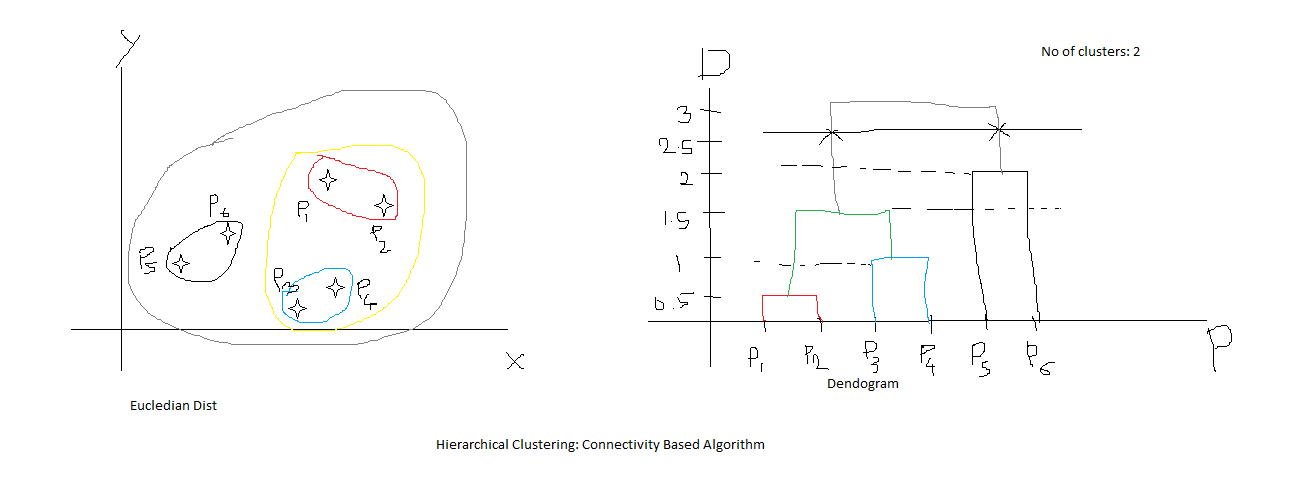
**What is Clustering:**

1. Unsupervised machine Learning
2. Involves discovering the natural grouping in data
3. Clustering algorithms are used to group data points based on certain similarities

**Hierarchical Clustering Algorithm**

1. Connectivity based algorithm
2. algorithms connect “objects” to form “clusters” based on their distance.
3. These algorithms provide a hierarchy of clusters that at certain distances are merged.
4. Dendrograms are used to represent the clusters.
5. Types: Agglomerative and Divisive
6. When to use this algorithms:
7. Parameters:
   1. n\_clustersint or None, default=2
   2. affinity: str or callable, default=’euclidean’
   3. linkage{‘ward’, ‘complete’, ‘average’, ‘single’}, default=’ward’
   4. distance\_thresholdfloat, default=None

**K-Means clustering algorithm:**

1. Centroid Based
2. Important step is to decide the no. of K (No. of centroids)
3. 2 methods to choose the value of K : Elbow and Silhouette
4. In the Elbow method we calculate the Within Cluster Sum of Square (WCSS) and plot WCSS against a range of K values, say from 1 to 20. And a point at the elbow is selected as value of K
5. Parameters:
   1. n\_clustersint, default=8
   2. init{‘k-means++’, ‘random’}, callable or array-like of shape (n\_clusters, n\_features), default=’k-means++’
6. Main Advantage: Better time performance

Mini Batch KMeans:

1. When the size of the dataset increases, k means result in memory issues.
2. To reduce the time and space complexity, mini batch is introduced
3. It tries to fit the data in the main memory in a way where the algorithm uses small batches of data that are of fixed size chosen at random

**DBSCAN: Density-Based Spatial Clustering of Application with Noise:**

1. Takes a point and draws a circle with the radius = epsilon.
2. If we get points = minPts inside this circle then our center point is called core point
3. If we do not get points = minPts inside circle but we get at least one core point inside the circle then we consider the center as border point
4. If we neither get points = minPts nor get any core point inside the circle then we consider the center as noise.
5. Noise are eliminated from the clusters
6. Parameters:
   1. eps: float, default=0.5
   2. min\_samples: int, default=5
   3. metric: str, or callable, default=’euclidean’
7. How to decide the values of eps and min\_samples: https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd

**Gaussian Mixture Model : Distribution Based Clustering Algorithm**

1. This assumes that there are a certain number of Gaussian distributions and each of these distributions represent a cluster
2. K-means is suitable for data when points are distributed in circular fashion.
3. a Gaussian Mixture Model tends to group the data points belonging to a single distribution together
4. GMM are probabilistic models.
5. GMM uses soft clustering techniques (a probability score is assigned to data points to be in those clusters.) When data points belong to more than one cluster at a time, we use soft clustering.
6. for a dataset with d features, we would have a mixture of k Gaussian distributions (where k is equivalent to the number of clusters), each having a certain mean vector and variance matrix
7. Value for mean and variance is determined by using a technique called Expectation-Maximization (EM)
8. EM has 2 steps:
   1. E-step: In this step, the available data is used to estimate (guess) the values of the missing variables
   2. M-step: Based on the estimated values generated in the E-step, the complete data is used to update the parameters

**Python Code to determine the optimum number of clusters:**

<https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad>

Can we identify which