Experiment No: 5 06/02/2025

title: Expectation Maximization Algorithm for clustering.

Aim: To Implement the Expection - Maximization (EM) algorithm for estimating the parameters of a gaussian mixture model (GMM) and apply it to cluster of data point.

Theory:

Expectation maximization (EM) Algorithm: The expectation maximization (EM) Algorithm is an iterative optimization technique used to estimate parameters in probabilistic models when some data is missing or hidden. It is widely used in unsupervised learning for clustering, particularly in Gaussian Mixture Model.

A gaussian mixture model assume that data points are generated from a mixture of multiple gaussian distributed, each with its own mean and variance. The EM algorithm finds the optimal parameter for each Gaussian components.

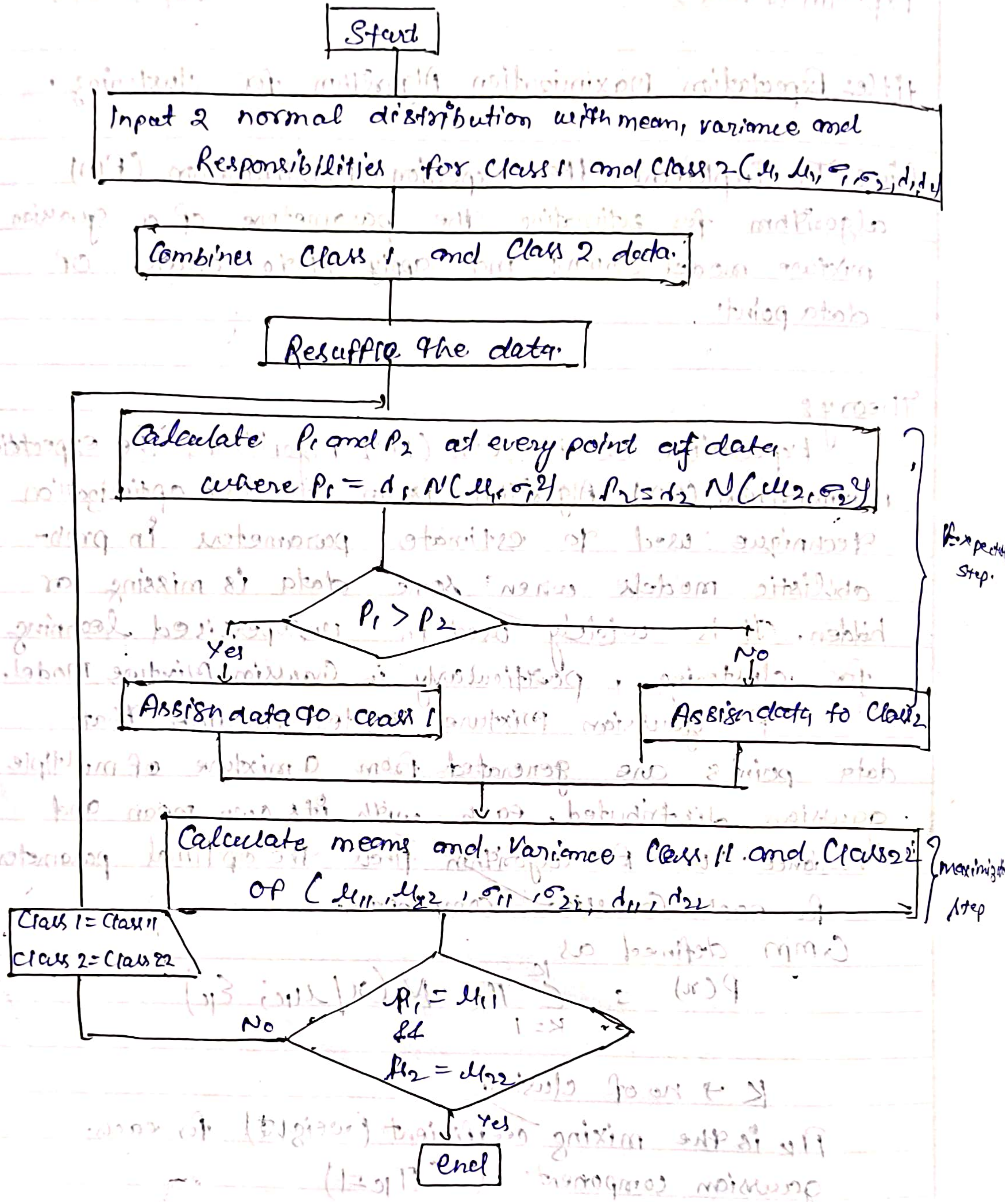
GMM defined as

$$P(x) = \sum_{K=1}^{K} \pi_K \cdot N(x \mid \mu_K ; \Sigma_K)$$

$K \rightarrow$ no of cluster

π_K is the mixing coefficient (weight) for each gaussian component ($\Sigma \pi_K = 1$)

Flowchart:


Start
Input 2 normal distribution with mean, variance and Responsibilities for Class 1 and Class 2 ($\mu_1, \mu_2, \sigma_1, \sigma_2, d_1, d_2$)
Combines Class 1 and Class 2 data.
Resuffle the data.
Calculate P_1 and P_2 at every point of data where $P_1 = d_1 N(\mu_1, \sigma_1^2)$ $P_2 = d_2 N(\mu_2, \sigma_2^2)$
Expectation Step.
$P_1 > P_2$
Yes
No
Assign data to class 1
Assign data to Class 2
Calculate means and Variance Class 11 and Class 22 of ($\mu_{11}, \mu_{22}, \sigma_{11}, \sigma_{22}, d_{11}, d_{22}$)
Maximization step
Class 1 = Class 11
Class 2 = Class 22
$\mu_1 = \mu_{11}$ && $\mu_2 = \mu_{22}$
No
Yes
End

Gaussian probability density function (Pdf):

$$N(x|\mu_k, \Sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(\frac{-(x_i-\mu_k)^2}{2\sigma_k^2}\right)$$

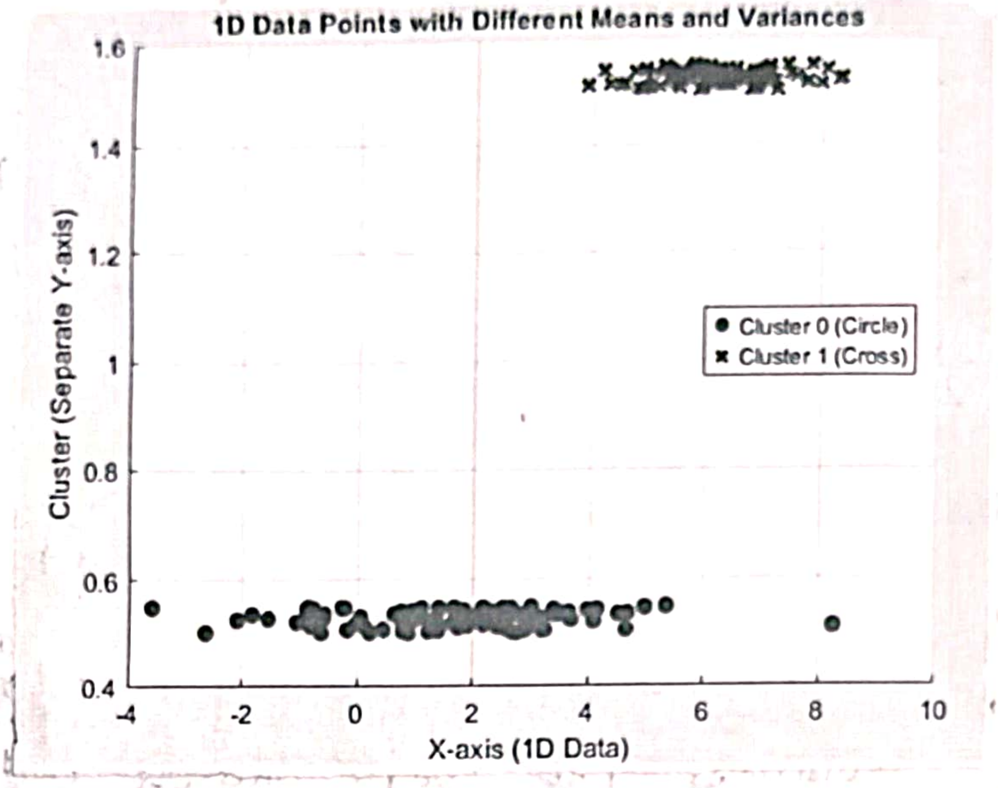
$P(x_i|b)$

$\sigma_k \rightarrow$ covariance (matrix)
$\mu_k \rightarrow$ mean (vector)

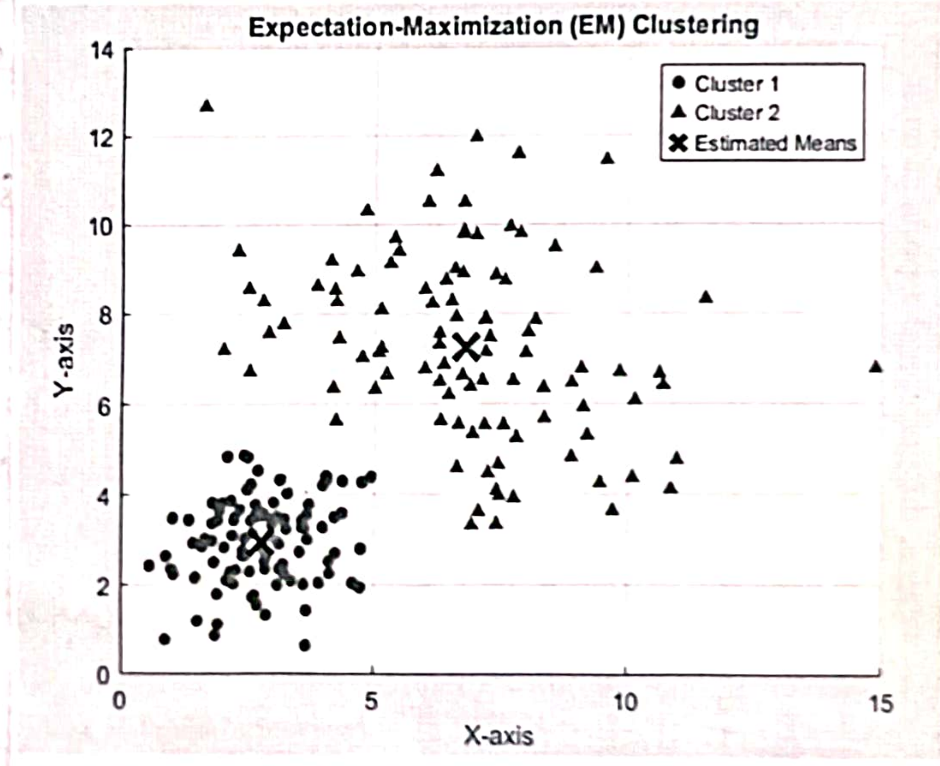
Algorithm:

⇒ Initilize the Input the normal distribution with means, variance and weights coefficient for 2 Clusters.
⇒ Combines the both clusters data.
⇒ Suffle the data.
⇒ Calculate the probabilites for the both cluster. where the $P_1 = \lambda_1 \cdot N(\mu_1, \sigma_1^2)$ and $P_2 = \lambda_2 N(\mu_2, \sigma_2^2)$
⇒ Assign the data with the highest Probabity cluster.
⇒ Calculate the means and variance the new assign clusters.
⇒ if the new means and new variance is same the previous means and variance Respectivily assign that particular cluster. Other wise iterate the Step-4.
→ end

Result:



1D Data Points with Different Means and Variances
Cluster (Separate Y-axis)
X-axis (1D Data)
Cluster 0 (Circle)
Cluster 1 (Cross)
1.6
1.4
1.2
1
0.8
0.6
0.4
-4
-2
0
2
4
6
8
10



Expectation-Maximization (EM) Clustering
Cluster 1
Cluster 2
Estimated Means
Y-axis
X-axis
14
12
10
8
6
4
2
0
0
5
10
15

Result: Taking random datapoint in 1D & 2D plane. (100 samples)
for 1D → initilize mean [2, 6]
variance randomly choose b/w the 0 - 2

for 2D → Initilize means [1, 5; 7, 8];
co-varience matrix 1:- $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, Covariance matrix:- $\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$

Discussion!

→ The algorithm starts with arbitrary intial means and variance.
→ A responsibility matrix is created to store the probability that each data point belongs to a given cluster.
→ Computes the probability of each data point belonging to each cluster, using Gaussian probability density function (Pdf)
→ Responsibilities (γ) are computed based on these probabilites.
→ Each datapoints is assigned to the cluster with the highest responsibilities value
→ The Expectation - Maximization Algorithm should be able to estimate these cluster assignments without prior knowledge of them based purly on the distribution of data.

Conclusion:

The Expectation maximization Algorithm clusters the given 2D and 1D data into two groups based on their underlying Gaussian distribution. It iteratively refines the cluster parameters. updating means and covariance matrics until convergence. The soft clustering approach assign probabilities. rather than strict label making Em effective for overlapping clusters. Although Em is sensitive to initilization. It perform well with properly chosen starting values. Enhancement like better initilization and regularization can further improve robustness, Overall, Em is a powerful clustering method widely used in the real world application.

13/02/25