Unsupervised Learning

Topics covered so far

Unsupervised Learning

- K-Means clustering
- PAM (K-Medoids) clustering
- Hierarchical clustering
- GMM
- DBSCAN

Discussion Questions

1. What is K-Means clustering and what are the advantages and disadvantages of using K-Means Clustering?

2. Why PAM (K-Medoids) clustering is a good alternative for K-Means clustering?

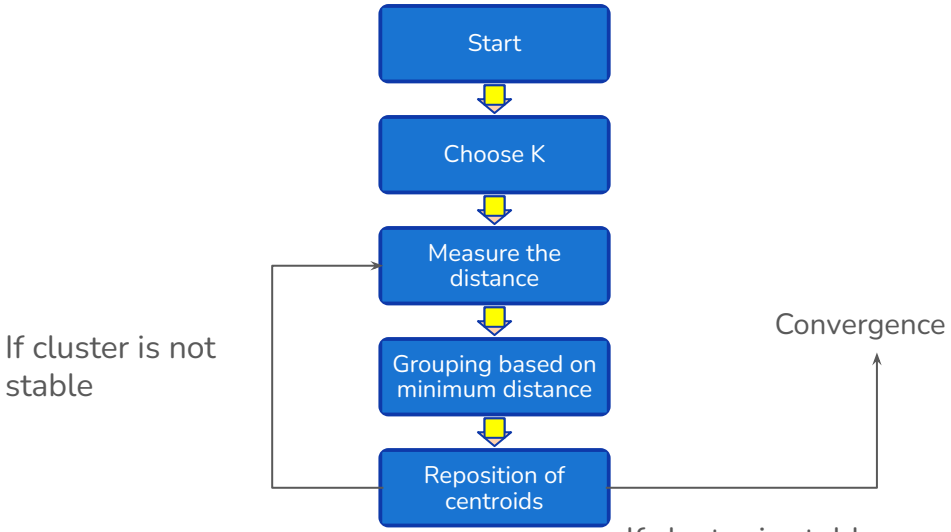
3. What is the expectation-maximization algorithm and how does it help in GMM clustering?

4. What is hierarchical clustering and how do we measure dissimilarity among clusters in hierarchical clustering?

5. How does DBSCAN work and what are the parameters that can be tuned in DBSCAN?

K-Means clustering

K-Means Clustering is an iterative **algorithm** that divides the unlabeled dataset into **K** different **clusters** in such a way that each point in the dataset belongs to only one group that has similar properties.



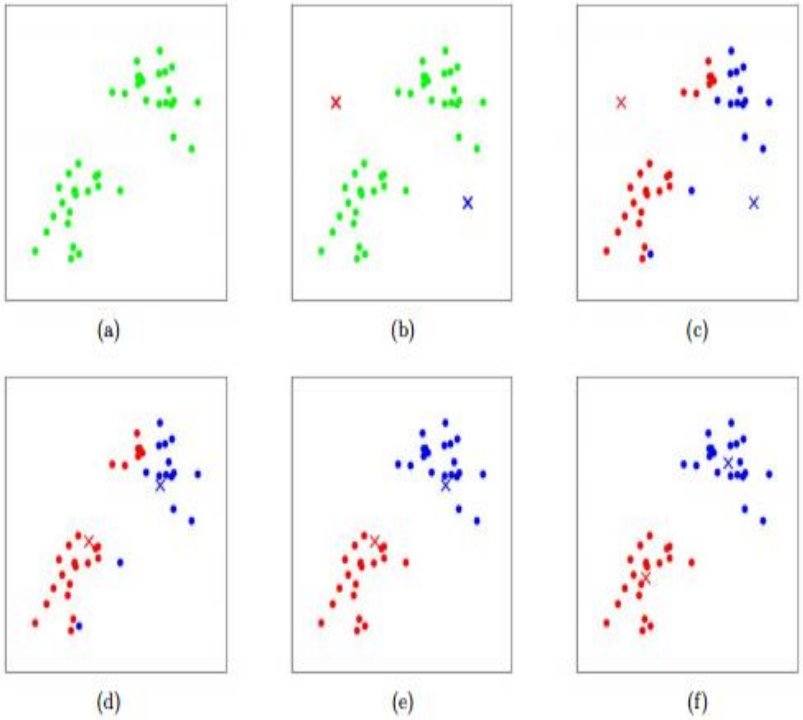


Image Source

Advantages and Disadvantages of using K-Means clustering

Advantages:

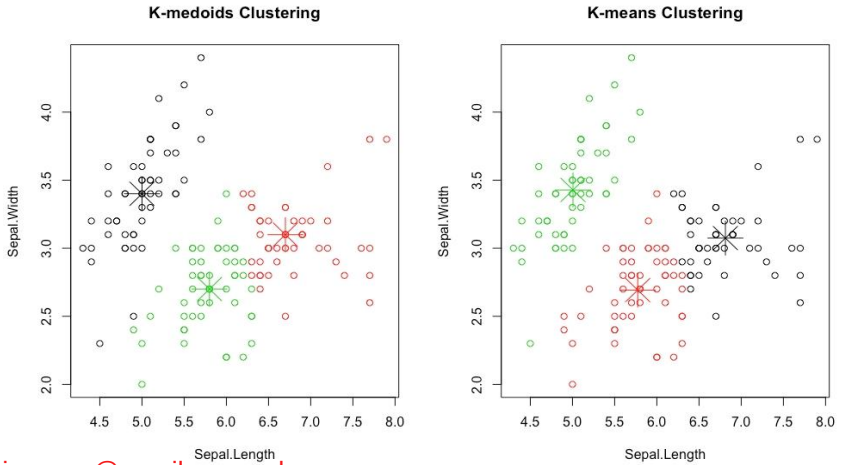
- K-Means is relatively simple to implement.
- It scales to large datasets.
- It also guarantees convergence.
- It can easily be adapted to new examples.

Disadvantages:

- It is difficult to identify the value of K.
- K-Means has trouble clustering data where clusters are of varying sizes and densities.
- It can easily be affected by outliers.
- It assumes the data shape to be spherical and does not perform well on arbitrary data.
- It depends on the initial values assigned to the centroids and gives different results for different initializations.

Alternative to K-Means - PAM (K-Medoids) clustering

- The problem with K-Means is that the final centroids are not interpretable, i.e., centroids are not actual points but the means of the points present in the cluster.
- The idea behind K-Medoids clustering is to make the final centroids as actual data points so that they are interpretable.
- In K-Medoids, we only change one step from K-Means which is to update the centroids. In this process, if there are m points in a cluster, swap the previous centroids with all other (m-1) points from the cluster and finalize the point as new centroid which has minimum loss.
- Because of this, unlike K-Means, it is robust to outliers and converges fast.
- You can see in this image that the centroids in K-Medoids are the actual data points represented as the cross, unlike K-Means.



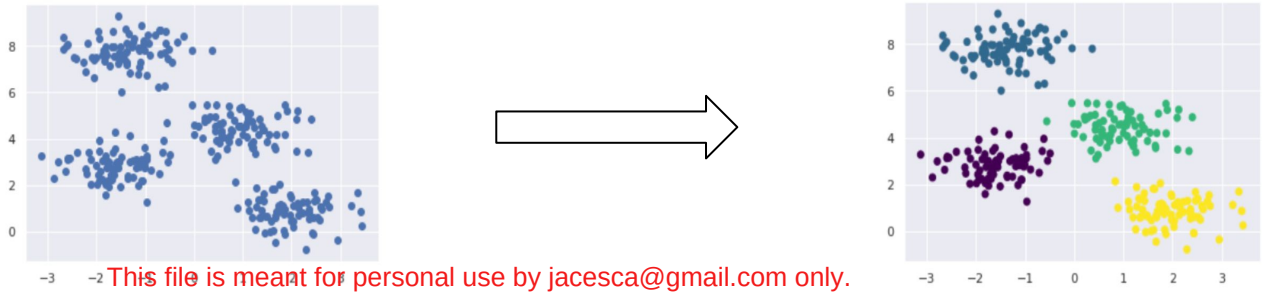
Expectation Maximization in GMM clustering

In GMM, we need the parameters of each Gaussian (mean, variance, etc.) in order to cluster our data, but we need to know which sample belongs to what Gaussian in order to estimate those very same parameters.
That is where we need the EM algorithm. There are two steps involved in this algorithm:

1. **The E-step:** It estimates the probability that a given observation is in a cluster/distribution. This value will be high when the point is assigned to the right cluster and lower otherwise.
2. **The M-step:** In this step, we want to maximize the likelihood that each observation came from the distribution.

After that, we reiterate these two steps and update the probabilities of observation being in a cluster.

Example of GMM clustering



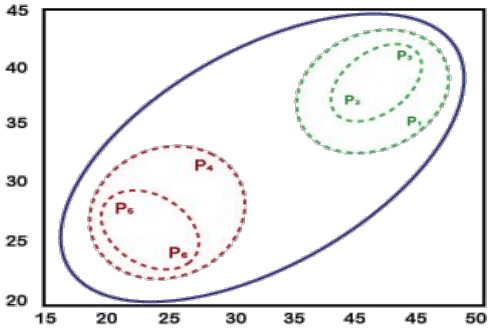
Hierarchical clustering

Hierarchical clustering is an unsupervised clustering algorithm that involves creating clusters that have predominant ordering from top to bottom. For example, all files and folders on our hard disk are organized in a hierarchy.

The algorithm groups similar objects into groups called "clusters." The endpoint is a set of clusters or groups, where each cluster is distinct from the other cluster, and the objects within each cluster are broadly similar to each other.

Steps:

- Make each data point a single-point cluster → that forms N clusters
- Take the two closest data points and make them one cluster → that forms N-1 clusters
- Take the two closest clusters and make them one cluster → that forms N-2 clusters.
- Repeat step-3 until you are left with only one cluster.



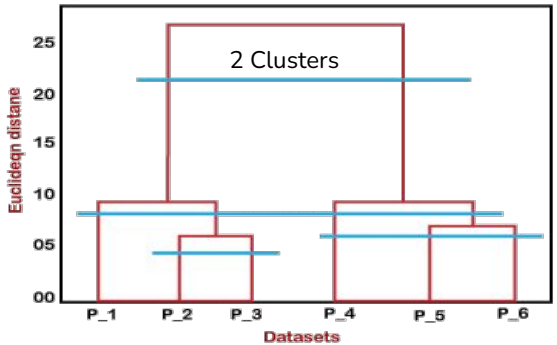
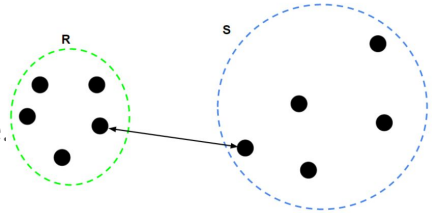


Image source

Dissimilarity among clusters in hierarchical clustering

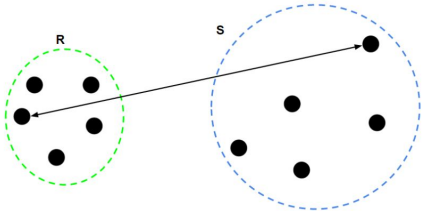
The below are some of the following ways by which we can measure dissimilarity among clusters in hierarchical clustering:

- **Single linkage:** It measures the closest pair of points, i.e., the minimum distance.



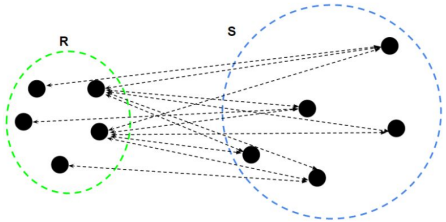
L(R,S) = min(d(i,j)) where i belongs to R and j belongs to S

- **Complete linkage:** It measures the farthest pair of points i.e the maximum distance



L(R,S) = max(d(i,j)) where i belongs to R and j belongs to S

- **Average linkage:** It measures the average dissimilarity over all pairs i.e. the average distance



$$L(R,S) = \frac{1}{n_R+n_S}\sum_{i=1}^{n_R}\sum_{j=1}^{n_S} D(i,j), i\epsilon R, j\epsilon S$$

Image Source

DBSCAN

DBSCAN stands for **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise.

It recognizes groups in the data by looking at the local density of a data point. Unlike K-means, **DBSCAN clustering is not sensitive to outliers** and also does not require the number of clusters to be told beforehand.

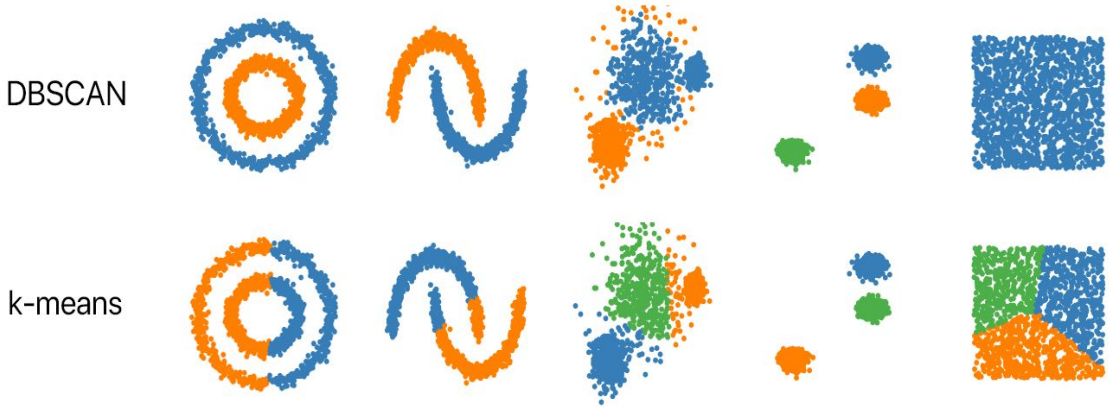
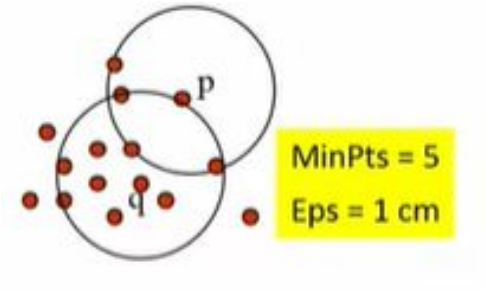


Image Source

Parameters in DBSCAN

- **eps (‘ε’):** It defines the neighborhood around a data point, i.e., if the distance between two points is lower or equal to 'eps', then they are considered neighbors. If the eps value is chosen too small, then a large part of the data will be considered outliers. If it is chosen very large, then the clusters will merge and the majority of the data points will be in the same clusters. One way to find the eps value is based on the k-distance graph.



- **MinPts:** Minimum number of neighbors (data points) within eps radius. The larger the dataset, the larger value of MinPts must be chosen. As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as MinPts >= D+1. The minimum value of MinPts must be chosen at least.

Case Study - Clustering

Great Learning

Happy Learning !