EM-Algorithm | ML LAB 8

What is Expectation-Maximization (EM Algorithm)?

- The expectation-maximization algorithm is an approach for performing maximum likelihood estimation in the presence of latent variables.
- Latent variables are variables that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed (directly measured).
- It is an effective and general approach and is most commonly used for density estimation with missing data, such as clustering algorithms like the Gaussian Mixture Model.

EM Algorithm

Step 1: ***Estimation*** **(E)** ***step:*** Calculate $Q(h'|h)$ using the current hypothesis h and the observed data X to estimate the probability distribution over Y.

$$Q(h' \mid h) \leftarrow \mathbf{E}[\ln P(Y|h')|h,X]$$

Step 2: ***Maximization (M) step:*** Replace hypothesis h by the hypothesis h' that maximizes this Q function.

$$h \leftarrow \underset{h'}{\operatorname{argmax}}\, Q(h'|h)$$

When the function Q is continuous, the EM algorithm converges to a stationary point of the likelihood function $P(Y|h')$. When this likelihood function has a single maximum, EM will converge to this global maximum likelihood estimate for **h'**. Otherwise, it is guaranteed only to converge to a local maximum.

Step 1 : Estimating the values for the latent variables

Sep 2 : Maximizes and optimizing the model

Repeat those two steps until convergence

What is Clustering?

- **Clustering** is one of the most common exploratory data analysis technique
- We can define it has the task of identifying subgroups in the data such that data points in the same subgroup (cluster) are very similar while data points in different clusters are very different
- We use similarity measure such as euclidean-based distance or correlation-based distance.

Clustering use cases

- **document classification**
- **delivery store optimization**
- **identifying crime localities**
- **customer segmentation**
- **cyber-profiling criminals**

What is K-Means Clustering?

- **Kmeans** algorithm is an iterative algorithm
- The algorithm tries to partition the dataset into **k** pre-defined distinct non-overlapping cluster
- Each data point will belong to only one cluster
- It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid
- The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

Steps in K-Means Clustering

Step 1:Initialize *k centroids=number of clusters* randomly or smartly

Step 2:Assign each data point to the closest centroid based on euclidian distance, thus forming the cluster

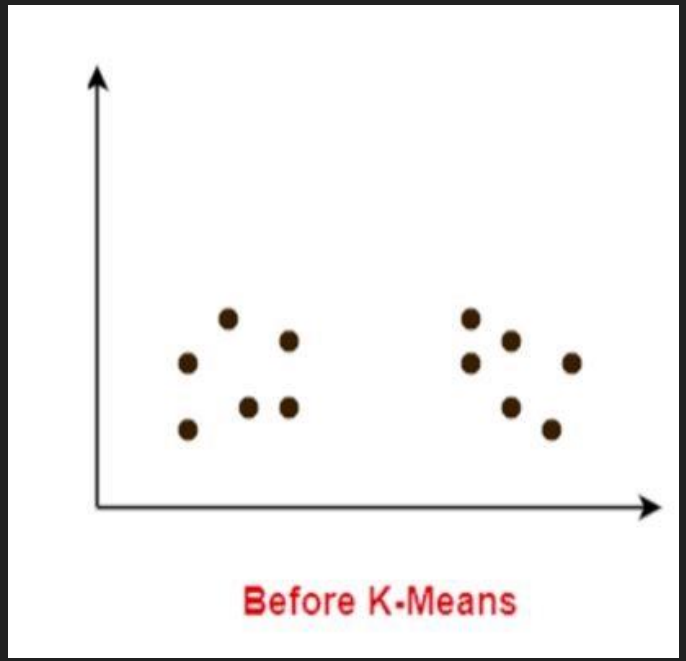
Step 3:Move centers to the average of all points in the cluster

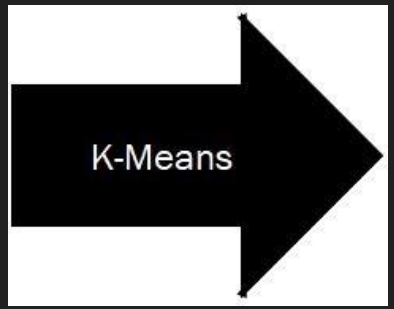
What does K-Means use to group the data points?

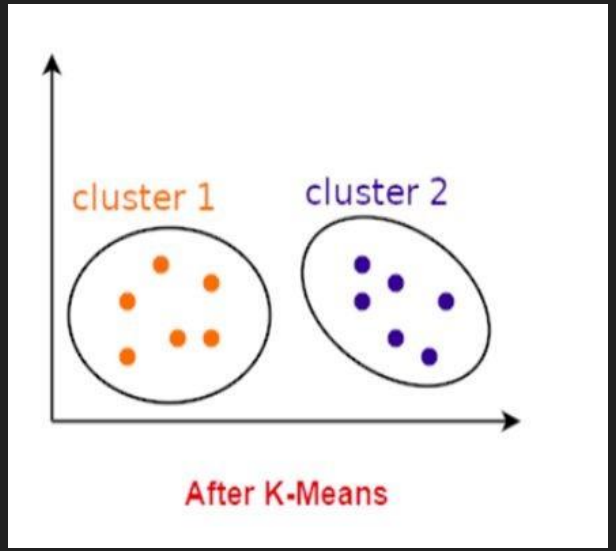
- It tries to minimize & optimize ***within-cluster sum-of-squared-distances*** or ***inertia*** of each cluster.
- When ***inertia*** value does not minimize further, algorithm converges. Thus, iteration stops.

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

Pictorial Representation of K-Means algorithm







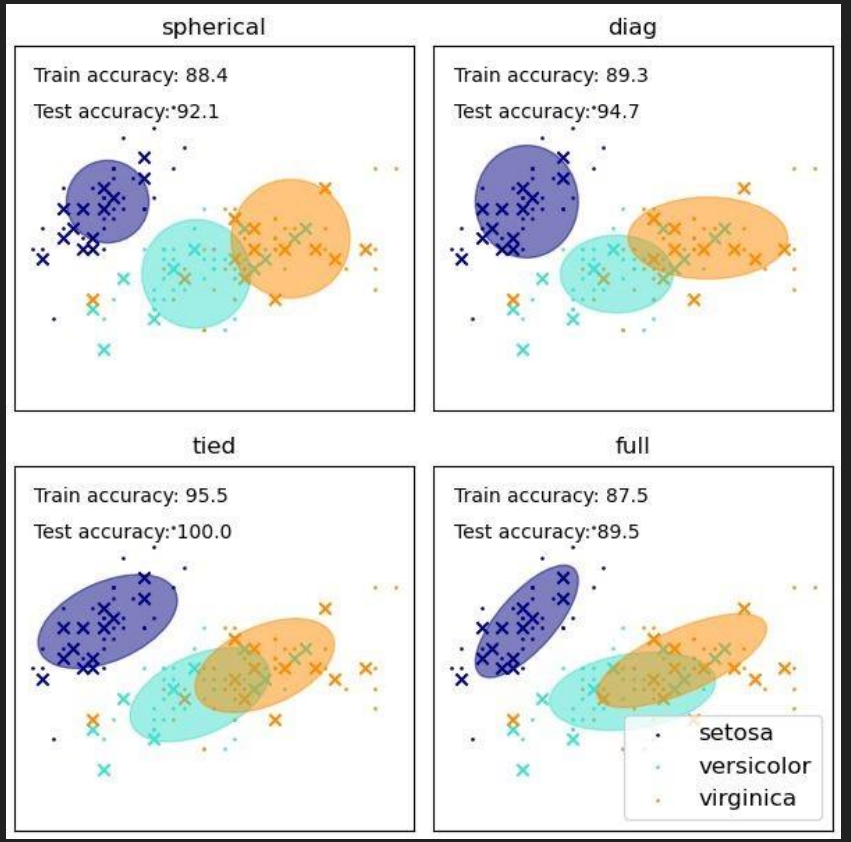
Disadavanatges of k-Means

- K-means is that it is a hard clustering method, which means that it will associate each point to one and only one cluster.
- A limitation to this approach is that there is no uncertainty measure or probability that tells us how much a data point is associated with a specific cluster.

What is Gaussian Mixture Model?

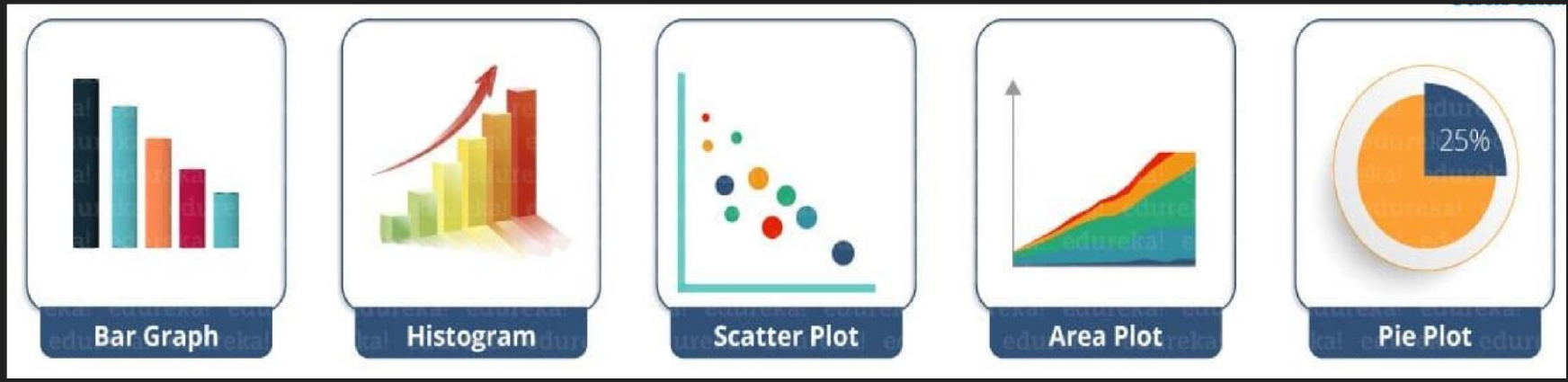
- A Gaussian Mixture is a function that is comprised of several Gaussians, each identified by **k ∈ {1,…, K}**, where **K** is the number of clusters of our dataset. Each **Gaussian k** in the mixture is comprised of the following parameters:
- A **mean μ** that defines its centre.
- A **covariance Σ** that defines its width.
- A **mixing probability π** that defines how big or small the Gaussian function will be.

Few example plots of gaussian mixture model



What is Matplotlib?

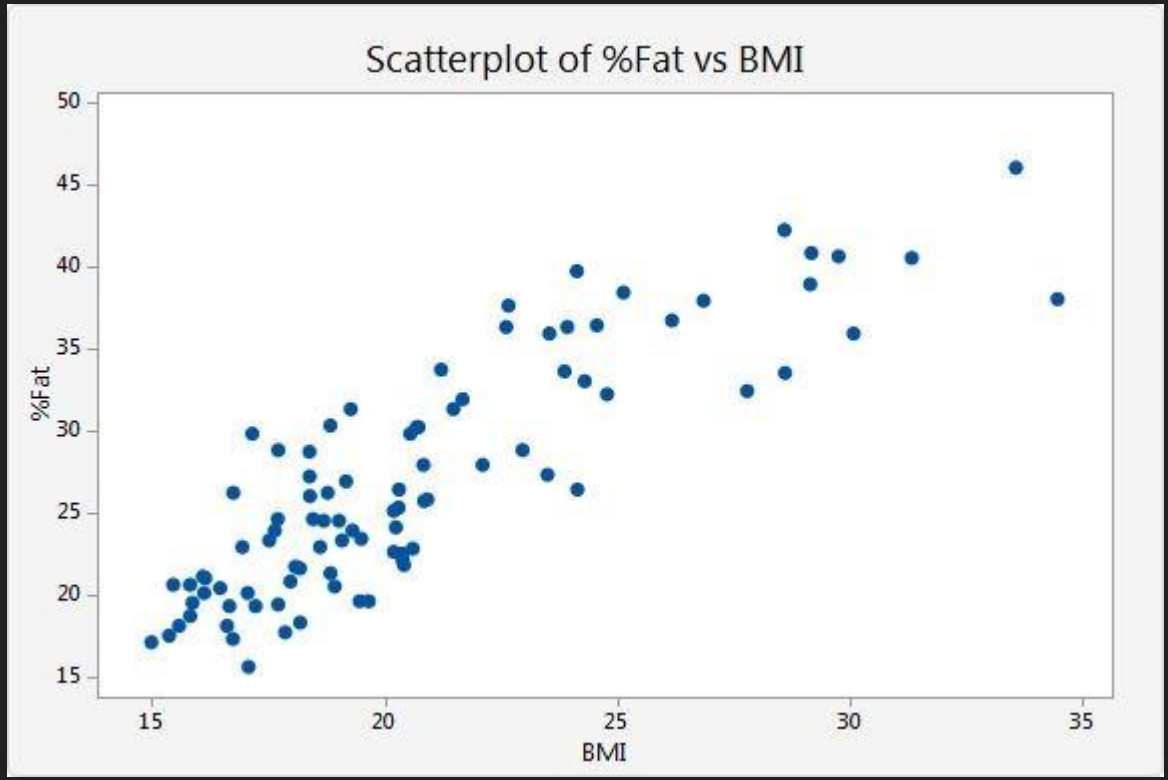
Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy.



Few functions to know in Matplotlib

- **plot():** plots x versus y graph
- **subplot(nrows, ncols, index)**: The two integer arguments to this function specify the number of rows and columns of the subplot grid.
- **xlim()**: is used to get or set the x-limits of the current axes.
- **ylim():** is used to get or set the y-limits of the current axes.
- **title():** is used give the title the plot
- **xlabel()**: sets the label for x axes
- **ylabel()**: sets the label for y axes
- **scatter()**: plots a scatter plot

Example of Scatter Plot


Scatterplot of %Fat vs BMI
%Fat
BMI
50
45
40
35
30
25
20
15
15
20
25
30
35