MACHINE LEARNING

1. **(a)**
2. **(c)**
3. **(d)**
4. **(a)**
5. **(b)**
6. **(d)**
7. **(a)**
8. **(b)**
9. **(a)**

10 .(a)

11 .(d)

12 .(a)

13 (I).clustering analysis is calculated by assigning k centres randomly.

(II) Calculate the distance of all the points from all the k centres and allocate the points to cluster based on the shortest distance .The model's inertia is the mean squared distance between each instance and its closest centroid .The goal is to have a model with the lowest inertia.

(III) Once all the points are assigned to clusters, recompute the centroids.

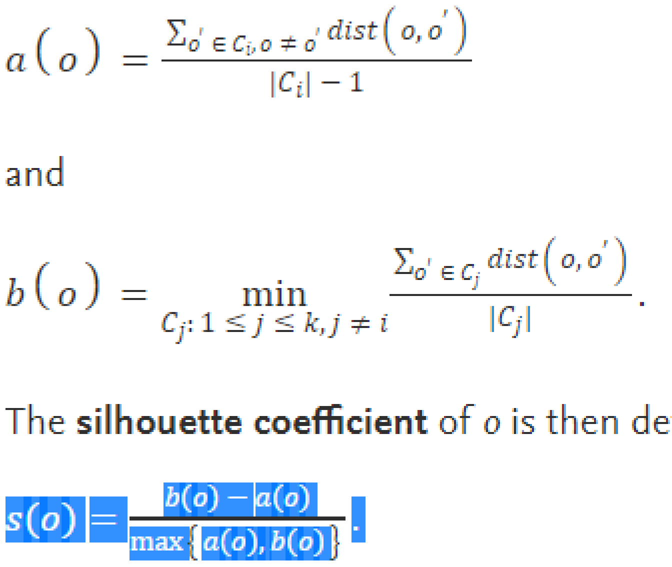
Repeat the steps (II) and (III) until the locations of the centroids stop changing and the cluster allocation of the points becomes constant.

14. Cluster quality is measured by examining how well the clusters are separated and hoe compact the cluster are .Many intrinsic methods have the advantage of a similarity metric between objects in the dataset.

The silhouette coefficient is a measure

Suppose D is a Dataset partitioned in to k clusters, C1,C2,…Ck .For each object oϵ D, we calculate a(o) as the average distance between o and all other objects in the cluster to which o belongs . Similarly, b(o) is the minimum average distance from o to all clusters to which o does not belong. Formally, suppose

o€Ci(1≤i≤k);then



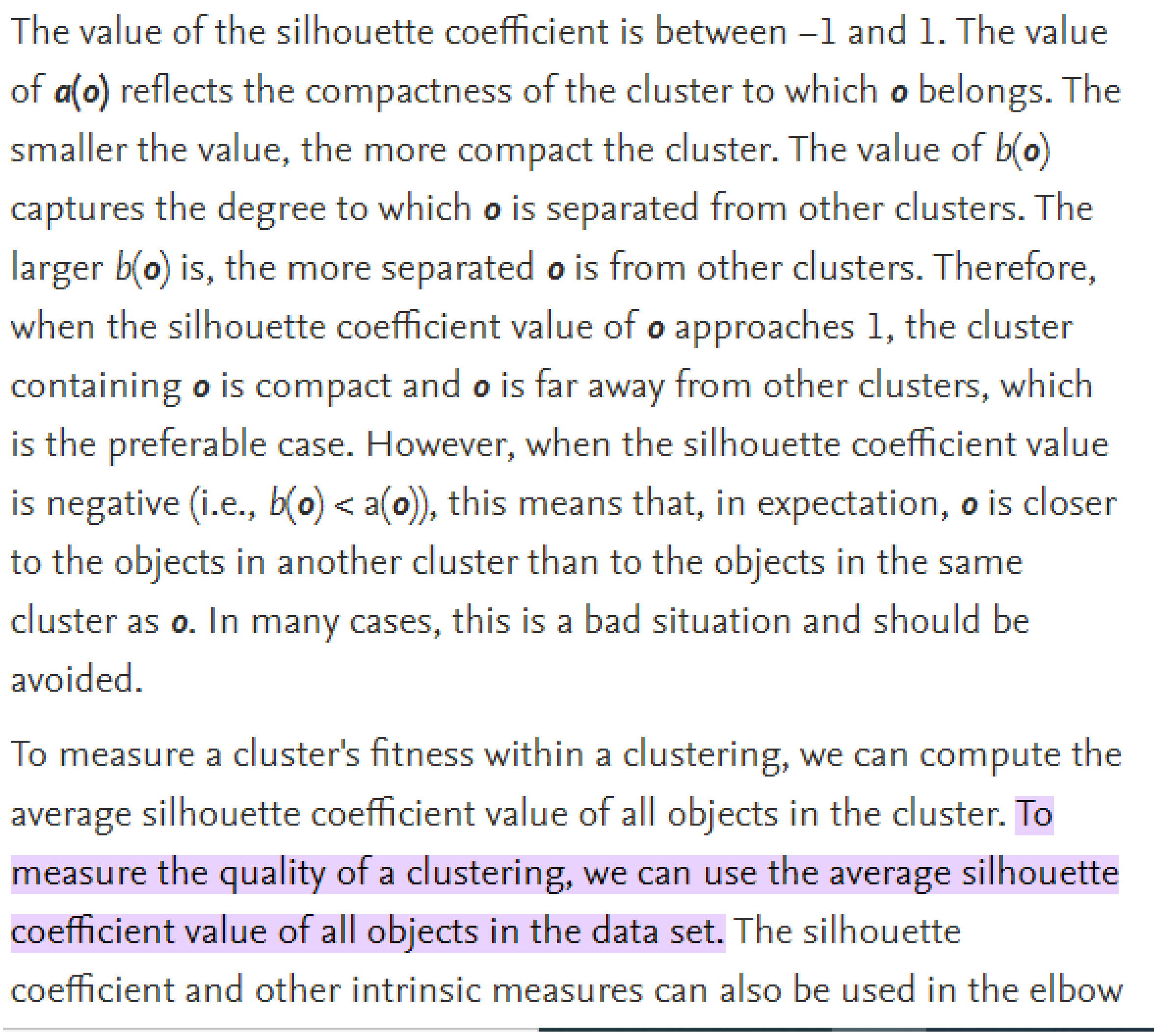
$$a(o) = \frac{\sum_{o' \in C_i, o \neq o'} dist(o, o')}{|C_i| - 1}$$

and

$$b(o) = \min_{C_j: 1 \leq j \leq k, j \neq i} \frac{\sum_{o' \in C_j} dist(o, o')}{|C_j|}.$$

The **silhouette coefficient** of o is then defined as

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}.$$

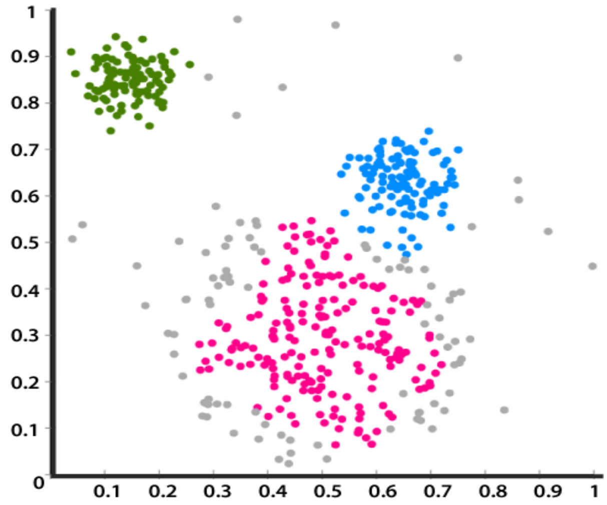


The value of the silhouette coefficient is between −1 and 1. The value of ***a***(***o***) reflects the compactness of the cluster to which ***o*** belongs. The smaller the value, the more compact the cluster. The value of *b*(***o***) captures the degree to which ***o*** is separated from other clusters. The larger *b*(***o***) is, the more separated ***o*** is from other clusters. Therefore, when the silhouette coefficient value of ***o*** approaches 1, the cluster containing ***o*** is compact and ***o*** is far away from other clusters, which is the preferable case. However, when the silhouette coefficient value is negative (i.e., *b*(***o***) < a(***o***)), this means that, in expectation, ***o*** is closer to the objects in another cluster than to the objects in the same cluster as ***o***. In many cases, this is a bad situation and should be avoided.

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 derive the no. of clusters in a data set by replacing the sum of within-cluster variances

15**.Cluster Analysis: 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 to each other than to those in other groups (clusters).It is a main task of exploratory data analysis ,and a common technique for statistical data analysis, used in many fields ,including pattern recognition ,image analysis ,computer graphics and machine learning.**



Clusters should exhibit high internal homogeneity and high external heterogeneity.

When plotted geometrically, objects with in clusters should be very close together and clusters will be far apart.

Types of cluster analysis:

Hierarchical cluster analysis:

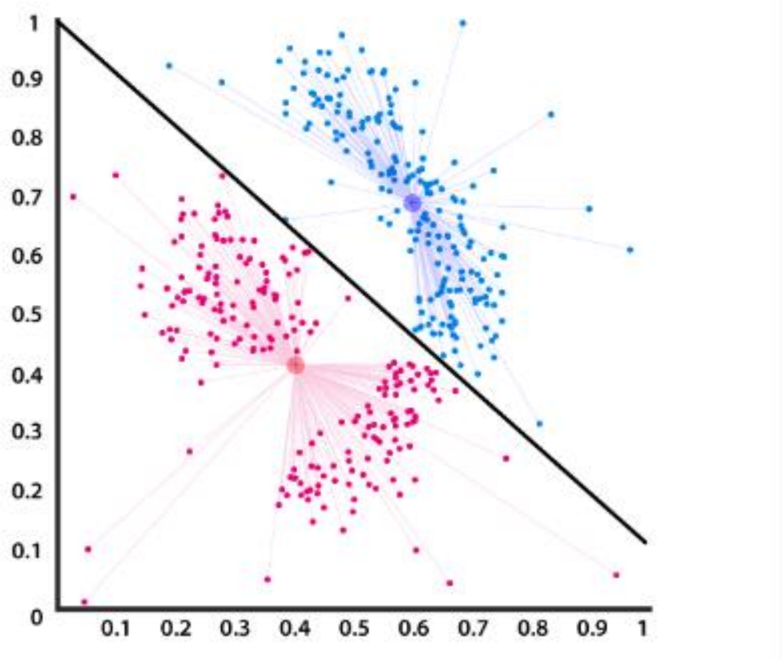
In this method, first a cluster is made and

Then added to another cluster 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 the complete data set and then starts with single objects and starts grouping them in to clusters.

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

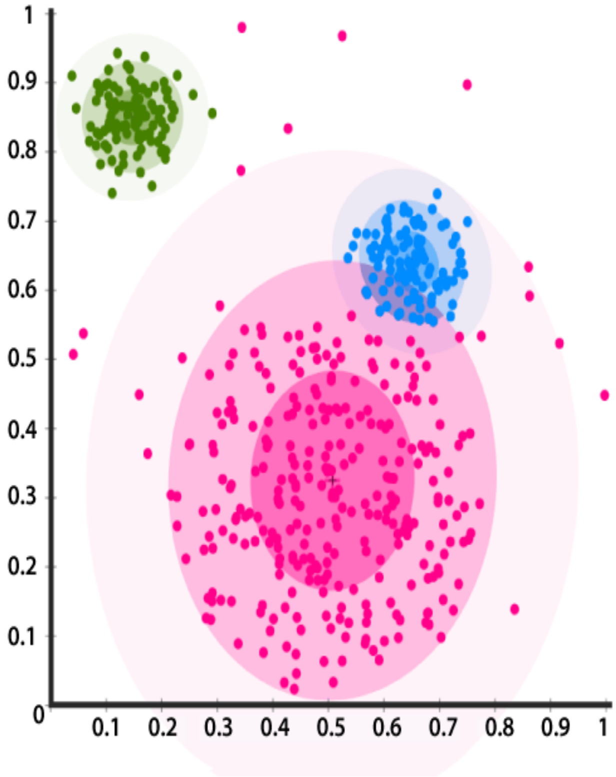
Centroid Based Clustering:

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. Dd



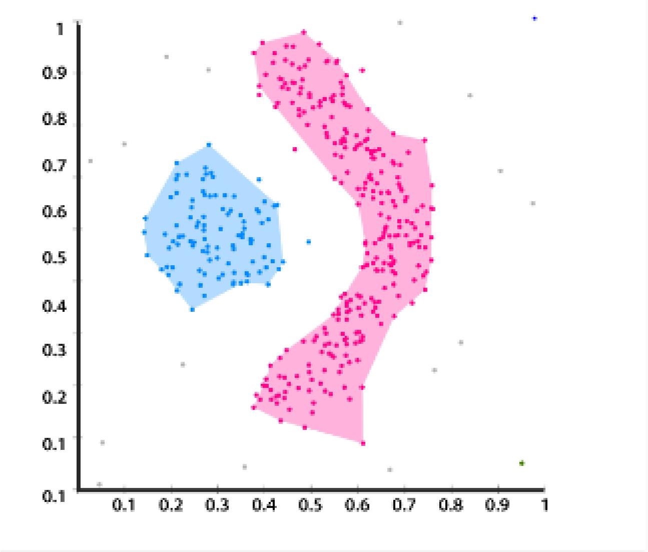
DDistribution-based Clustering

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