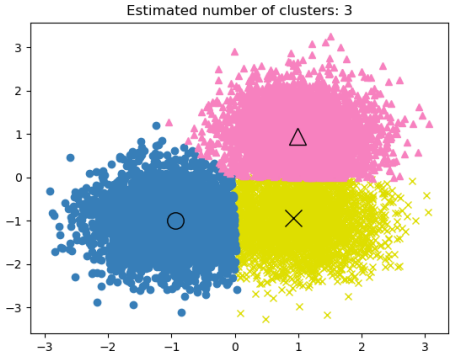
Lab Activity – K-Means and Agglomerative Hierarchical Clustering

K-Means & Hierarchical Clustering

Clustering of unlabeled data can be performed with the module **sklearn.cluster**.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more likely to have certain similarities.



Each clustering algorithm comes in two variants:

- a class, that implements the fit method to learn the clusters on train data,
- and a function, that, given train data, returns an array of integer labels corresponding to the different clusters.
  For the class, the labels over the training data can be found in the labels_ attribute.

K-Means is one of the most commonly used clustering techniques that is based on the concept of centroids.

The **KMeans** algorithm clusters data by trying to separate samples in n groups of equal variance, minimizing a criterion known as the *inertia* or within-cluster sum-of-squares.

This algorithm requires the number of clusters to be specified. It scales well to large numbers of samples and has been used across a large range of application areas in many different fields.

The k-means algorithm divides a set of N samples X into K disjoint clusters C, each described by the mean μ_j of the samples in the cluster. The means are commonly called the cluster “centroids”; note that they are not, in general, points from X, although they live in the same space.

The K-means algorithm aims to choose centroids that minimize the **inertia**, or **within-cluster sum-of-squares criterion**:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

Inertia can be recognized as a measure of how internally coherent clusters are.

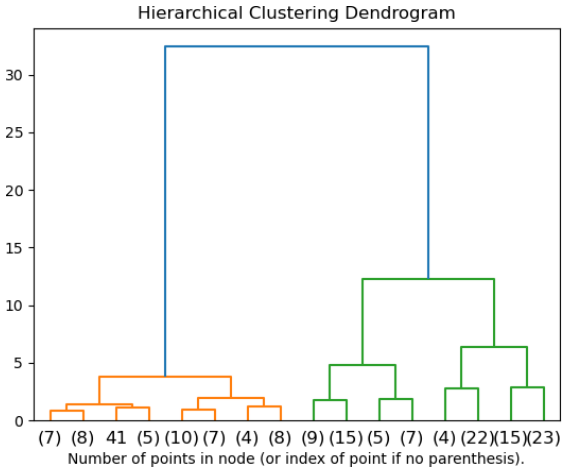
Hierarchical clustering is a general family of clustering algorithms that build nested clusters by merging or splitting them successively. This hierarchy of clusters is represented as a tree (or dendrogram). The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample.

The **AgglomerativeClustering** object performs a hierarchical clustering using a bottom up approach: each observation starts in its own cluster, and clusters are successively merged together. The linkage criteria determines the metric used for the merge strategy:

- **Ward** minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach and in this sense is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.
- **Maximum** or **complete linkage** minimizes the maximum distance between observations of pairs of clusters.
- **Average linkage** minimizes the average of the distances between all observations of pairs of clusters.
- **Single linkage** minimizes the distance between the closest observations of pairs of clusters.

AgglomerativeClustering can also scale to large number of samples when it is used jointly with a connectivity matrix, but is computationally expensive when no connectivity constraints are added between samples: it considers at each step all the possible merges.

It's possible to visualize the tree representing the hierarchical merging of clusters as a dendrogram. Visual inspection can often be useful for understanding the structure of the data, though more so in the case of small sample sizes.



Additional reading and references:
Lecture – Unsupervised Machine Learning
https://scikit-learn.org/stable/modules/clustering.html
https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering

Use case 1: K-means Clustering of Drivers
Use case 2: K-means & AHC Clustering of Mall Customers

In this lab activity, you will analyze two use cases.

Use case 1: K-means Clustering of Drivers

In the first use case, we group drivers into several clusters based on their driving speed and distance, using K-means clustering technique.

Use case 2: K-means & AHC Clustering of Mall Customers

In the second use case, we cluster mall customers based on information related to their genre, age, income and spending score, using both K-means and Agglomerative Clustering techniques.

Explore the implementation of k-means and agglomerative hierarchical clustering algorithms to solve this problem starting by loading the dataset, making some exploratory analysis, apply clustering technique on your data, use elbow and/or silhouette methods to identify the optimal k for kmeans, visualize your clustered datapoints, try to get insights on the generated clusters.