ML

LECTURE-17

BY
Dr. Ramesh Kumar Thakur
Assistant Professor (II)
School Of Computer Engineering

Hierarchical clustering

- 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**.

- Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. **As there is no requirement to predetermine the number of clusters.**

- The hierarchical clustering technique has two 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.**

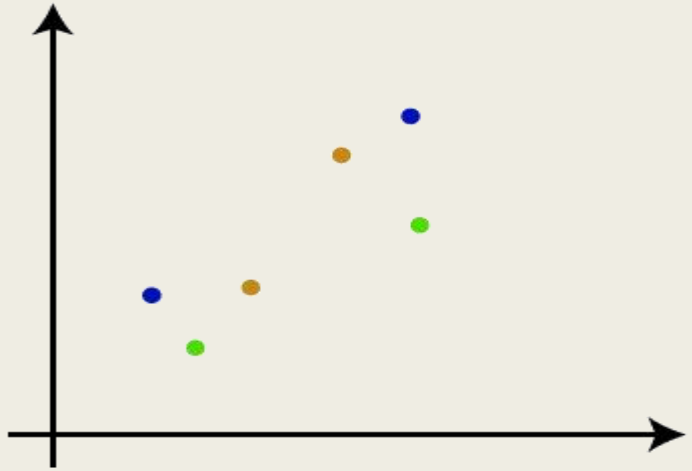
- **Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is a **top-down approach.**

Agglomerative Hierarchical clustering

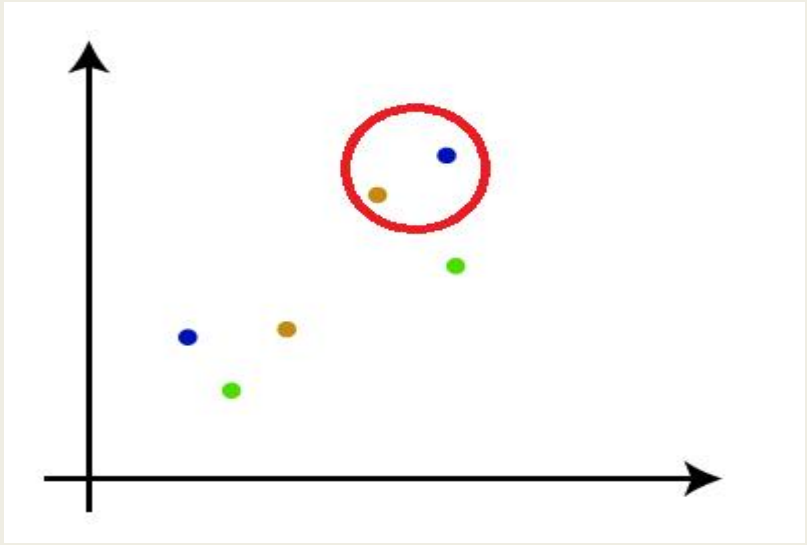
- The agglomerative hierarchical clustering algorithm is a popular example of HCA.
- To group the datasets into clusters, it follows the bottom-up approach.
- It means, this algorithm considers each dataset as a single cluster at the beginning, and then start combining the closest pair of clusters together.
- It does this until all the clusters are merged into a single cluster that contains all the datasets.
- This hierarchy of clusters is represented in the form of the dendrogram.

Steps of Agglomerative Hierarchical clustering

- 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.

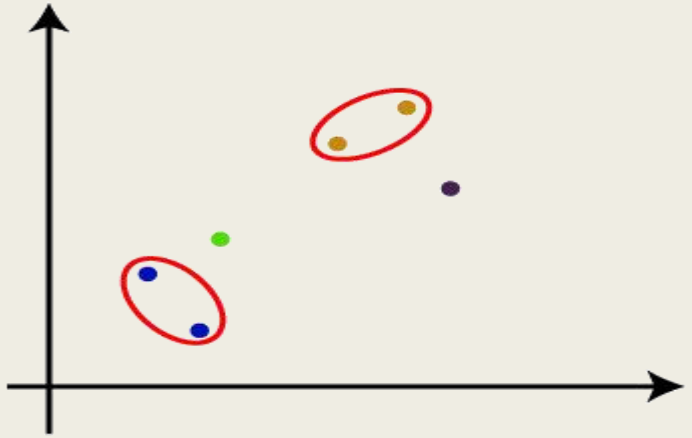


- Step-2: Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.

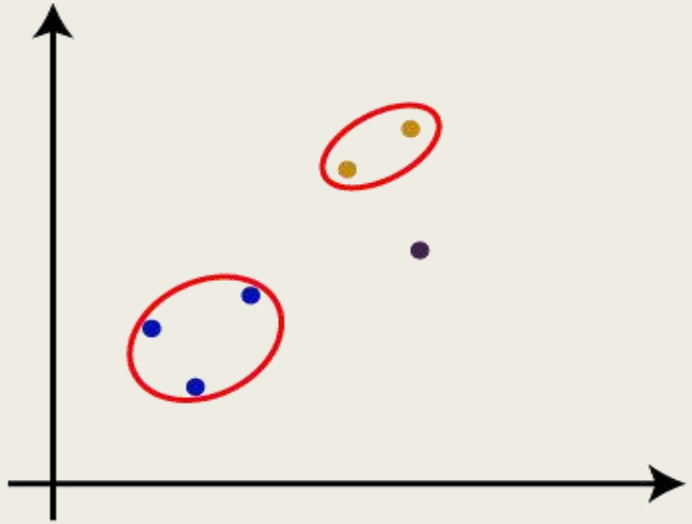


Steps of Agglomerative Hierarchical clustering

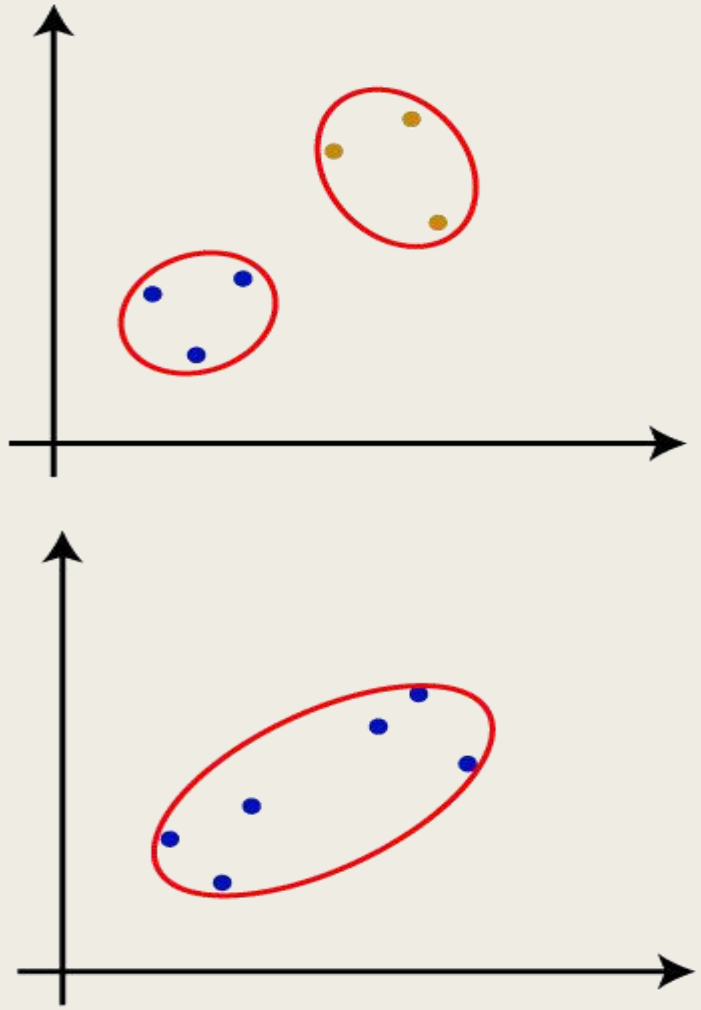
- Step-3: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.



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



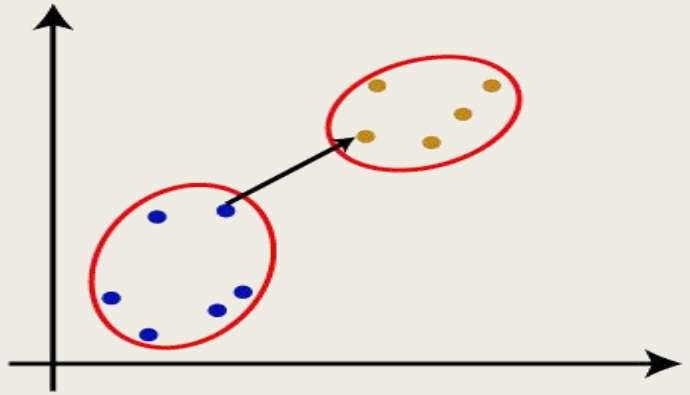
Steps of Agglomerative Hierarchical clustering



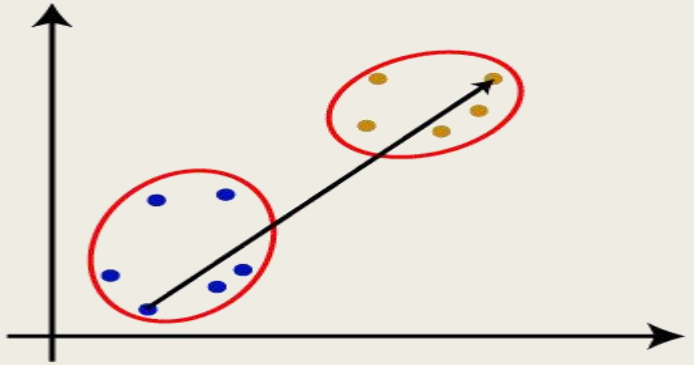
- Step-5: Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

Measure for the distance between two clusters

- 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**. Some of the popular linkage methods are given below:
- **Single Linkage:** It is the **Shortest Distance between the closest points of the clusters.** Consider the below image:

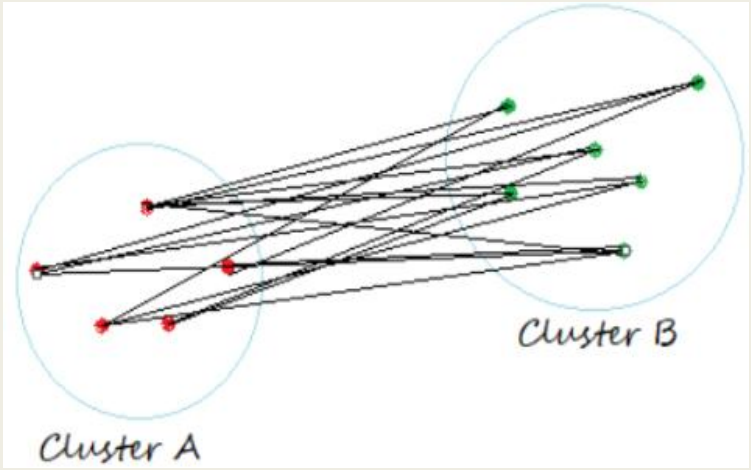


- **Complete Linkage:** It is the **farthest distance between the two points of two different clusters.** It is one of the popular linkage methods as **it forms tighter clusters than single-linkage.**

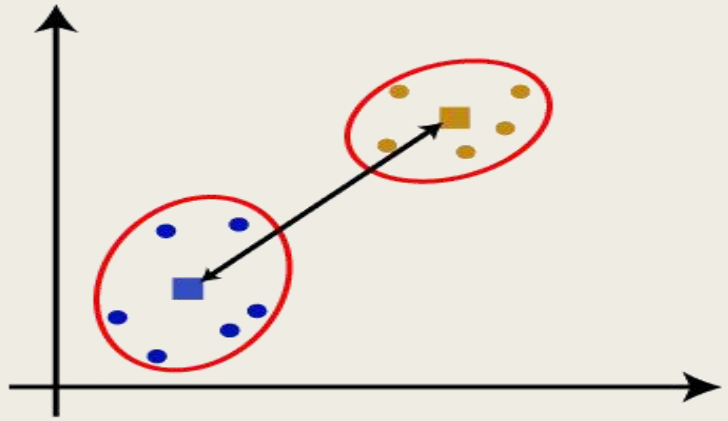


Measure for the distance between two clusters

- **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.

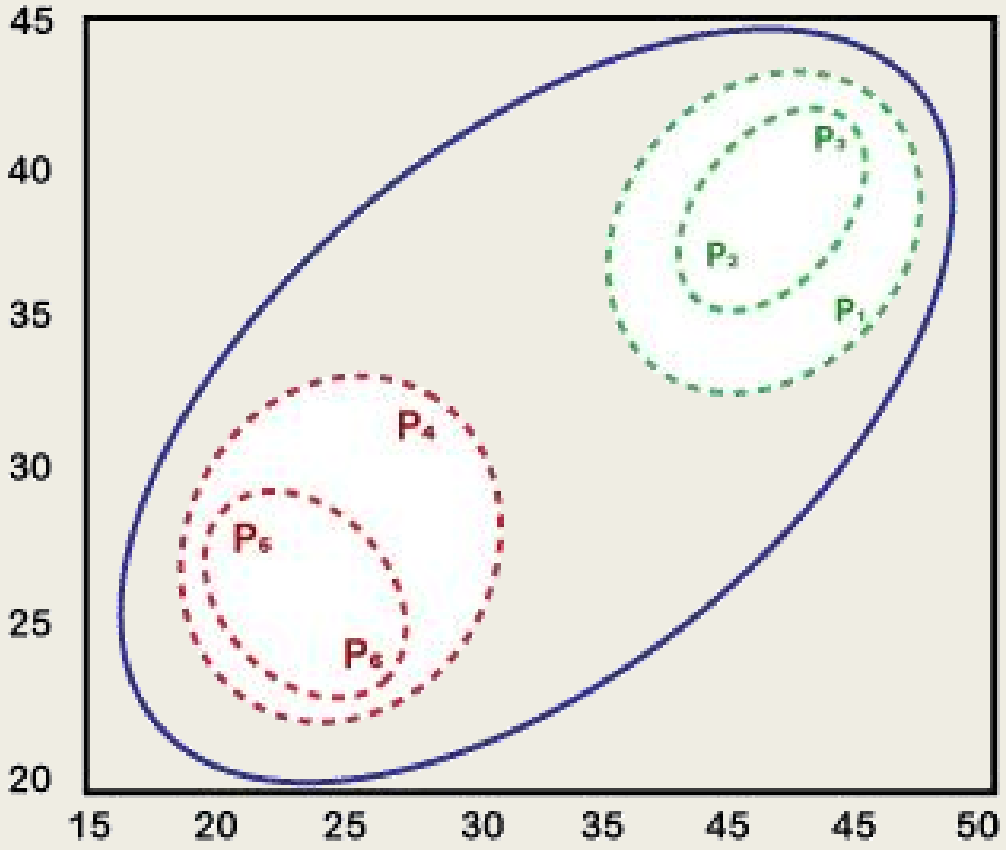


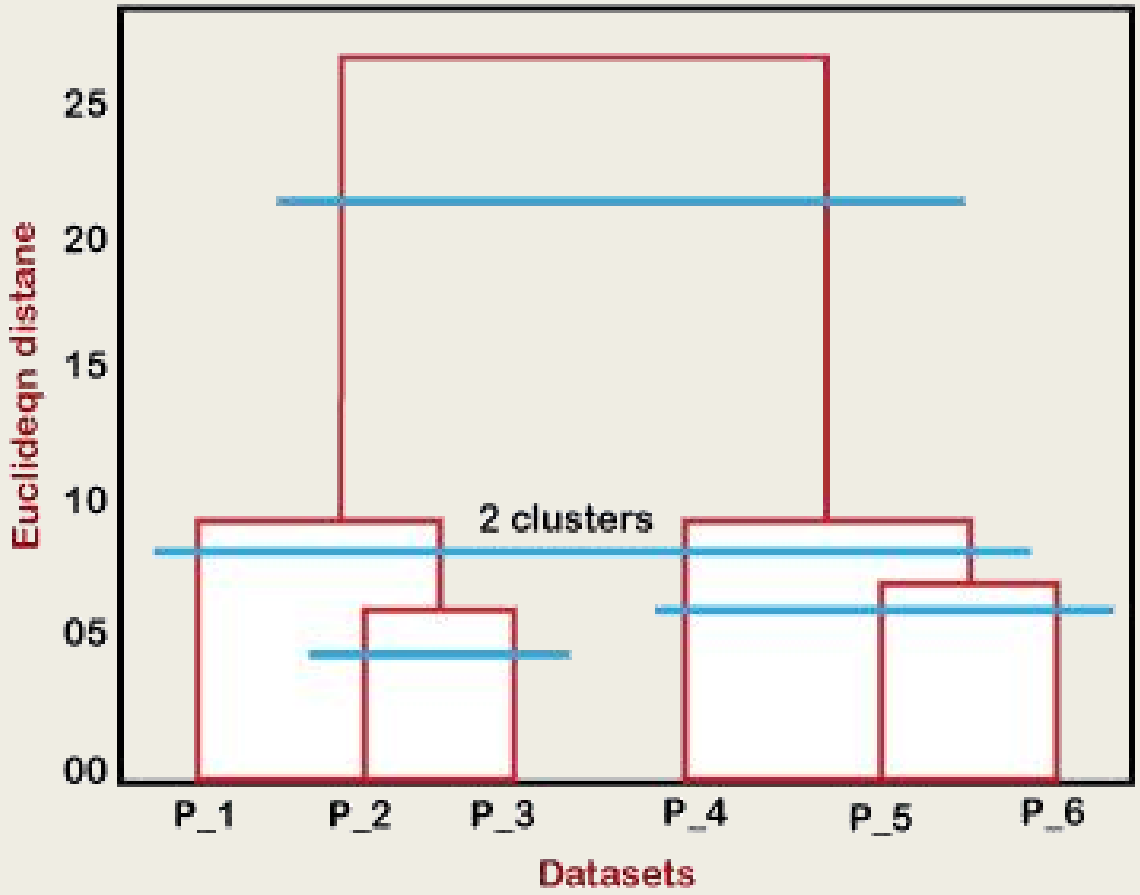
- **Centroid Linkage:** It is the linkage method in which the **distance between the centroid of the clusters is calculated.** Consider the below image:



Woking of Dendrogram in Hierarchical clustering

- The dendrogram is a tree-like structure that is mainly used to store each step as a memory that the HC algorithm performs. In the dendrogram plot, the Y-axis shows the Euclidean distances between the data points, and the x-axis shows all the data points of the given dataset.
- The working of the dendrogram can be explained using the below diagram:

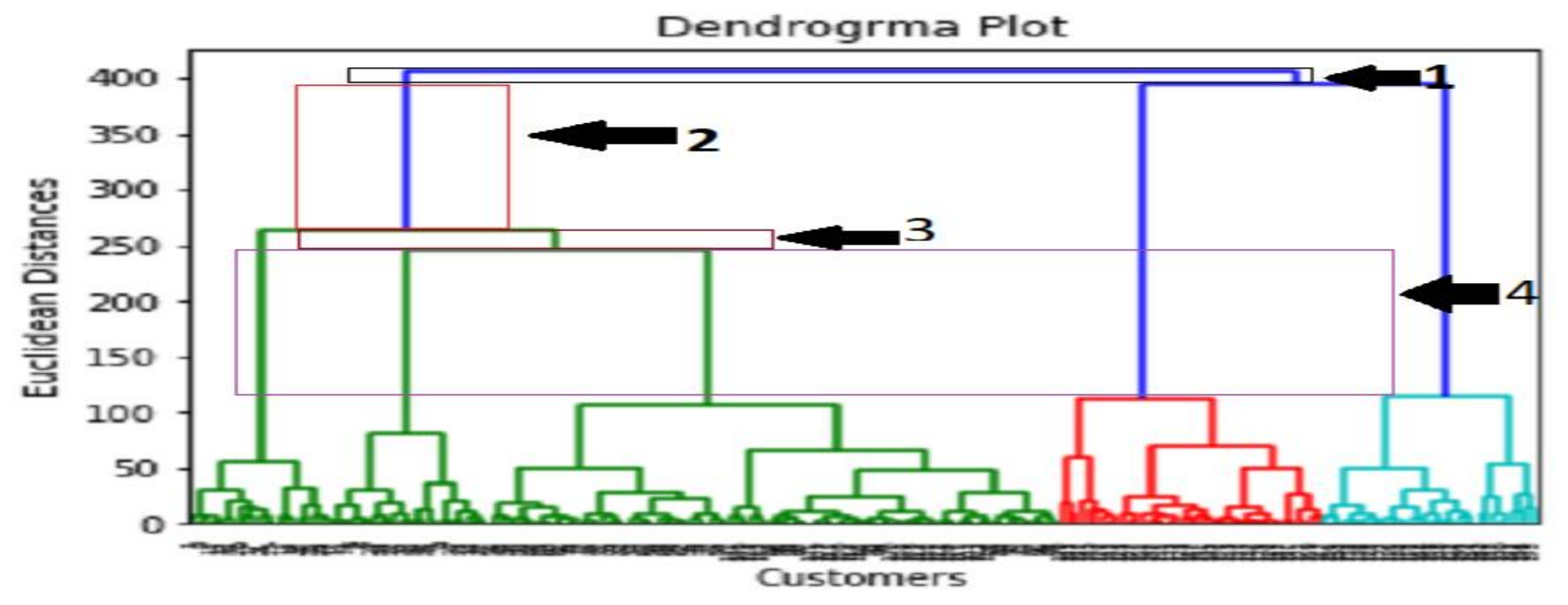




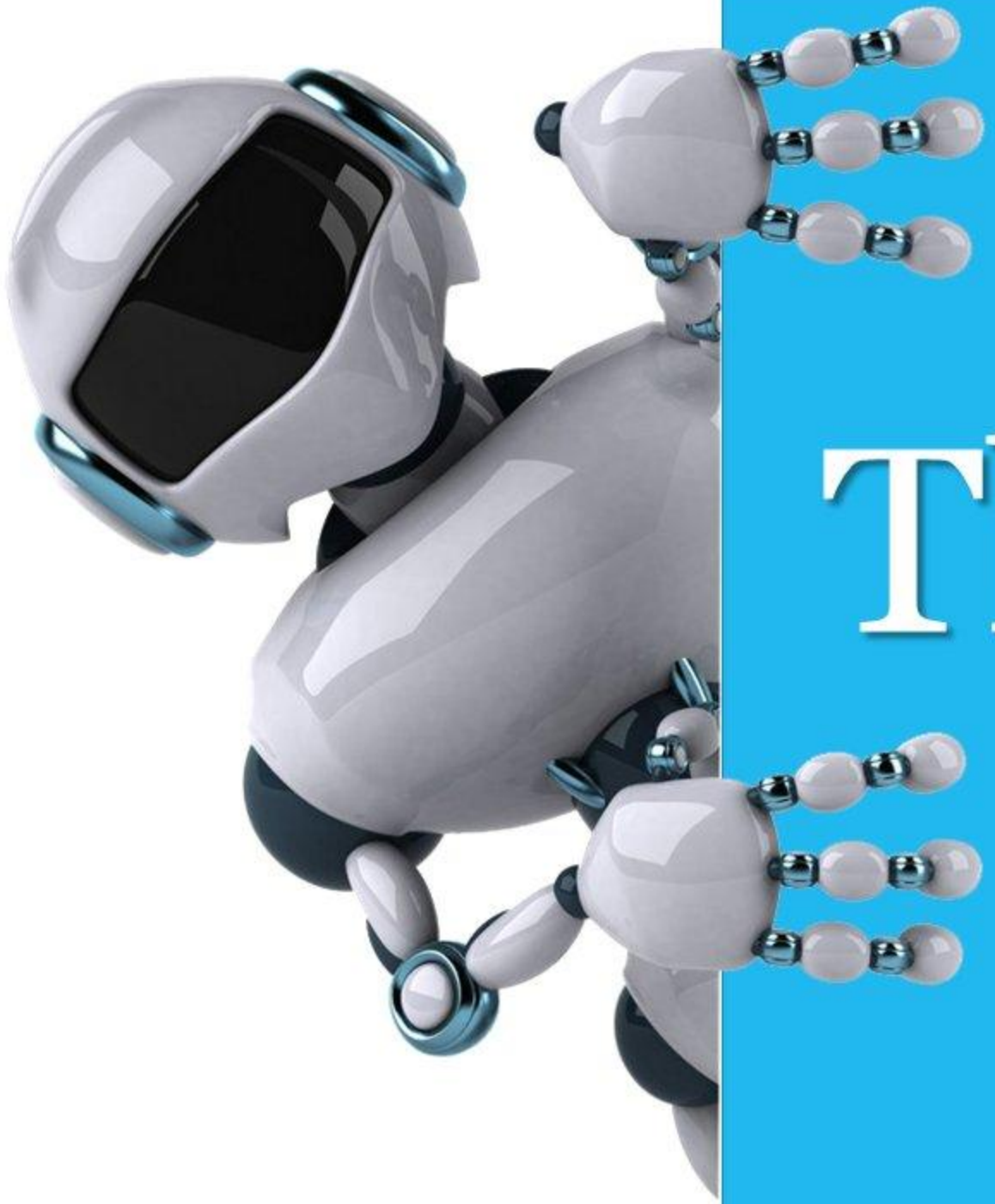
Woking of Dendrogram in Hierarchical clustering

- In the previous diagram, the left part is showing how clusters are created in agglomerative clustering, and the right part is showing the corresponding dendrogram.
- Firstly, the datapoints P2 and P3 combine together and form a cluster, correspondingly a dendrogram is created, which connects P2 and P3 with a rectangular shape. The hight is decided according to the Euclidean distance between the data points.
- In the next step, P5 and P6 form a cluster, and the corresponding dendrogram is created. It is higher than of previous, as the Euclidean distance between P5 and P6 is a little bit greater than the P2 and P3.
- Again, two new dendrograms are created that combine P1, P2, and P3 in one dendrogram, and P4, P5, and P6, in another dendrogram.
- At last, the final dendrogram is created that combines all the data points together.
- We can cut the dendrogram tree structure at any level as per our requirement.

Finding the optimal number of clusters using the Dendrogram



- In the above diagram, we have shown the **vertical distances that are not cutting their horizontal bars.**
- As we can visualize, the **4th distance is looking the maximum, so according to this, the number of clusters will be 5(the vertical lines in this range).**
- So, **the optimal number of clusters will be 5**, and we will train the model in the next step, using the same.



Thank you