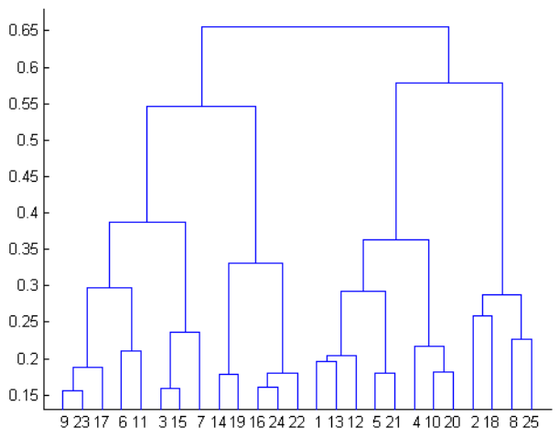
MACHINE LEARNING

ASSIGNMENT – 1

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

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 commonly 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: a) Divide the data points into groups

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: 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
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

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

13. How is cluster analysis calculated?

Answer: The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

14. How is cluster quality measured?

Answer: If the cluster's data items are all very similar to one another, the cluster is of high quality. In most cases, the Dissimilarity/Similarity metric can be used to assess the quality of clustering. If the clusters are similar, there are additional ways to assess the qualities of good clustering.

1. **Dissimilarity/Similarity metric:** The distance function, denoted by d, can be used to express the similarity between the clusters (i , j). Different data formats and data variables have different distance functions. For continuous-valued variables, categorical variables, and vector variables, the distance function measure varies. For various forms of data, the distance function can be expressed as Euclidean, Mahalanobis, or Cosine distance.

2. **Cluster completeness**: The crucial criterion is cluster completeness. If any two data objects have comparable qualities, they are placed in the same category of the cluster in accordance with the ground truth for effective clustering. If the objects belong to the same category, the cluster completion rate is high.
Let's take a look at the clustering C1, which has the sub-clusters s1 and s2, whose members, based on ground truth, are in the same category. Consider another clustering C2, which is the same as C1 but has merged s1 and s2 into a single cluster. After that, we construct the clustering quality measure, Q, and according to cluster completeness, Q(C2, Cg) > Q indicates that C2 has a higher cluster quality than C1 (C1, Cg).

3. Ragbag: In certain circumstances, there may be a few categories whose objects cannot be combined with those of other categories. The Rag Bag technique is then used to gauge the calibre of those cluster categories. We should classify the heterogeneous object into a rag bag category in accordance with the rag bag approach. Consider a clustering C1 and a cluster C C1 where, according to the ground truth, all items in C fall into the same category as the cluster C1 but for object o. Consider a clustering C2 that is similar to C1 but where o is assigned to a cluster D that contains objects from various categories. The reality is that the situation is noisy and the quality the rag bag criteria are used to gauge the degree of grouping. According to the rag bag technique criteria C2, which we define as Q(C2, Cg)>Q, the clustering quality measure will have a higher cluster quality than the C1 criteria (C1, Cg).

4. **Small cluster preservation**: If a small category of clustering is further divided into smaller pieces, those smaller bits of cluster become noise to the overall clustering, making it challenging to separate that small category from the clustering. According to the small cluster preservation criterion, it is not advisable to divide a tiny category into pieces since the pieces of the clusters are distinct, thus lowering the quality of the clusters. Assume that cluster C1 has broken up into three clusters: C11 = d1,... dn, C12 = dn+1, and C13 = dn+2.Let's say that clustering C2 similarly breaks up into three clusters, with C1 being C2 and C3 being dn+1 and dn+2. The clustering quality measure Q should give C2 a better score since C1 splits the little category of items whereas C2 splits the big category, which is preferable in accordance with the criterion given above. This is because Q(C2, Cg) > Q. (C1, Cg).

15. What is cluster analysis and its types?

Answer: Clustering is similar to classification, but the basis is different. In Clustering we don't know what you are looking for, and we are trying to identify some segments or clusters in our data. When we use clustering algorithms on our dataset, unexpected things can suddenly pop up like structures, clusters and groupings we would have never thought of otherwise.

TYPES OF CLUSTERING

1. K-Means Clustering
2. Hierarchical Clustering
3. Distribution-based Clustering
4. Density-based Clustering

1. **K-Means Clustering (Centroid-based Clustering):** Contrary to the hierarchical clustering described below, centroid-based clustering groups the data into non-hierarchical clusters. The most popular centroid-based clustering algorithm is k-means. Although efficient, centroid-based algorithms are sensitive to beginning conditions and outliers. K-means is the primary clustering algorithm covered in this course because it is straightforward, effective, and efficient.

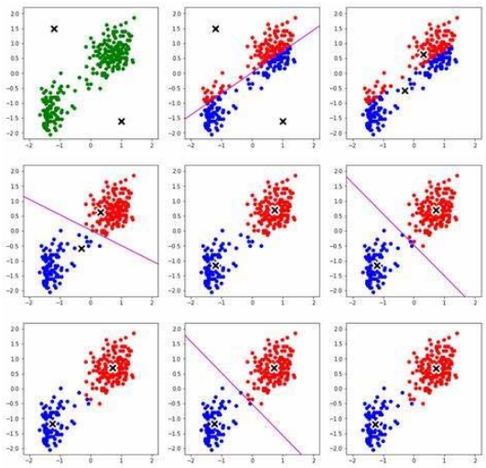


Fig: K-Means Clustering (Centroid-based Clustering):

2. **Hierarchical Clustering:** This approach involves first creating a cluster, which is then combined with another cluster (the most comparable and nearby one) to create a single cluster. Until every subject is in a single cluster, this process is repeated. Agglomerative method is the name of this specific approach. Agglomerative clustering begins by collecting individual objects into clusters. Another type of Hierarchical technique is the divisive method, which begins clustering with the entire data set before breaking it into partitions.

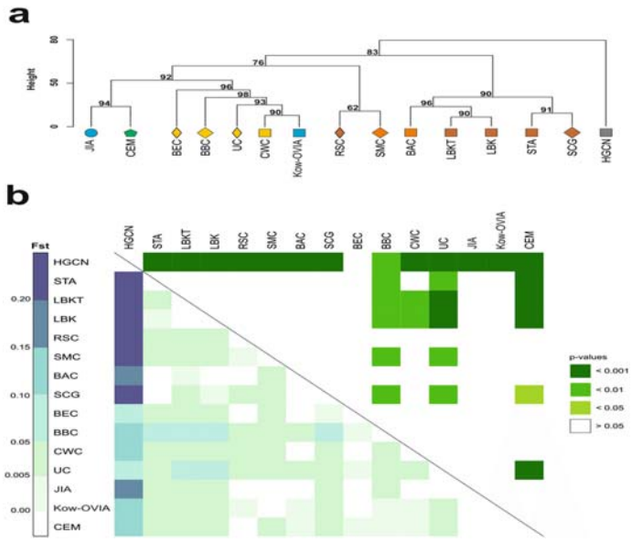


Fig: Hierarchical Clustering:

3. **Distribution-based Clustering:** It is a kind of clustering model that is strongly related to statistics that uses distribution modalities. A cluster is created by grouping together objects that fall within the same distribution. Some complicated characteristics of objects, such as correlation and reliance between qualities, can be captured by this kind of grouping.

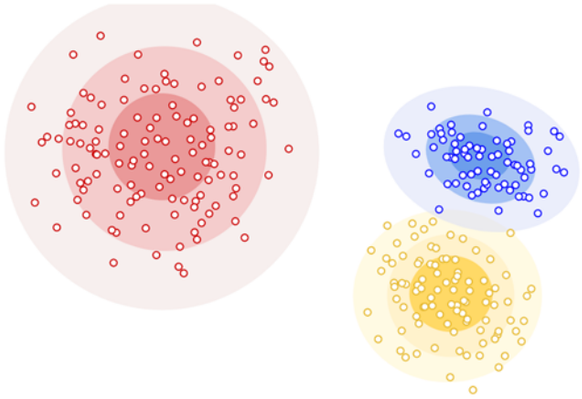


Fig: Distribution-based Clustering:

4. **Density-based Clustering**: Areas of higher density than the rest of the data set are what define clusters in this sort of clustering. Clusters are typically separated by objects in sparse areas. Typically, the graph's border points and noise are the items in these sparse spots DBSCAN is the most often used technique for this kind of clustering.