Hierarchical Clustering

Hierarchical Clustering is an 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.

In this algorithm, there is no need to predetermine the number of clusters as we did in the k-means algorithm. The hierarchical clustering has two approaches :-

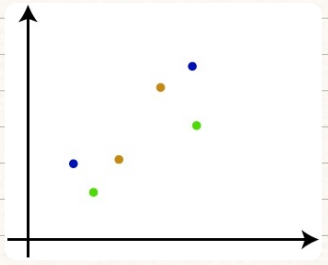
(i) Agglomerative :- Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.

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

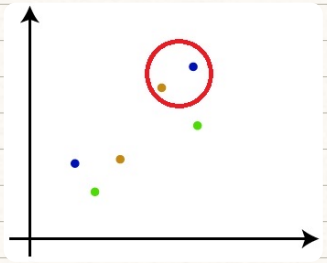
Note- k-means clustering always try to create clusters of same size and we need to predetermine number of clusters also. To solve these issues, we use hierarchical clustering.

Agglomerative Hierarchical Clustering :- It follows the bottom-up approach, which 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 dataset. The hierarchy of clusters is represented in the form of dendrogram. Working :-

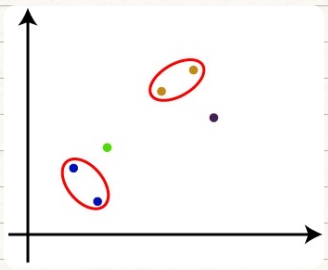
Step 1 :- Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.



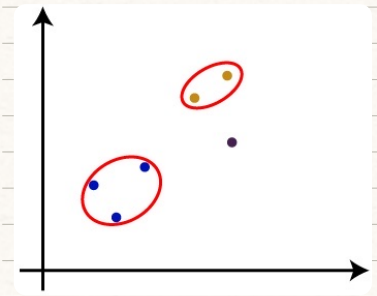
Step 2 :- Take two closest data points or clusters and merge them to form one cluster. So, there will now be $N-1$ clusters.

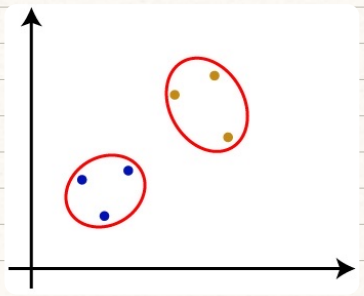


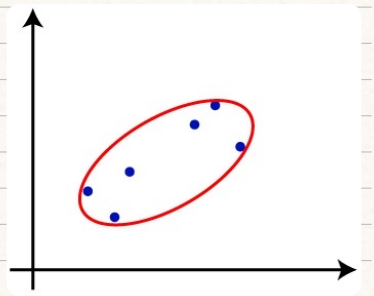
Step 3 :- Again, take the two closest clusters and merge them together to form one cluster. There will be $N-2$ clusters.



Step 4 :- Repeate Step 3 until only one cluster is left.



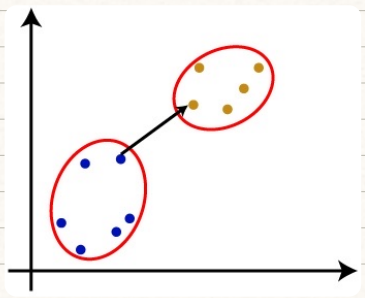




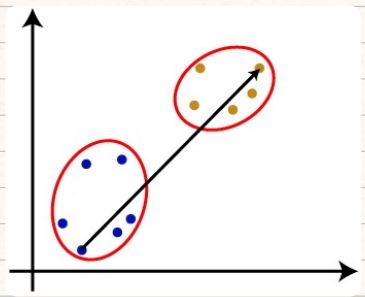
Step 5 :- Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.

Measure for the distance between two clusters :- The closest distance between two clusters is very 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. Types :-

(i) Single Linkage :- It is the shortest distance between the closest points of the clusters.

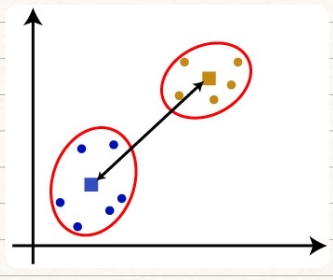


(ii) Complete Linkage :- It is the farthest distance between the two points of two different clusters. It is one of the popular methods as it forms tighter clusters than single-linkage.



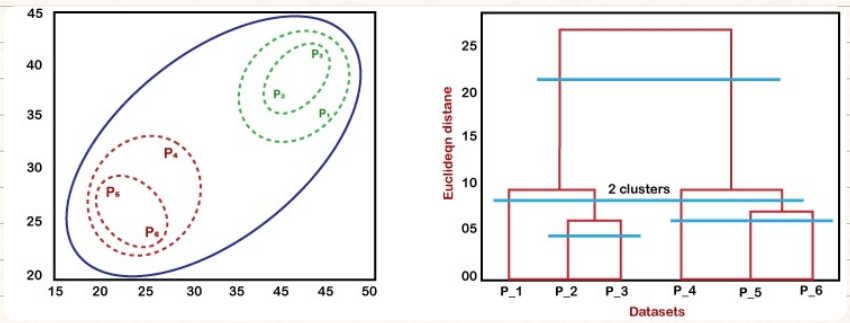
(iii) Average Linkage :- It is the linkage method in which the distance between each pair of data points is added up and then divided by the total number of data points to calculate the average distance between two clusters.

(iv) Centroid Linkage :- It is the linkage method in which the distance between the centroid of the clusters is calculated.



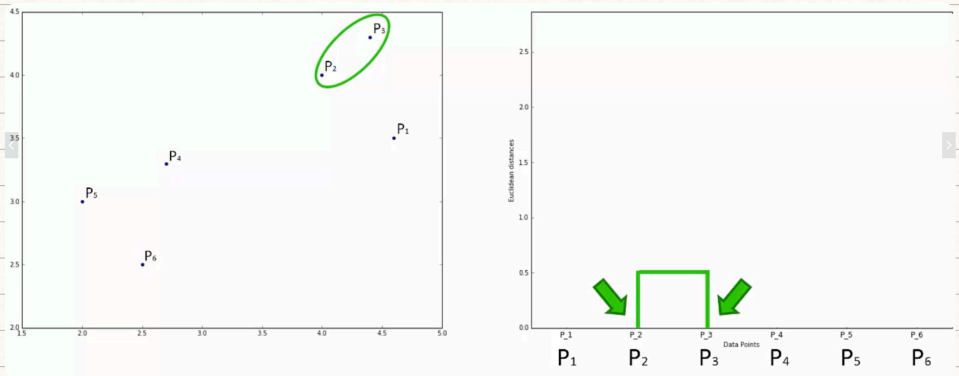
Working of Dendrogram in Hierarchical Clustering :-

The dendrogram is a tree-like structure that is mainly used to store each step as a memory that the Hierarchical Clustering 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.

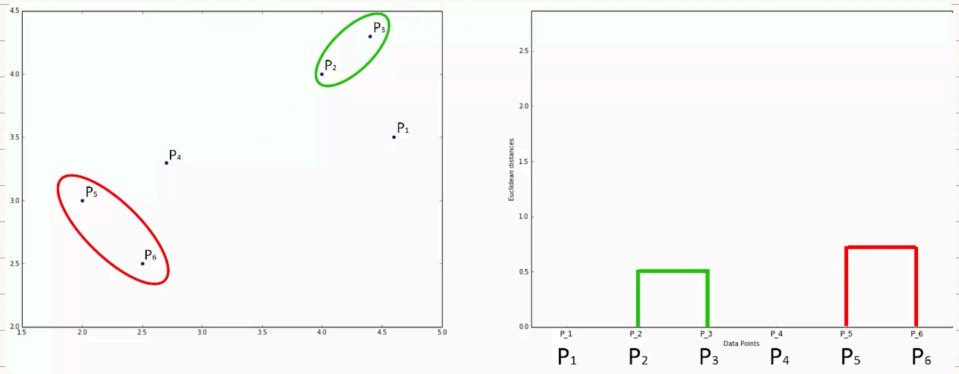


In the above diagram, the left part is showing how clusters are created in agglomerative clustering, and the right part is showing the corresponding dendrogram. Steps :-

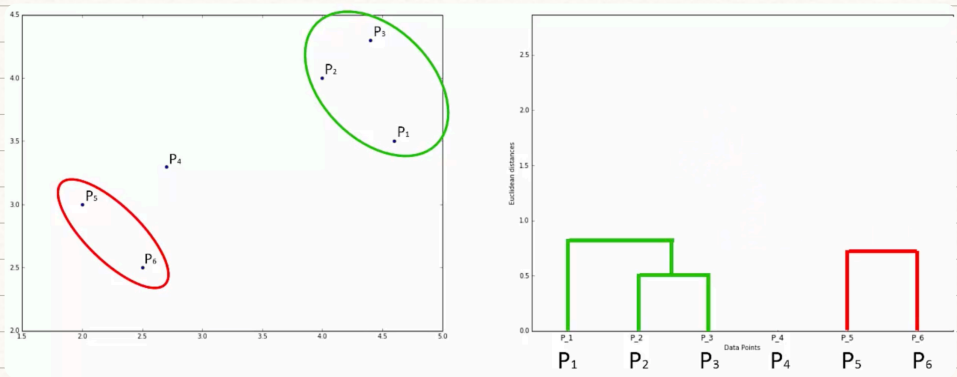
(i) Firstly, the datapoints P2 and P3 combine together and form a cluster, correspondingly a datagram is created, which connects P2 and P3 with a rectangular shape. The height is decided according to the Euclidean distance between the data points.

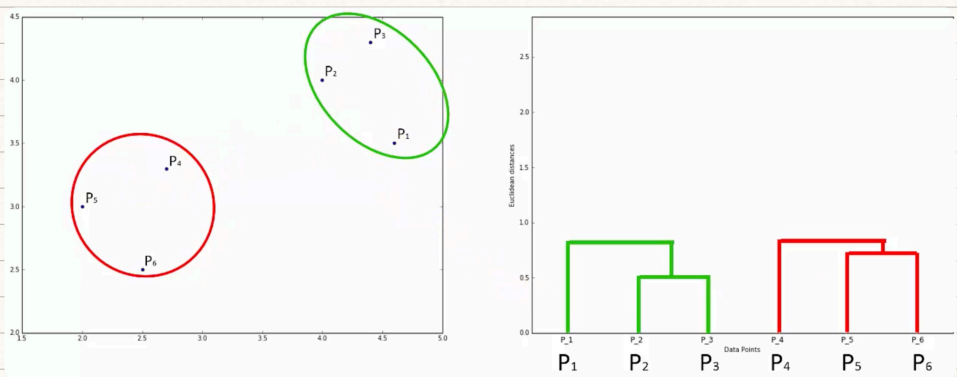


(ii) In the next step, P5 and P6 form a cluster, and the corresponding dendrogram is created. It is higher than of previous, as the Euclidean Distance between P5 and P6 is a little bit greater than the P2 and P3.



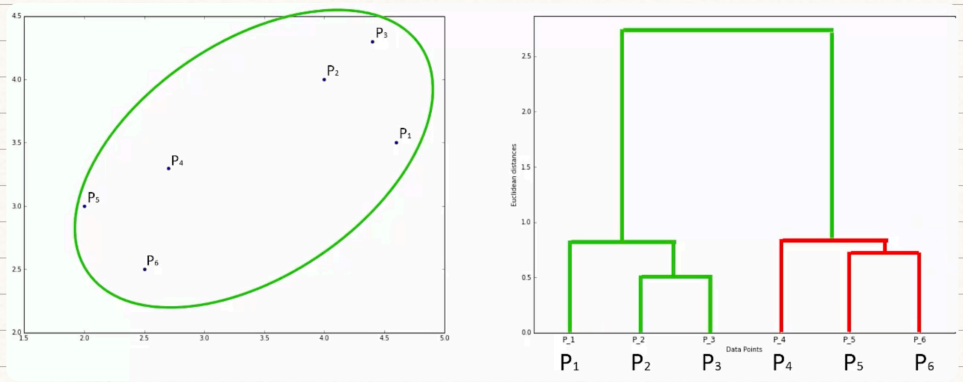
(iii) Again, two new dendrograms are created that combines P1, P2, and P3 in one dendrogram, and P4, P5, and P6 in another dendrogram.



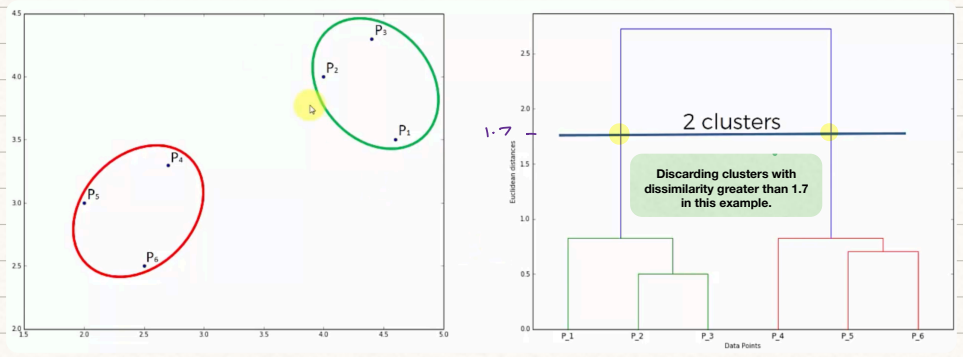


(iv) At last, the final dendrogram is created that combines all the data points together.

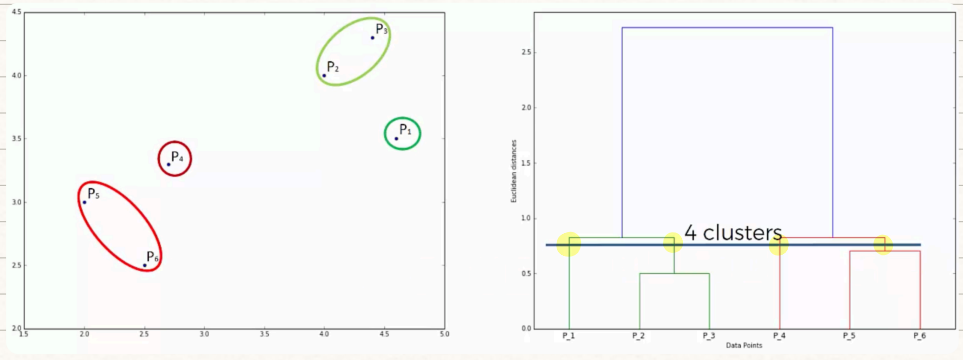
We can cut the dendrogram tree structure at any level as per our requirement.

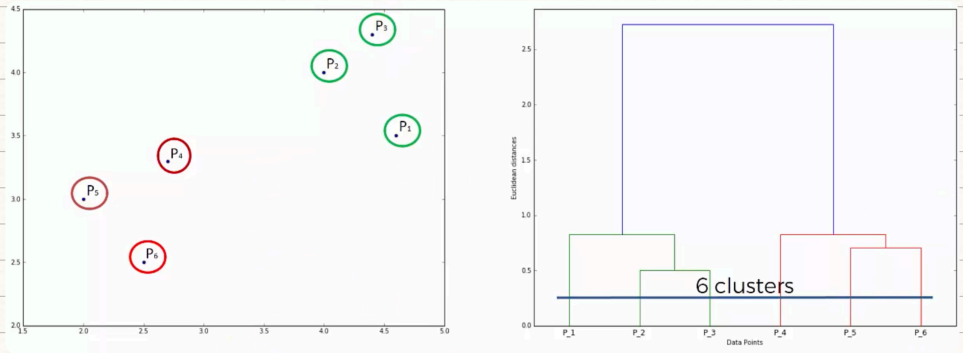


Now, what we can do is by looking at the horizontal levels of dendrograms, we can set distance threshold (also called dissimilarity threshold). We will not create clusters above the threshold which we have setted.

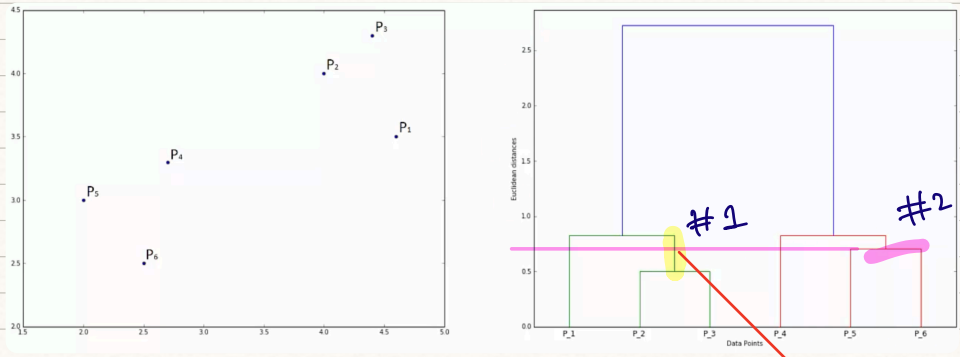


We can quickly tell how many clusters we will have by looking at the vertical lines the horizontal threshold actually crosses.



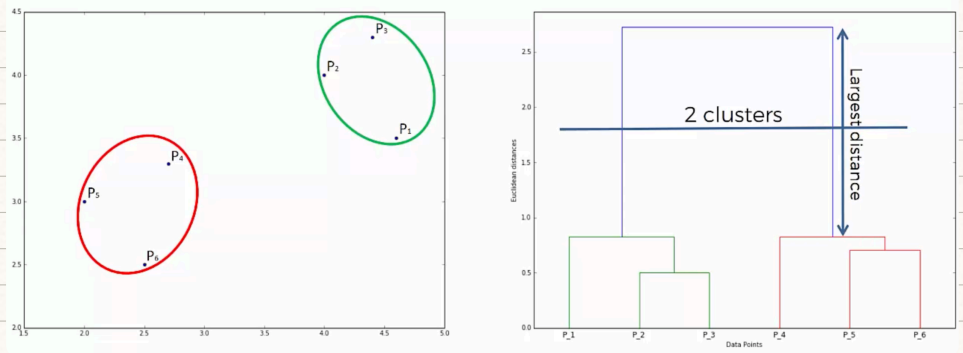


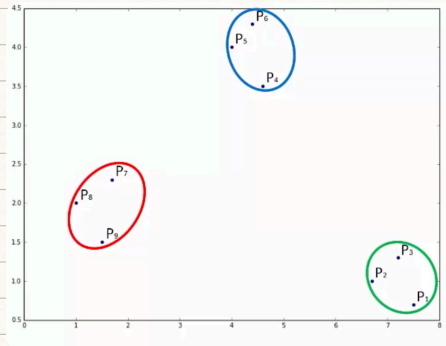
Selecting optimal number of clusters :- The standard approach is to look at the highest vertical distance that you can find in the dendrogram. Basically any vertical line which will not cross any horizontal line.

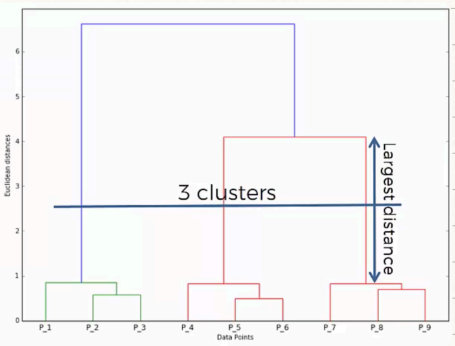


even this line cannot be considered as #2 will hypothically cross the #1 vertical line.

Dendrograms – Optimal # of Clusters




P6
P5
P4
P7
P8
P9
P3
P2
P1


3 clusters
Largest distance
Euclidean distances
Data Points
P_1
P_2
P_3
P_4
P_5
P_6
P_7
P_8
P_9