SNS COLLEGE OF ENGINEERING
(Autonomous)
DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

Artificial Intelligence & Machine Learning

Unit 3 – Unsupervised Learning

Hierarchical Clustering in Machine Learning

Prepared by,
P.Ramya
Assistant Professor/ECE
SNS College of Engineering

Hierarchical Clustering in Machine Learning

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the **dendrogram**.

2 Approaches

Agglomerative: Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

Divisive: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

Why hierarchical clustering?

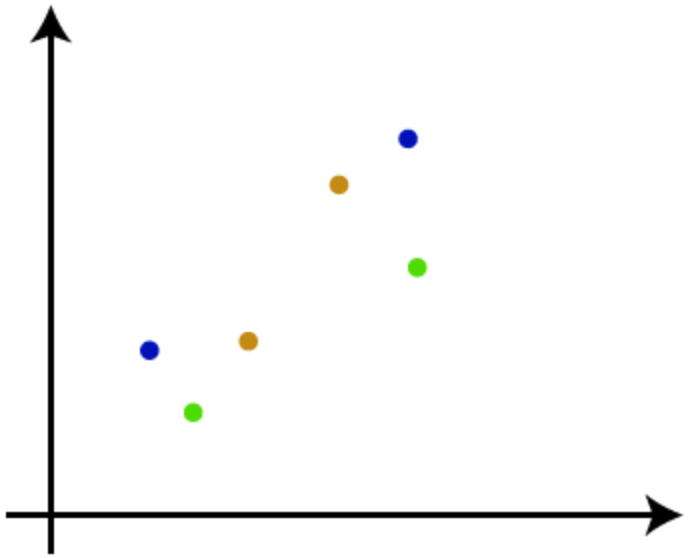
•As we already have other clustering algorithms such as K-Means Clustering, then why we need hierarchical clustering? So, as we have seen in the K-means clustering that there are some challenges with this algorithm, which are a predetermined number of clusters, and it always tries to create the clusters of the same size.

Agglomerative Hierarchical clustering

The agglomerative hierarchical clustering algorithm is a popular example of HCA. To group the datasets into clusters, it follows the bottom-up approach.

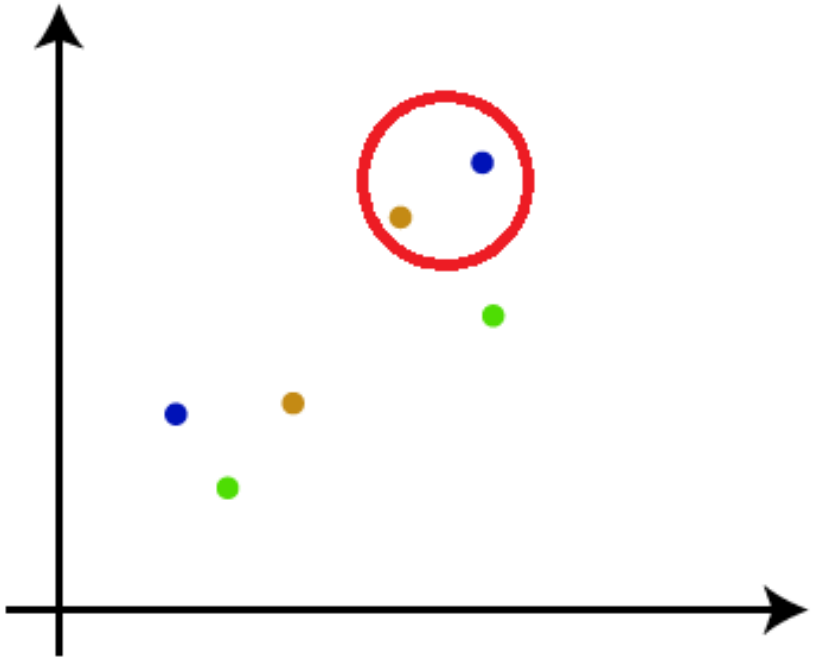
The working of the AHC algorithm can be explained using the below steps:

Step-1: Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.



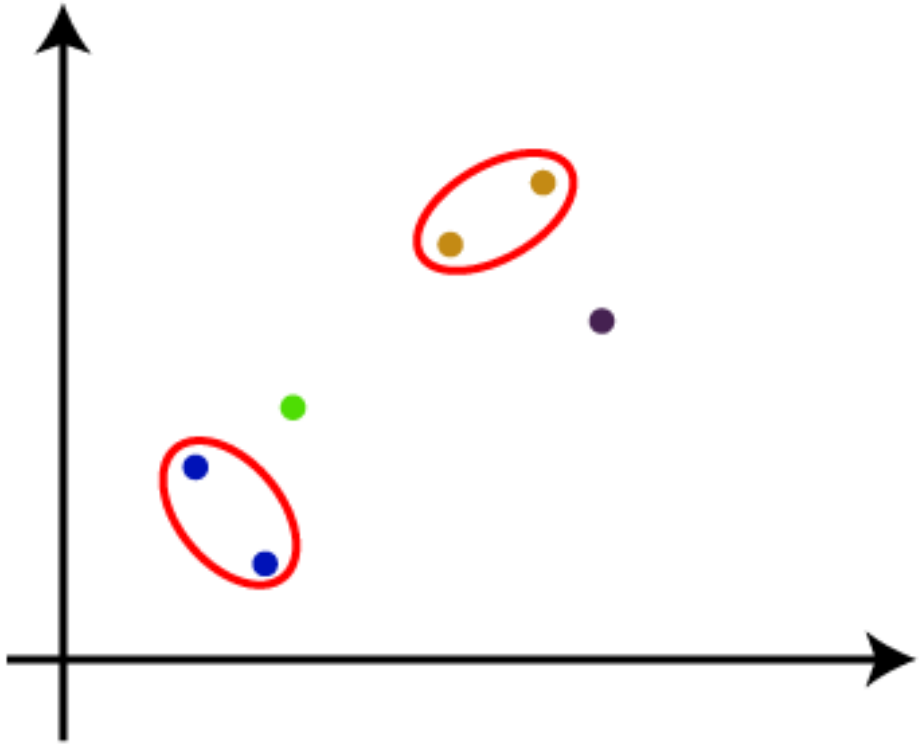
Contd…

Step-2: Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.



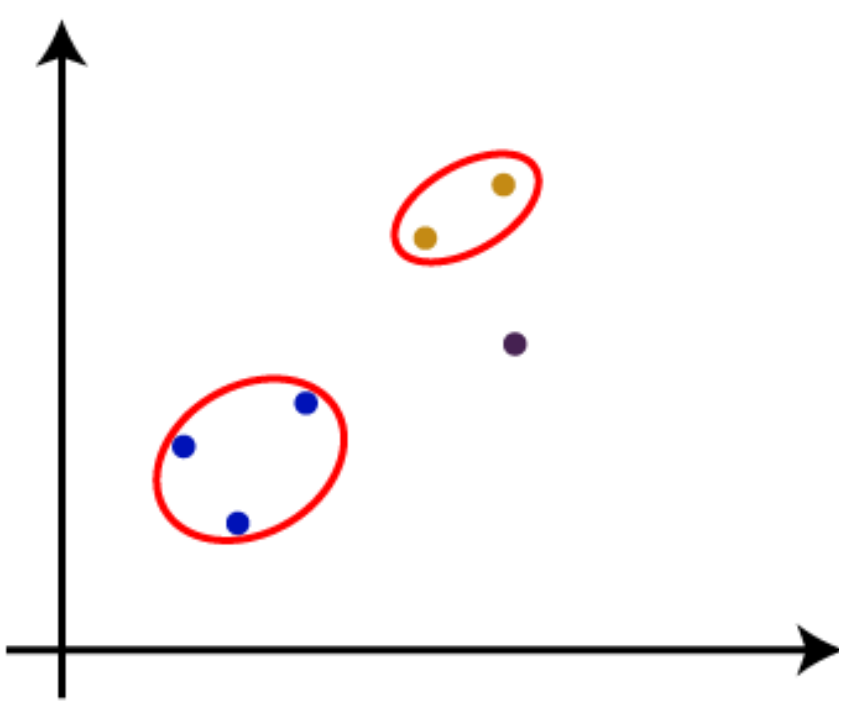
Contd…

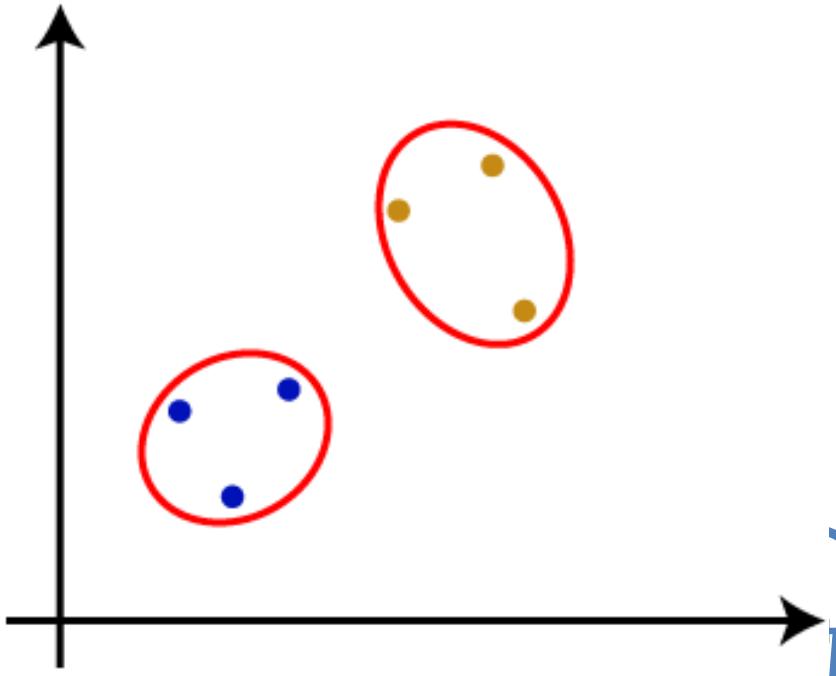
•**Step-3**: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.



Contd…

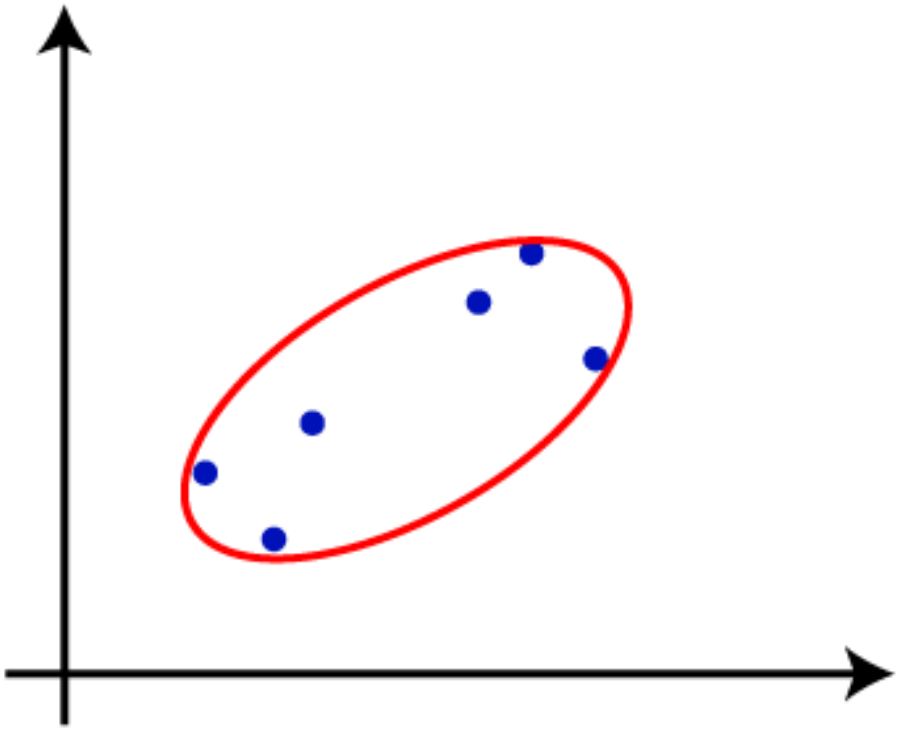
•**Step-4:** Repeat Step 3 until only one cluster left. So, we will get the following clusters. Consider the below images:





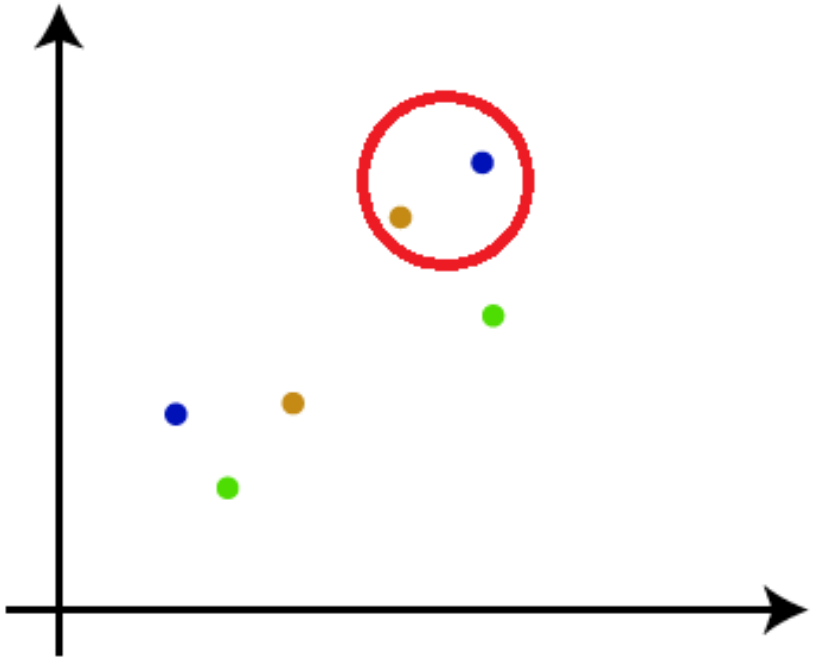
Contd…

•**Step-5:** Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.



Contd…

Step-2: Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.

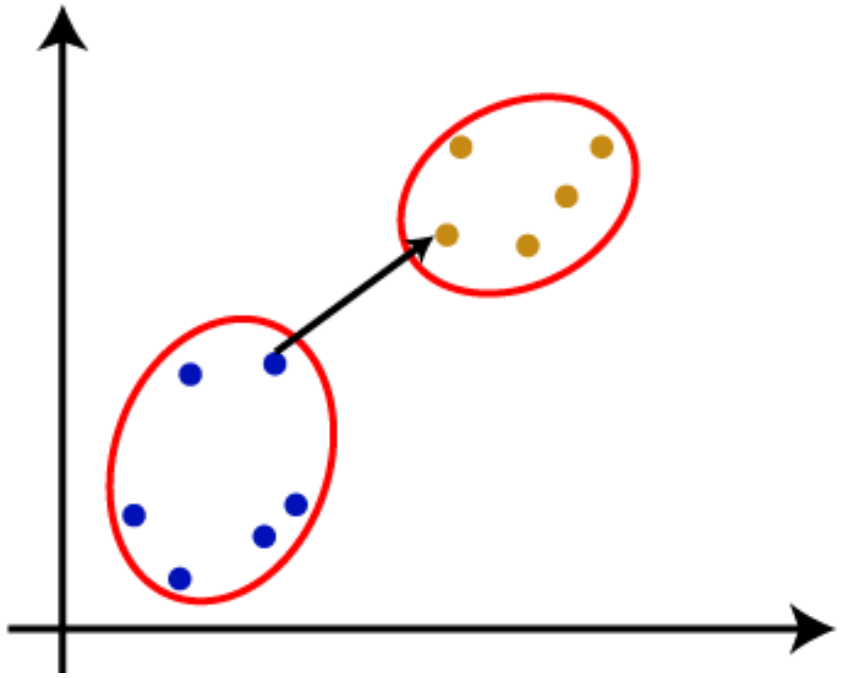


Measure for the distance between two clusters

As we have seen, the **closest distance** between the two clusters is crucial for the hierarchical clustering. There are various ways to calculate the distance between two clusters, and these ways decide the rule for clustering. These measures are called **Linkage methods**.

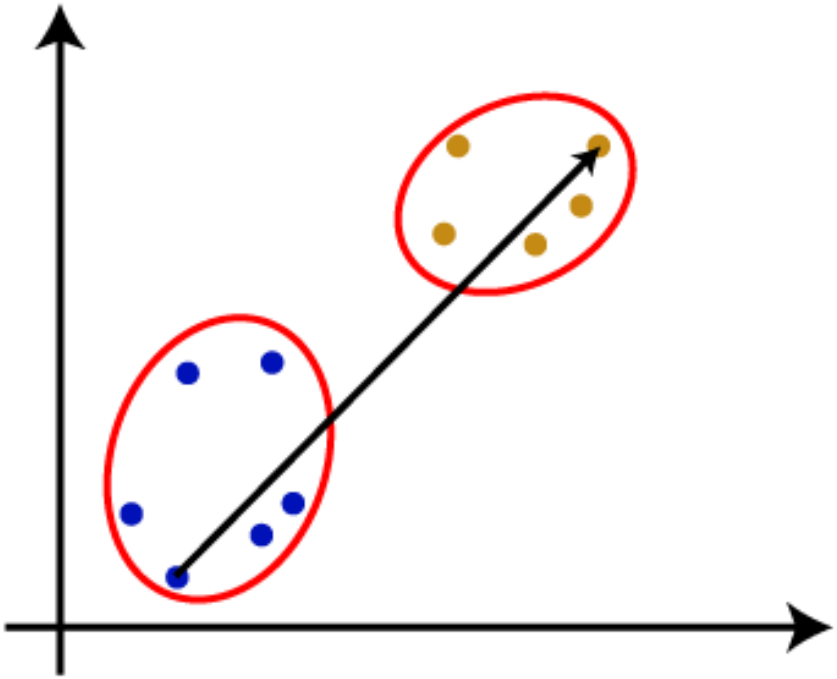
Contd…

Single Linkage: It is the Shortest Distance between the closest points of the clusters. Consider the below image:



Contd…

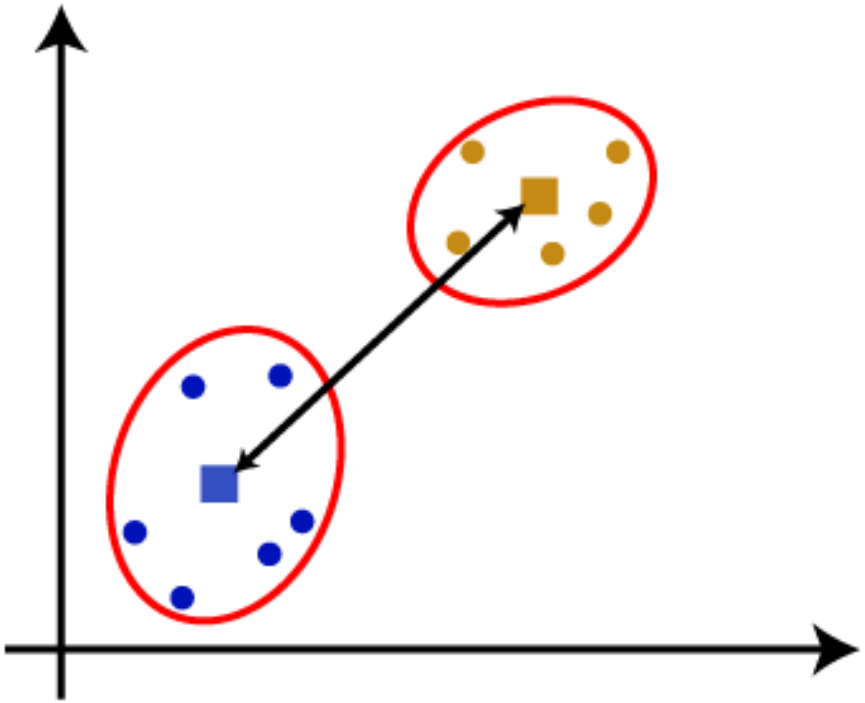
Complete Linkage: It is the farthest distance between the two points of two different clusters. It is one of the popular linkage methods as it forms tighter clusters than single-linkage.



Contd…

Average Linkage: It is the linkage method in which the distance between each pair of datasets is added up and then divided by the total number of datasets to calculate the average distance between two clusters. It is also one of the most popular linkage methods.

Centroid Linkage: It is the linkage method in which the distance between the centroid of the clusters is calculated. Consider the below image:




thank
you