MACHINE LEARNING

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

Answer :- B

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:

Answer :- D

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

Answer :- A

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

Answer :- A

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.**

Answer :-C

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

Answer :-C

7. **The goal of clustering is to**

Answer :-D

8. **Clustering is a**

Answer :-D

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

Answer :-C

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

Answer :-A

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

Answer :-D

12. For clustering, we do not require

Answer :-A

13. **How is cluster analysis calculated?**

Answer :- Cluster analysis is an exploratory analysis that tries to identify structures within the data. Cluster analysis is also called segmentation analysis or taxonomy analysis. More specifically, it tries to identify homogenous groups of cases if the grouping is not previously known. Because it is exploratory, it does not make any distinction between dependent and independent variables. The different cluster analysis methods that SPSS offers can handle binary, nominal, ordinal, and scale (interval or ratio) data.

Cluster analysis is often used in conjunction with other analyses (such as discriminant analysis). The researcher must be able to interpret the cluster analysis based on their understanding of the data to determine if the results produced by the analysis are actually meaningful.

The hierarchical cluster analysis follows three basic steps: 1) calculate the distances, 2) link the clusters, and 3) choose a solution by selecting the right number of clusters.

First, we have to select the variables upon which we base our clusters. In the dialog window we add the math, reading, and writing tests to the list of variables. Since we want to cluster cases we leave the rest of the tick marks on the default.

In the dialog box Statistics… we can specify whether we want to output the proximity matrix (these are the distances calculated in the first step of the analysis) and the predicted cluster membership of the cases in our observations. Again, we leave all settings on default.

In the dialog box Plots… we should add the Dendrogram. The Dendrogram will graphically show how the clusters are merged and allows us to identify what the appropriate number of clusters is.

The dialog box Method… allows us to specify the distance measure and the clustering method. First, we need to define the correct distance measure. SPSS offers three large blocks of distance measures for interval (scale), counts (ordinal), and binary (nominal) data.

For interval data, the most common is Square Euclidian Distance. It is based on the Euclidian Distance between two observations, which is the square root of the sum of squared distances. Since the Euclidian Distance is squared, it increases the importance of large distances, while weakening the importance of small distances.

If we have ordinal data (counts) we can select between Chi-Square or a standardized Chi-Square called Phi-Square. For binary data, the Squared Euclidean Distance is commonly used.

In our example, we choose Interval and Square Euclidean Distance.

Next, we have to choose the Cluster Method. Typically, choices are between-groups linkage (distance between clusters is the average distance of all data points within these clusters), nearest neighbor (single linkage: distance between clusters is the smallest distance between two data points), furthest neighbor (complete linkage: distance is the largest distance between two data points), and Ward's method (distance is the distance of all clusters to the grand average of the sample). Single linkage works best with long chains of clusters, while complete linkage works best

with dense blobs of clusters. Between-groups linkage works with both cluster types. It is recommended is to use single linkage first. Although single linkage tends to create chains of c…

A last consideration is standardization. If the variables have different scales and means we might want to standardize either to Z scores or by centering the scale. We can also transform the values to absolute values if we have a data set where this might be appropriate.

14. **How is cluster quality measured?**

Answer :- If all the data objects in the cluster are highly similar then the cluster has high quality. We can measure the quality of Clustering by using the Dissimilarity/Similarity metric in most situations. But there are some other methods to measure the Qualities of Good Clustering if the clusters are alike.

1. **Dissimilarity/Similarity metric:** The similarity between the clusters can be expressed in terms of a distance function, which is represented by d(i, j). Distance functions are different for various data types and data variables. Distance function measure is different for continuous-valued variables, categorical variables, and vector variables. Distance function can be expressed as Euclidean distance, Mahalanobis distance, and Cosine distance for different types of data.

2. **Cluster completeness:** Cluster completeness is the essential parameter for good clustering, if any two data objects are having similar characteristics then they are assigned to the same category of the cluster according to ground truth. Cluster completeness is high if the objects are of the same category.

Let us consider the clustering C1, which contains the sub-clusters s1 and s2, where the members of the s1 and s2 cluster belong to the same category according to ground truth. Let us consider another clustering C2 which is identical to C1 but now s1 and s2 are merged into one cluster. Then, we define the clustering quality measure, Q, and according to cluster completeness C2, will have more cluster quality compared to the C1 that is, $Q(C2, Cg) > Q(C1, Cg)$.

3. **Ragbag:** In some situations, there can be a few categories in which the objects of those categories cannot be merged with other objects. Then the quality of those cluster categories is measured by the Rag Bag method. According to the rag bag method, we should put the heterogeneous object into a rag bag category.

Let us consider a clustering C1 and a cluster $C \in C1$ so that all objects in C belong to the same category of cluster C1 except the object o according to ground truth. Consider a clustering C2 which is identical to C1 except that o is assigned to a cluster D which holds the objects of different categories. According to the ground truth, this situation is noisy and the quality of clustering is measured using the rag bag criteria. we define the clustering quality measure, Q, and according to rag bag method criteria C2, will have more cluster quality compared to the C1 that is, $Q(C2, Cg) > Q(C1, Cg)$.

4. **Small cluster preservation:** If a small category of clustering is further split into small pieces, then those small pieces of cluster become noise to the entire clustering and thus it becomes difficult to identify that small category from the clustering. The small cluster preservation criterion states that are splitting a small category into pieces is not advisable and it further decreases the quality of clusters as the pieces of clusters are distinctive. Suppose clustering C1 has split into three clusters, C11 = {d1, . . . , dn}, C12 = {dn+1}, and C13 = {dn+2}.

Let clustering C2 also split into three clusters, namely C1 = {d1, . . . , dn−1}, C2 = {dn}, and C3 = {dn+1,dn+2}. As C1 splits the

small category of objects and C2 splits the big category which is preferred according to the rule mentioned above the clustering quality measure Q should give a higher score to C2, that is, Q(C2, Cg) > Q(C1, Cg).

15. What is cluster analysis and its types?

Answer :- Cluster analysis is a multivariate data mining technique whose goal is to groups objects (eg., products, respondents, or other entities) based on a set of user selected characteristics or attributes. It is the basic and most important step of data mining and a common technique for statistical data analysis, and it is used in many fields such as data compression, machine learning, pattern recognition, information retrieval etc.

Types of Cluster Analysis

The clustering algorithm needs to be chosen experimentally unless there is a mathematical reason to choose one cluster method over another.It should be noted that an algorithm that works on a particular set of data will not work on another set of data. There are a number of different methods to perform cluster analysis. Some of them are,

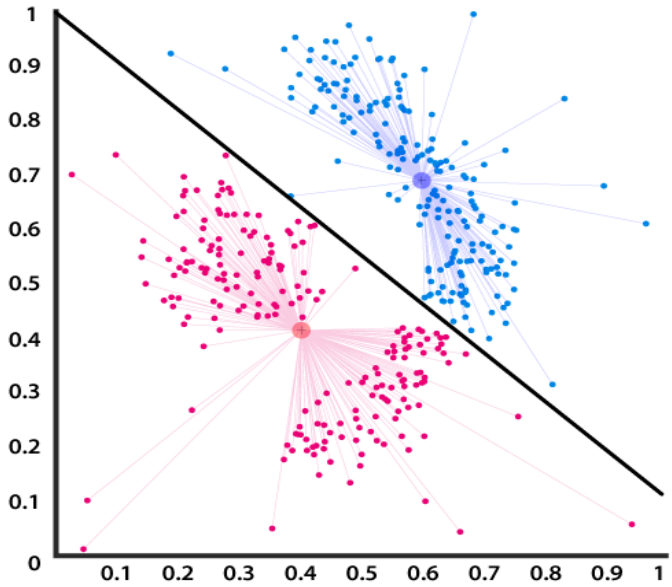
Hierarchical Cluster Analysis

In this method, first, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster. This process is repeated until all subjects are in one cluster. This particular method is known as **Agglomerative method**. Agglomerative clustering starts with single objects and starts grouping them into clusters.

The divisive method is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions.

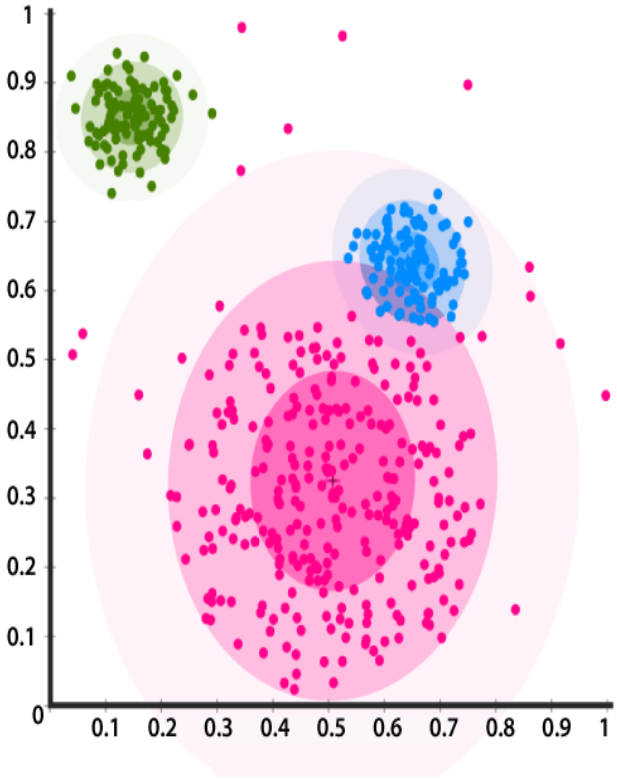
Centroid-based Clustering

In this type of clustering, clusters are represented by a central entity, which may or may not be a part of the given data set. K-Means method of clustering is used in this method, where k are the cluster centers and objects are assigned to the nearest cluster centres.



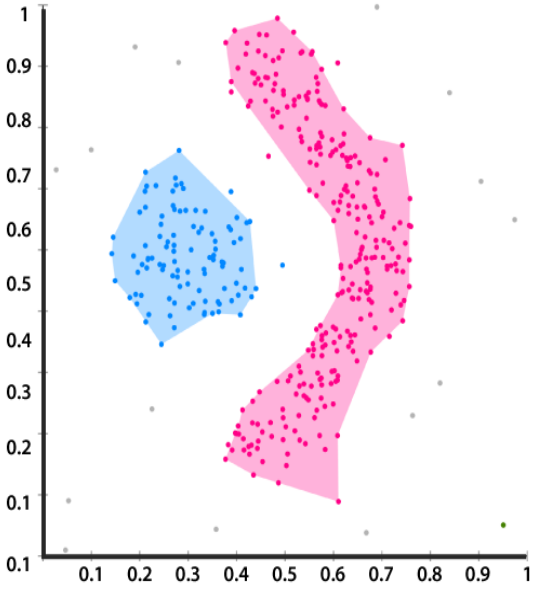
Distribution-based Clustering

It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster.This type of clustering can capture some complex properties of objects like correlation and dependence between attributes.



Density-based Clustering

In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters.The objects in these sparse points are usually noise and border points in the graph.The most popular method in this type of clustering is DBSCAN.


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1