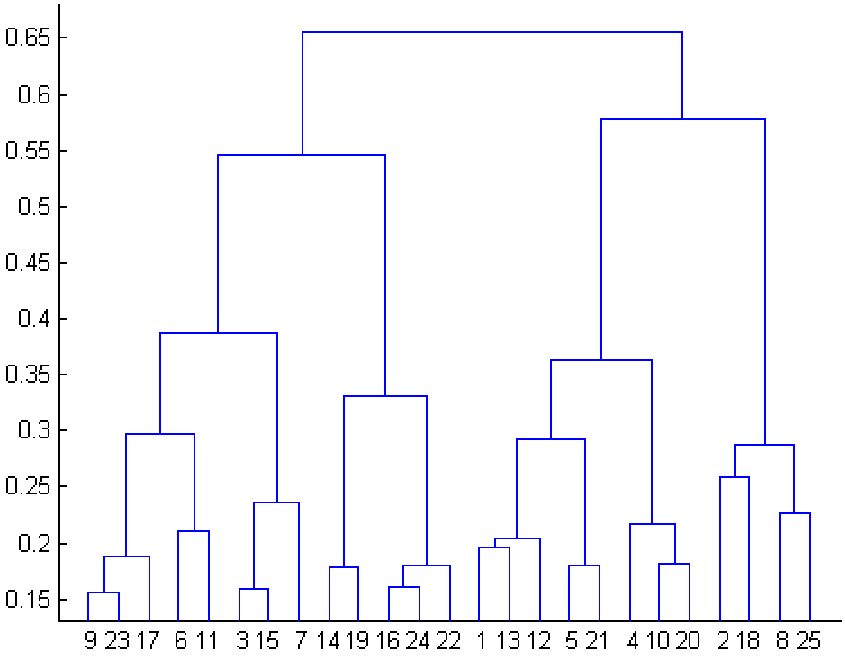
Worksheet-1 Machine Learning

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
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
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
Answer: d) formulating the clustering problem

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

a) Euclidean distance
b) city-block distance
c) Chebyshev's distance
d) Manhattan distance

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

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

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

Answer: d) All of the above

8. Clustering is a-

a) Supervised learning
b) Unsupervised learning
c) Reinforcement learning
d) None

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

Answer: a) K-Means clustering

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
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
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
Answer: a) Labeled data

13. How is cluster analysis calculated?

The objective of cluster analysis is to divide the observations into homogeneous and distinct groups. Following steps are followed to calculate cluster analysis:

- First, an initial partition with k clusters (given number of clusters) is created.
- Then, starting with the first object in the first cluster, Euclidean distances of all objects to all cluster foci are calculated.
- If an object is detected whose distance to the centre of gravity of the own cluster is greater than the distance to the centre of gravity (centroid) of another cluster, this object is shifted to the other cluster.
- Finally, the centroids of the two changed clusters are calculated again since the compositions have changed here.
- These steps are repeated until each object is in a cluster with the smallest distance to its centroid (centre of the cluster) (optimal solution)

14. How is cluster quality measured?

- Usually, clustering algorithms are non-supervised which often means that the correct partition into classes is not known and the number of classes is also unknown.
- In these cases, there are measures of cluster validity based in maximizing inter-cluster distances and minimizing intra-cluster distances. If the correct partition is known, then one can use a distance measure between partitions to measure how like the correct clustering is the clustering obtained by the algorithm.
- The choice of metric rather depends on what you consider the purpose of clustering to be.
- Here you have a couple of measures, but there are many more:

- **<u>SSE:</u>** sum of the square error from the items of each cluster.
- **<u>Inter cluster distance:</u>** sum of the square distance between each cluster centroid.
- **<u>Intra cluster distance for each cluster:</u>** sum of the square distance from the items of each cluster to its centroid.
- **<u>Maximum Radius:</u>** largest distance from an instance to its cluster centroid.
- **<u>Average Radius:</u>** sum of the largest distance from an instance to its cluster centroid divided by the number of clusters.

15. **What is cluster analysis and its types?**

Cluster analysis is a class of techniques that are used to classify objects or cases into relative groups called clusters. Cluster analysis is also called classification analysis or numerical taxonomy. In cluster analysis, there is no prior information about the group or cluster membership for any of the objects.

There are four types of cluster analysis:

1. **<u>Centroid Clustering</u>**

- In centroid cluster analysis you choose the number of clusters that you want to classify. For example, if you are a pet store owner you may choose to segment your customer list by people who bought dog and/or cat products.
- The algorithm will start by randomly selecting centroids (cluster centres) to group the data points into the two pre-defined clusters. A line is then drawn separating the data points into the two clusters based on their proximity to the centroids. The algorithm will then reposition the centroid relative to all the points within each cluster. The centroids and points in a cluster will adjust through all iterations, resulting in optimized clusters. The result of this analysis is the segmentation of your data into the two clusters. In this example, the data set will be segmented into customers who are own dogs and cats.

2. **<u>Density Clustering</u>**

- Density clustering groups data points by how densely populated they are. To group closely related data points, this algorithm leverages the understanding that the more dense the data points...the more related they are. To determine this, the algorithm will select a random point then start measuring the distance between each point around it. For most density algorithms a predetermined distance between data points is selected to benchmark how closely points need to be to one another to be considered related.. Then, the algorithm will identify all other points that are within the allowed distance of relevance. This process will continue to iterate by selecting different random data points to start with until the best clusters can be identified.

3. **<u>Distribution Clustering</u>**

- Distribution clustering identifies the probability that a point belongs to a cluster. Around each possible centroid the algorithm defines the density distributions for each cluster, quantifying the probability of belonging based on those distributions the algorithm optimizes the characteristics of the distributions to best represent the data.
- These maps look a lot like targets at an archery range. If a data point hits the bull’s eye on the map, then the probability of that person/object belonging to that cluster is 100%. Each ring around the bull’s eye represents lessening percentage or certainty.

- Distribution clustering is a great technique to assign outliers to clusters, whereas density clustering will not assign an outlier to a cluster.

4. **Connectivity Clustering**

- Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its own cluster. The primary premise of this technique is that points closer to each other are more related. The iterative process of this algorithm is to continually incorporate a data point or group of data points with other data points and/or groups until all points are engulfed into one big cluster. The critical input for this type of algorithm is determining where to stop the grouping from getting bigger.