Ho Le Thong 17110077

Huynh Thanh Dung 17110016

HierarchicalClustering

# Defination

***Hierarchical clustering***, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

## Steps to Perform Hierarchical Clustering

Following are the steps involved in agglomerative clustering:

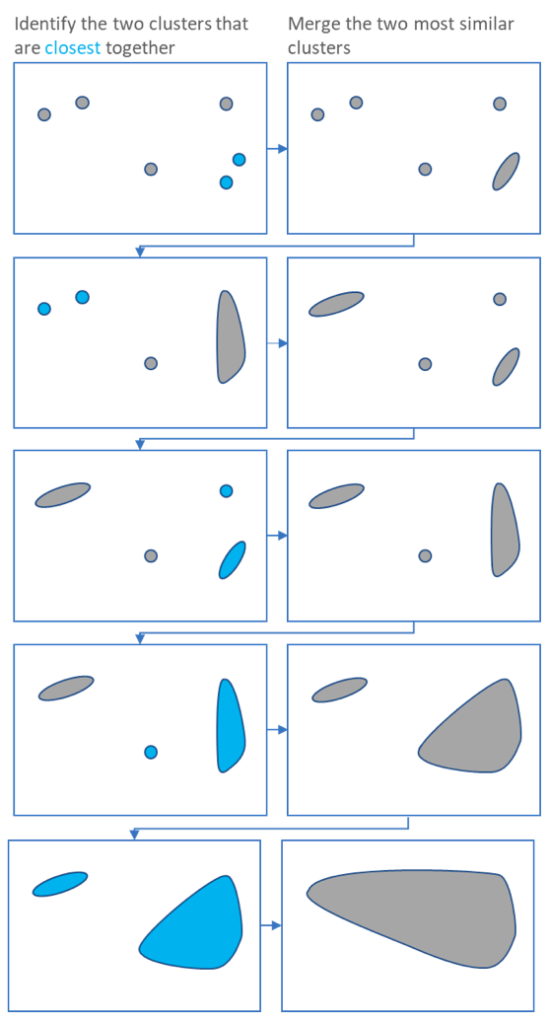
1. At the start, treat each data point as one cluster. Therefore, the number of clusters at the start will be K, while K is an integer representing the number of data points.
2. Form a cluster by joining the two closest data points resulting in K-1 clusters.
3. Form more clusters by joining the two closest clusters resulting in K-2 clusters.
4. Repeat the above three steps until one big cluster is formed.
5. Once single cluster is formed, [dendrograms](https://en.wikipedia.org/wiki/Dendrogram) are used to divide into multiple clusters depending upon the problem. We will study the concept of dendrogram in detail in an upcoming section.

There are different ways to find distance between the clusters. The distance itself can be Euclidean or Manhattan distance. Following are some of the options to measure distance between two clusters:

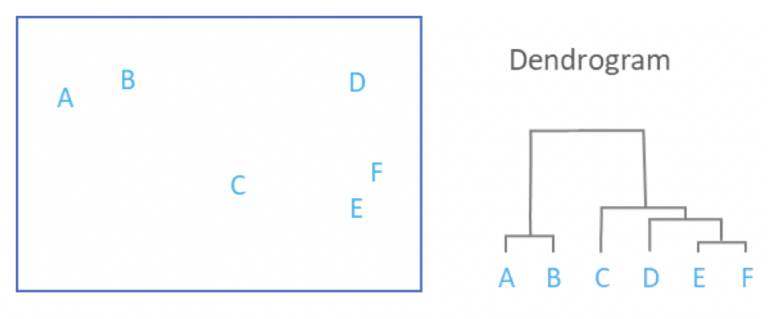
1. Measure the distance between the closes points of two clusters.
2. Measure the distance between the farthest points of two clusters.
3. Measure the distance between the centroids of two clusters.
4. Measure the distance between all possible combination of points between the two clusters and take the mean.

## **How hierarchical clustering works**

Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This iterative process continues until all the clusters are merged together. This is illustrated in the diagrams below.



The main output of Hierarchical Clustering is a [*dendrogram*](https://www.displayr.com/what-is-dendrogram/), which shows the hierarchical relationship between the clusters:



# Advantages

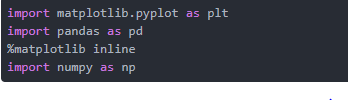
The strengths of hierarchical clustering are that it is easy to understand and easy to do

# Disadvantages

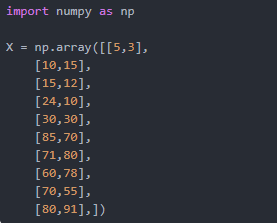
The weaknesses are that it rarely provides the best solution, it involves lots of arbitrary decisions, it does not work with missing data, it works poorly with mixed data types, it does not work well on very large data sets, and its main output, the dendrogram, is commonly misinterpreted.

# Example

The process of clustering is similar to any other unsupervised machine learning algorithm. We start by importing the required libraries:

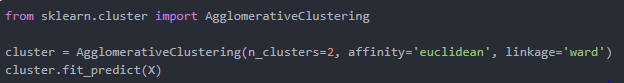


The next step is to import or create the dataset. In this example, we'll use the following example data:



The next step is to import the class for clustering and call its fit\_predict method to predict the clusters that each data point belongs to.

Take a look at the following script:



In the code above we import the AgglomerativeClustering class from the "sklearn.cluster" library. The number of parameters is set to 2 using the n\_clusters parameter while the affinity is set to "euclidean" (distance between the datapoints). Finally linkage parameter is set to "ward", which minimizes the variant between the clusters.

Next we call the fit\_predict method from the AgglomerativeClustering class variable cluster. This method returns the names of the clusters that each data point belongs to. Execute the following script to see how the data points have been clustered.



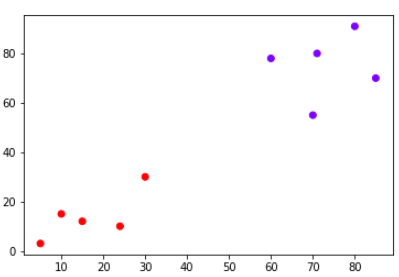
The output is a one-dimensional array of 10 elements corresponding to the clusters assigned to our 10 data points.



As expected the first five points have been clustered together while the last five points have been clustered together. It is important to mention here that these ones and zeros are merely labels assigned to the clusters and have no mathematical implications.

Finally, let's plot our clusters. To do so, execute the following code:





You can see points in two clusters where the first five points clustered together and the last five points clustered together.