UNSUPERVISED MACHINE LEARNING - CLUSTERING

A task involving machine learning may not be linear, but it has a number of well-known steps:

- Problem definition.
- Preparation of Data.
- Learn an underlying model.
- Improve the underlying model by quantitative and qualitative evaluations.
- Present the model.

One good way to come to terms with a new problem is to work through identifying and defining the problem in the best possible way and learn a model that captures meaningful information from the data. While problems in Pattern Recognition and Machine Learning can be of various types, they can be broadly classified into three categories:

- **Supervised Learning:**
  The system is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

- **Unsupervised Learning:**
  No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

- **Reinforcement Learning:**
  A system interacts with a dynamic environment in which it must perform a certain goal (such as driving a vehicle or playing a game against an opponent). The system is provided feedback in terms of rewards and punishments as it navigates its problem space.

Unsupervised Learning can be further classified into two categories:

Parametric Unsupervised Learning

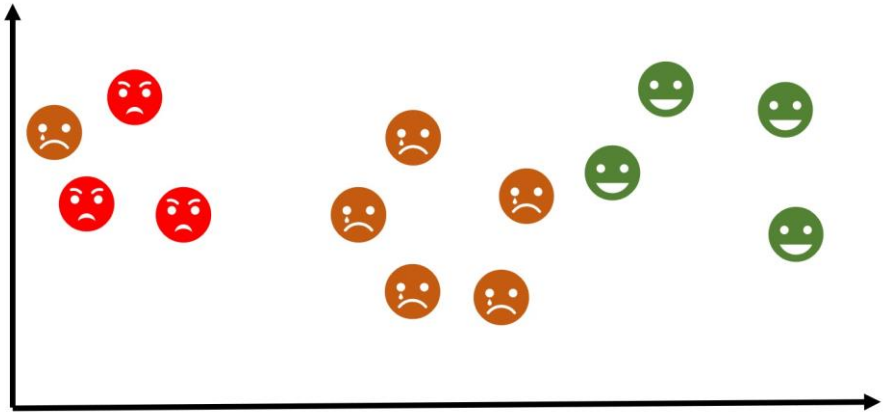
In this case, we assume a parametric distribution of data. It assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters. Theoretically, in a normal family of distributions, all members have the same shape and are parameterized by mean and standard deviation. That means if you know the mean and standard deviation, and that the distribution is normal, you know the probability of any future observation. Parametric Unsupervised Learning involves construction of Gaussian Mixture Models and using Expectation-Maximization algorithm to predict the class of the sample in question. This case is much harder than the standard supervised learning because there are no answer labels available and hence there is no correct measure of accuracy available to check the result.

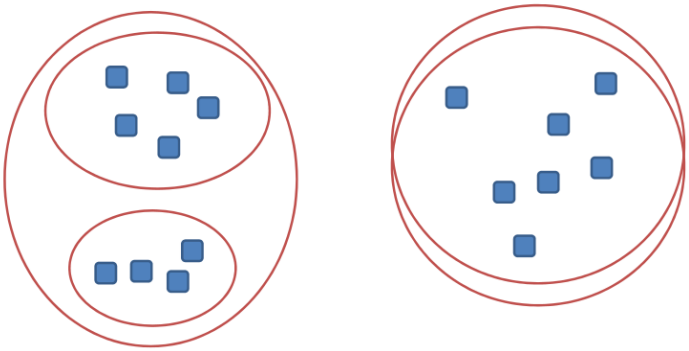
Non-parametric Unsupervised Learning

In non-parameterized version of unsupervised learning, the data is grouped into clusters, where each cluster(hopefully) says something about categories and classes present in the data. This method is commonly used to model and analyze data with small sample sizes. Unlike parametric models, nonparametric models do not require the modeler to make any assumptions about the distribution of the population, and so are sometimes referred to as a distribution-free method.

What is Clustering?

Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.





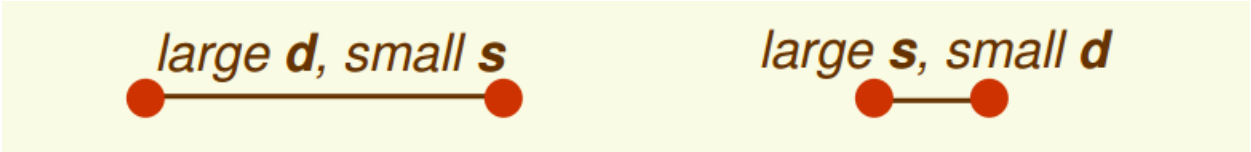
In the above image, how do we know what is the best clustering solution?

To find a particular clustering solution, we need to define the similarity measures for the clusters.

Proximity Measures

For clustering, we need to define a proximity measure for two data points. Proximity here means how similar/dissimilar the samples are with respect to each other.

- Similarity measure S(xi,xk): large if xi,xk are similar
- Dissimilarity (or distance) measure D(xi,xk): small if xi,xk are similar



There are various similarity measures which can be used.

- Vectors: Cosine Distance

$$s(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^t \mathbf{x}'}{\|\mathbf{x}\| \ \|\mathbf{x}'\|}$$

- Sets: Jaccard Distance

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

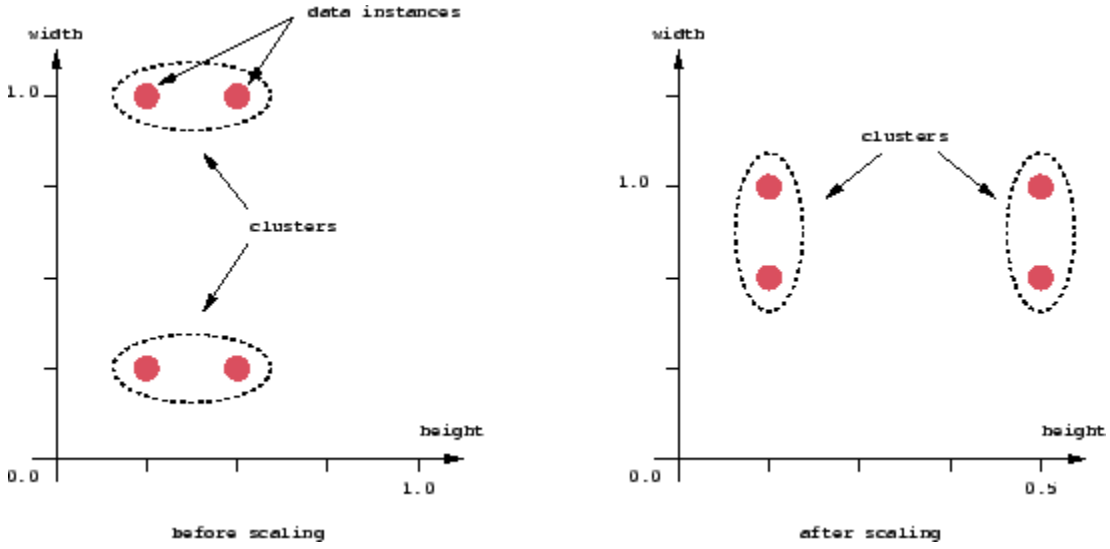
(If A and B are both empty, we define $J(A,B) = 1$.)

$$0 \leq J(A,B) \leq 1.$$

- Points: Euclidean Distance
  q=2

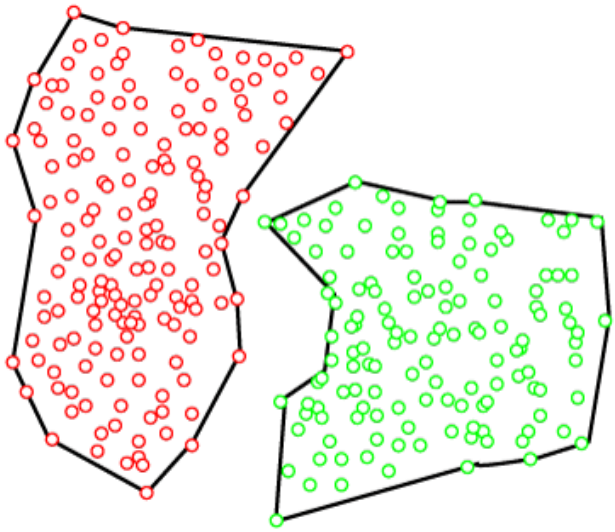
$$d(\mathbf{x}, \mathbf{x}') = \left(\sum_{k=1}^{d} |x_k - x'_k|^q \right)^{1/q},$$

A “good” proximity measure is VERY application dependent. The clusters should be invariant under the transformations “natural” to the problem. Also, while clustering it is not advised to normalize data that are drawn from multiple distributions.

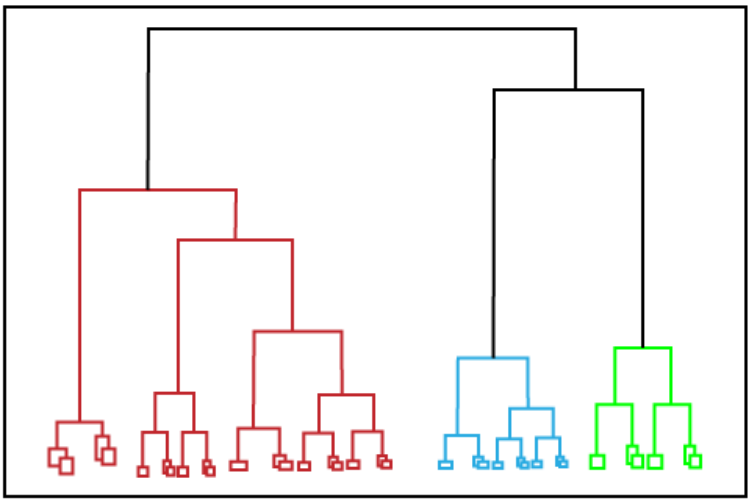


Clustering Methods:

- **Density-Based Methods:** These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters. Example *DBSCAN (Density-Based Spatial Clustering of Applications with Noise)*, *OPTICS (Ordering Points to Identify Clustering Structure)*, etc.

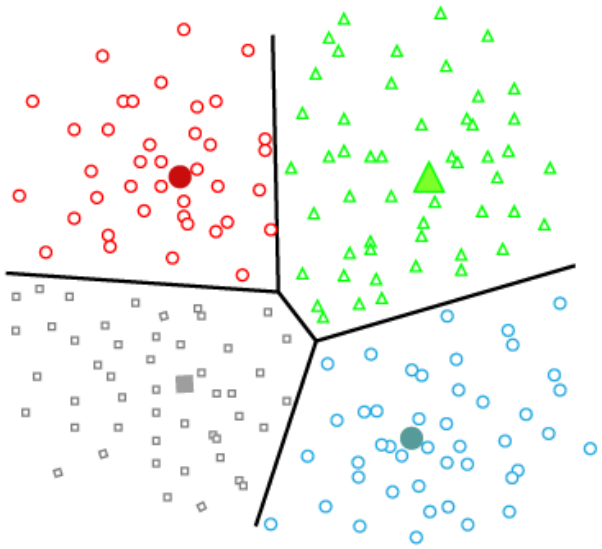


- **Hierarchical Based Methods:** The clusters formed in this method form a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two categories:
    - **Agglomerative** (bottom-up *approach*)
    - **Divisive** (top-down *approach*)



Examples: *CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies)*, etc.

- **Partitioning Methods:** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example *K-means, CLARANS (Clustering Large Applications based upon Randomized Search)*, etc.



- **Grid-based Methods:** In this method, the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operations done on these grids are fast and independent of the number of data objects example *STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest)*, etc.

Applications of Clustering in different fields:

- **Marketing:** It can be used to characterize & discover customer segments for marketing purposes.
- **Biology:** It can be used for classification among different species of plants and animals.
- **Libraries:** It is used in clustering different books on the basis of topics and information.
- **Insurance:** It is used to acknowledge the customers, their policies and identifying the frauds.
- **City Planning:** It is used to make groups of houses and to study their values based on their geographical locations and other factors present.
- **Earthquake studies:** By learning the earthquake-affected areas we can determine the dangerous zones.
- **Image Processing**: Clustering can be used to group similar images together, classify images based on content, and identify patterns in image data.
- **Genetics:** Clustering is used to group genes that have similar expression patterns and identify gene networks that work together in biological processes.
- **Finance:** Clustering is used to identify market segments based on customer behavior, identify patterns in stock market data, and analyze risk in investment portfolios.
- **Customer Service:** Clustering is used to group customer inquiries and complaints into categories, identify common issues, and develop targeted solutions.
- **Manufacturing**: Clustering is used to group similar products together, optimize production processes, and identify defects in manufacturing processes.
- **Medical diagnosis:** Clustering is used to group patients with similar symptoms or diseases, which helps in making accurate diagnoses and identifying effective treatments.
- **Fraud detection:** Clustering is used to identify suspicious patterns or anomalies in financial transactions, which can help in detecting fraud or other financial crimes.
- **Traffic analysis:** Clustering is used to group similar patterns of traffic data, such as peak hours, routes, and speeds, which can help in improving transportation planning and infrastructure.
- **Social network analysis:** Clustering is used to identify communities or groups within social networks, which can help in understanding social behavior, influence, and trends.
- **Cybersecurity:** Clustering is used to group similar patterns of network traffic or system behavior, which can help in detecting and preventing cyberattacks.
- **Climate analysis:** Clustering is used to group similar patterns of climate data, such as temperature, precipitation, and wind, which can help in understanding climate change and its impact on the environment.
- **Sports analysis:** Clustering is used to group similar patterns of player or team performance data, which can help in analyzing player or team strengths and weaknesses and making strategic decisions.

- **Crime analysis:** Clustering is used to group similar patterns of crime data, such as location, time, and type, which can help in identifying crime hotspots, predicting future crime trends, and improving crime prevention strategies.

Clustering Algorithms (or Methods):

1. K-Means clustering

In K-means clustering, data is grouped in terms of characteristics and similarities. K is a letter that represents the number of clusters. For example, if K=5, then the number of desired clusters is 5. If K=10, then the number of desired clusters is 10.

Key concepts

Squared Euclidean distance and cluster inertia are the two key concepts in K-means clustering. Learning these concepts will help understand the algorithm steps of K-means clustering.

- *Squared Euclidean distance*: If we have two points x and y, and the dimensional space given by m, the squared Euclidean distance will be given as:

$$d(\boldsymbol{x}, \boldsymbol{y})^2 = \sum_{j=1}^{m} \left(x_j - y_j\right)^2 = \left\|\boldsymbol{x} - \boldsymbol{y}\right\|_2^2$$

- *Cluster inertia*: This refers to the Sum of Squared Errors in the cluster. We give the cluster inertia as:

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{k} w^{(i,j)} \left\|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}^{(j)}\right\|_2^2$$

In the equation above, $\mu(j)$ represents cluster j centroid. If x(i) is in this cluster(j), then w(i,j)=1. If it's not, then w(i,j)=0.

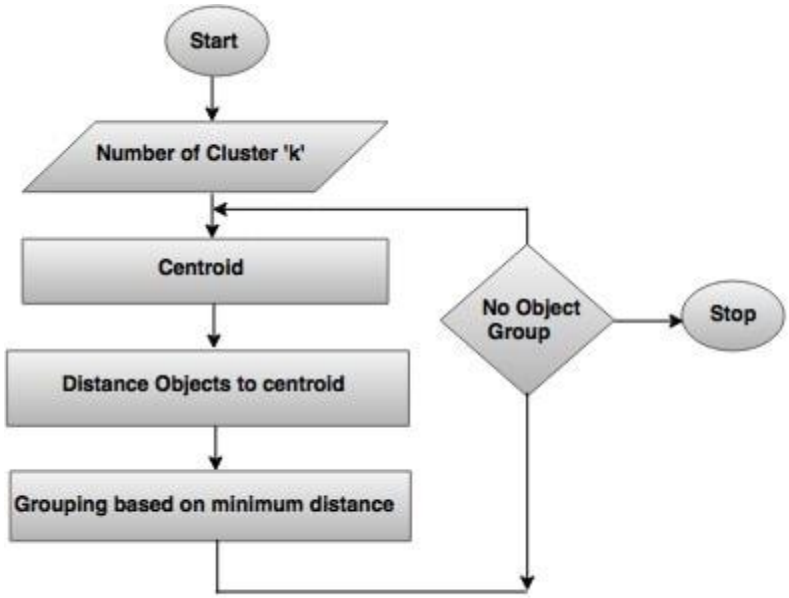
Based on this information, we should note that the K-means algorithm aims at keeping the cluster inertia at a minimum level.

Algorithm steps

1. Choose the value of K (the number of desired clusters). We can choose the optimal value of K through three primary methods: field knowledge, business

decision, and elbow method. The elbow method is the most commonly used. We can find more information about this method here.

2. Select K number of cluster centroids randomly.
3. Use the Euclidean distance (between centroids and data points) to assign every data point to the closest cluster.
4. Recalculate the centers of all clusters (as an average of the data points have been assigned to each of them).
5. Steps 3-4 should be repeated until there is no further change.



K- mean Algorithm

Advantages

- It is very efficient in terms of computation
- K-Means algorithms can be implemented easily

Disadvantages

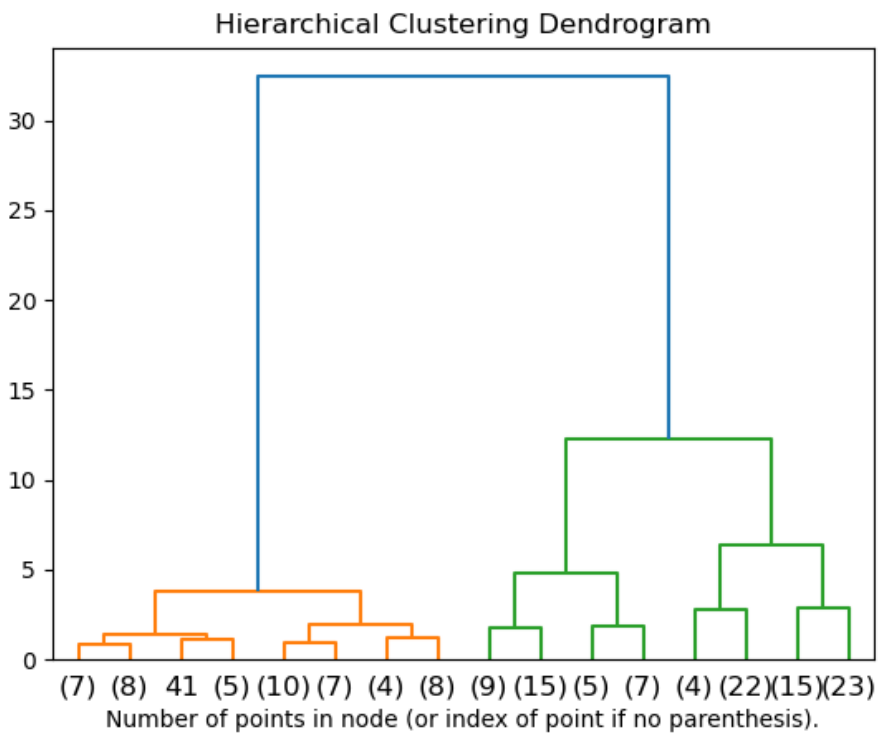
- K-Means algorithms are not effective in identifying classes in groups that are spherically distributed.
- The random selection of initial centroids may make some outputs (fixed training set) to be different. This may affect the entire algorithm process.

2. Hierarchical Clustering

In this type of clustering, an algorithm is used when constructing a hierarchy (of clusters). This algorithm will only end if there is only one cluster left.

Unlike K-means clustering, hierarchical clustering doesn't start by identifying the number of clusters. Instead, it starts by allocating each point of data to its cluster.

A dendrogram is a simple example of how hierarchical clustering works.



In the diagram above, the bottom observations that have been fused are similar, while the top observations are different.

Algorithm steps

1. Allocate each data point to its cluster.
2. Use Euclidean distance to locate two closest clusters. We should merge these clusters to form one cluster.
3. Determine the distance between clusters that are near each other. We should combine the nearest clusters until we have grouped all the data items to form a single cluster.

Advantages

- The representations in the hierarchy provide meaningful information.
- It doesn't require the number of clusters to be specified.
- It's resourceful for the construction of dendrograms.

Disadvantages

- Hierarchical models have an acute sensitivity to outliers. In the presence of outliers, the models don't perform well.
- The computation need for Hierarchical clustering is costly.

3. Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

This is a density-based clustering that involves the grouping of data points close to each other. We mark data points far from each other as outliers. It then sort data based on commonalities.

Key concepts

MinPts: This is a certain number of neighbors or neighbor points

Epsilon neighbourhood: This is a set of points that comprise a specific distance from an identified point. The distance between these points should be less than a specific number (epsilon).

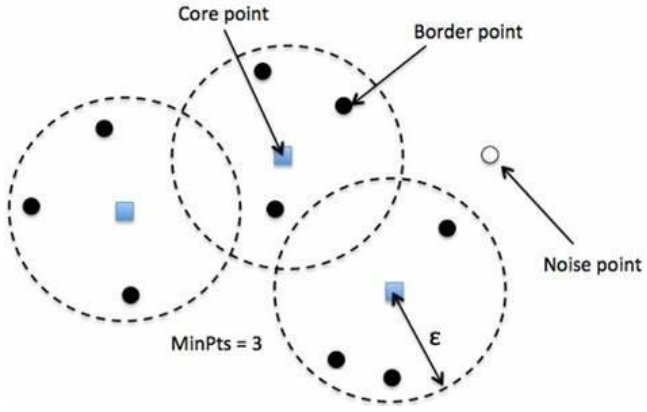
Core Point: This is a point in the density-based cluster with at least MinPts within the epsilon neighborhood.

Border point: This is a point in the density-based cluster with fewer than MinPts within the epsilon neighborhood.

Noise point: This is an outlier that doesn't fall in the category of a core point or border point. It's not part of any cluster.

Algorithm steps

1. In the first step, a core point should be identified. The core point radius is given as ε. Create a group for each core point.
2. Identify border points and assign them to their designated core points.
3. Any other point that's not within the group of border points or core points is treated as a noise point.



Advantages

- It doesn't require a specified number of clusters.
- It's very resourceful in the identification of outliers.
- It offers flexibility in terms of the size and shape of clusters.

Disadvantages

- It's not effective in clustering datasets that comprise varying densities.
- Failure to understand the data well may lead to difficulties in choosing a threshold core point radius.
- In some rare cases, we can reach a border point by two clusters, which may create difficulties in determining the exact cluster for the border point.

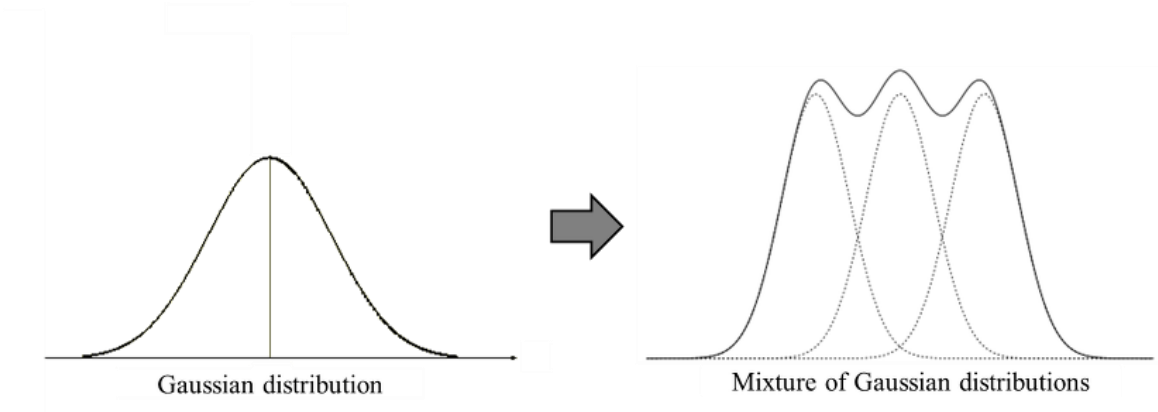
4. Gaussian Mixture Models (GMM)

This is an advanced clustering technique in which a mixture of Gaussian distributions is used to model a dataset. These mixture models are probabilistic. GMM clustering models are used to generate data samples.

In these models, each data point is a member of all clusters in the dataset, but with varying degrees of membership. The probability of being a member of a specific cluster is between 0 and 1.

In Gaussian mixture models, the key information includes the latent Gaussian centers and the covariance of data. This makes it similar to K-means clustering.

The following diagram shows a graphical representation of these models.



Algorithm steps

1. Initiate K number of Gaussian distributions. This is done using the values of standard deviation and mean.

2. Expectation Phase-Assign data points to all clusters with specific membership levels.
3. Maximization Phase-The Gaussian parameters (mean and standard deviation) should be re-calculated using the ‘expectations’.
4. Evaluate whether there is convergence by examining the log-likelihood of existing data.
5. Repeat steps 2-4 until there is convergence.

Advantages

- It offers flexibility in terms of size and shape of clusters.
- Membership can be assigned to multiple clusters, which makes it a fast algorithm for mixture models.

Disadvantages

- If a mixture consists of insufficient points, the algorithm may diverge and establish solutions that contain infinite likelihood. This may require rectifying the covariance between the points (artificially).
- A sub-optimal solution can be achieved if there is a convergence of GMM to a local minimum.