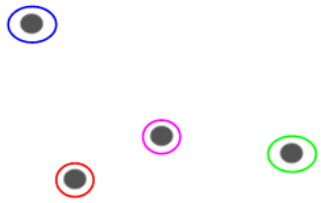
Hierarchical Clustering

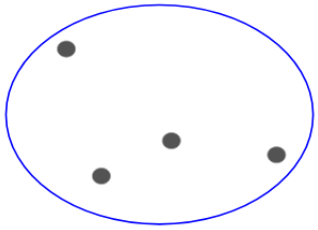
Hierarchical clustering is a type of unsupervised learning that groups similar data points or objects into groups called clusters.

Let's say we have the below points and we want to cluster them into groups:

We can assign each of these points to a separate cluster:



Now, based on the similarity of these clusters, we can combine the most similar clusters together and repeat this process until only a single cluster is left:



We are essentially building a hierarchy of clusters. That's why this algorithm is called hierarchical clustering.

Types of Hierarchical Clustering

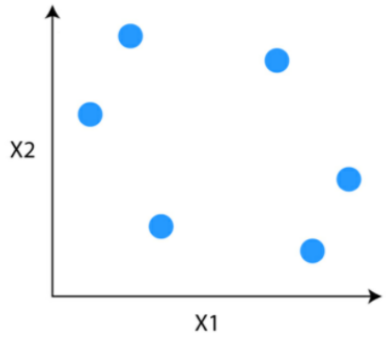
There are two types of hierarchical clustering:

- Agglomerative hierarchical clustering
- Divisive hierarchical clustering

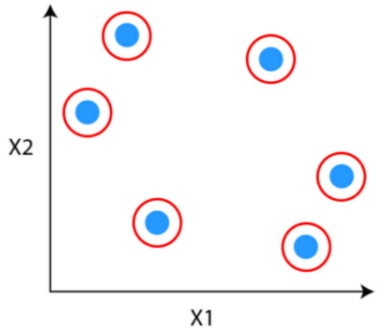
Agglomerative Hierarchical Clustering

The Agglomerative Hierarchical Clustering is the most common type of hierarchical clustering used to group objects in clusters based on their similarity. It's a bottom-up approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

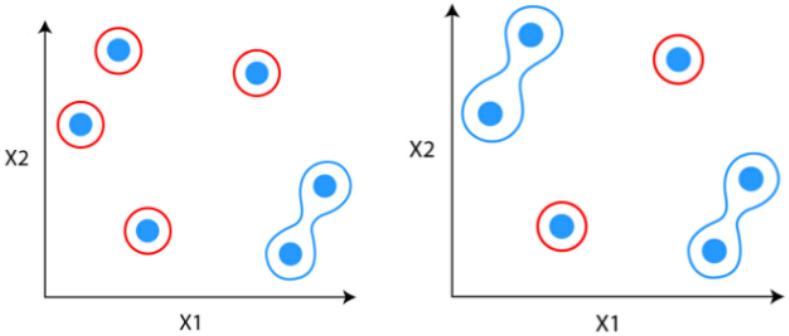
How does Agglomerative Hierarchical Clustering work?
Suppose you have data points which you want to group in similar clusters.



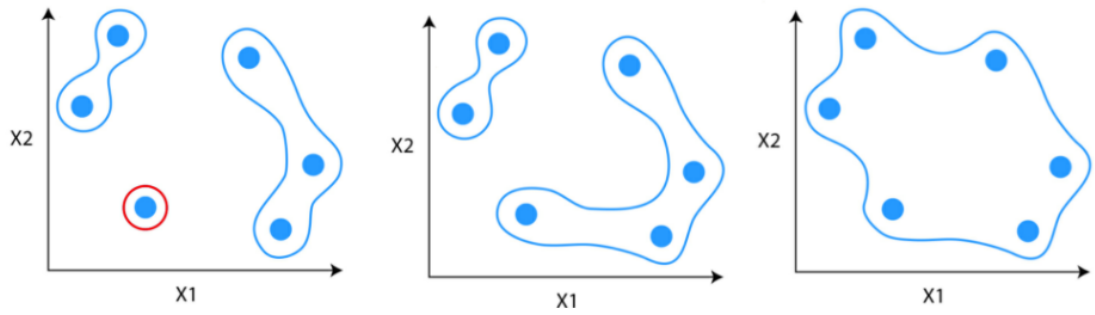
Step 1: The first step is to consider each data point to be a cluster.



Step 2: Identify the two clusters that are similar and make them one cluster.



Step 3: Repeat the process until only single clusters remains



How to identify if two clusters are similar?

One of the ways to do so is to find the distance between clusters.

Measure of distance (similarity)

- Euclidean Distance
- Hamming Distance
- Manhattan Distance (Taxicab or City Block)
- Minkowski Distance

Linkage Criterion

After selecting a distance metric, it is necessary to determine from where distance is computed. Some of the common linkage methods are:

Single-Linkage: Single linkage or nearest linkage is the shortest distance between a pair of observations in two clusters.
Complete-linkage: Complete linkage or farthest linkage is the farthest distance between a pair of observations in two clusters.
Average-linkage: Average linkage is the distance between each observation in one cluster to every observation in the other cluster.
Centroid-linkage: Centroid linkage is the distance between the centroids of two clusters. In this, you need to find the centroid of two clusters and then calculates the distance between them before merging.
Ward's-linkage: Ward's method or minimum variance method or Ward's minimum variance clustering method calculates the distance between two clusters as the increase in the error sum of squares after merging two clusters into a single cluster. This method seeks to choose the successive clustering steps so as to minimize the increase in sum of squares error at each step.

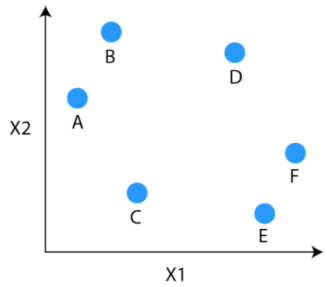
How to choose the number of clusters?

To choose the number of clusters in hierarchical clustering, we make use of concept called dendrogram.

What is a Dendrogram?

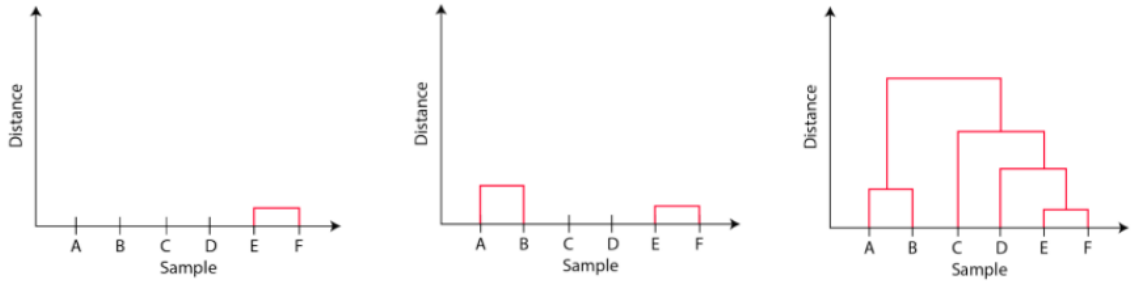
Dendrogram is a tree like diagram that shows the hierarchical relationship between the observations. It contains the memory of hierarchical clustering algorithms.

Just by looking at the Dendrogram you can tell how the cluster is formed. Let see how to form the dendrogram for the below data points.



The observations E and F are closest to each other by any other points. So, they are combined into one cluster and also the height of the link that joins them together is the smallest. The next observations that are closest to each other are A and B which are combined together.

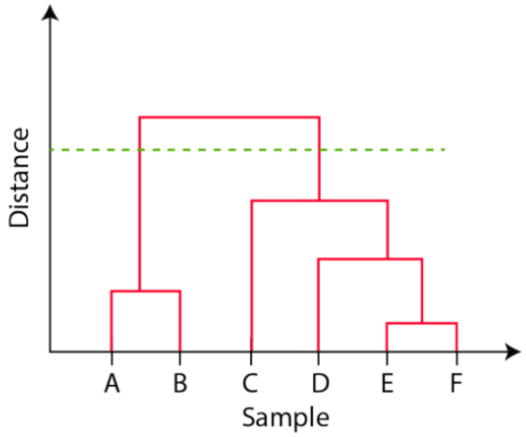
This can also be observed in the dendrogram as the height of the block between A and B is slightly bigger than E and F. Similarly, D can be merged into E and F clusters and then C can be combined to that. Finally A and B combined to C, D, E and F to form a single cluster.



The important point to note while reading dendrogram is that:

1. Height of the blocks represents the distance between clusters, and
2. Distance between observations represents dissimilarities.

But the question still remains the same, how do we find the number of clusters using a dendrogram or where should we stop merging the clusters. Observations are allocated to clusters by drawing a horizontal line through the dendrogram.



Generally, we cut the dendrogram in such a way that it cuts the tallest vertical line. In the above example, we have two clusters. One cluster has observations A and B, and a second cluster has C, D, E, and F.

Divisive Hierarchical Clustering

Divisive hierarchical clustering is not used much in solving real-world problems. It works in the opposite way of agglomerative clustering. In this, we start with all the data points as a single cluster.

At each iteration, we separate the farthest points or clusters which are not similar until each data point is considered as an individual cluster. Here we are dividing the single clusters into n clusters, therefore the name divisive clustering.

Pros and Cons of Hierarchical Clustering

1. Like K-means clustering, we need not to specify the number of clusters required for the algorithm.

2. It doesn't work well on the large dataset. It is generally applicable to the smaller data. If you have a large dataset, it can become difficult to determine the correct number of clusters by the dendrogram.

3. In comparison to K-means, hierarchical clustering is computationally heavy and takes longer time to run.

Example:

	P1	P2	P3	P4	P5
P1	0				
P2	9	0			
P3	3	7	0		
P4	6	5	9	0	
P5	11	10	2	8	0