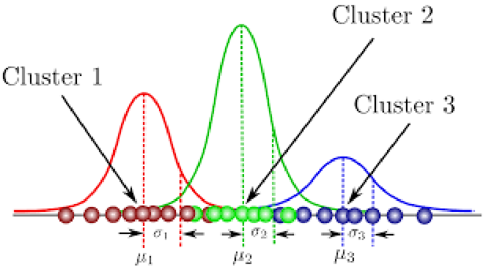
Summer of Code 2024 - Final Project
Gaussian Mixture Model using Expectation-Maximization

Tanmay Mandaliya
22B1037

1 Introduction

Gaussian Mixture Models (GMMs) are probabilistic models that assume all the data points are generated from a mixture of several Gaussian distributions with unknown parameters. GMMs are used for clustering, where each Gaussian represents a cluster.



2 How Gaussian Mixture Model (GMM) Algorithm Works

GMM can be considered a probabilistic version of KMeans. While KMeans uses a distance-based approach, GMM uses a probabilistic approach, assuming that the dataset consists of multiple Gaussian distributions (a mixture of Gaussians).

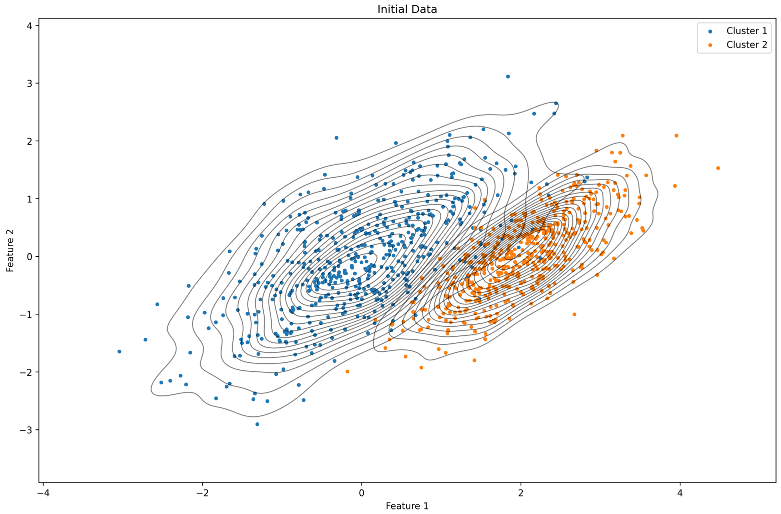


Figure 1: Initial Data

2.1 Initialization

Initialize the parameters (weights, means, and covariances) of the Gaussian components.

2.2 E-Step (Expectation)

Calculate the probability (responsibility) that each data point belongs to each Gaussian component.

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i|\mu_j, \Sigma_j)} \tag{1}$$

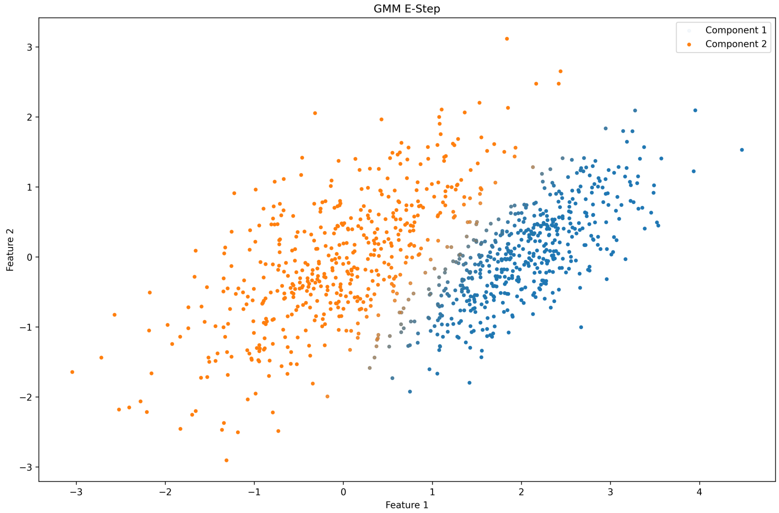


Figure 2: GMM E-Step

2.3 M-Step (Maximization)

Update the parameters using the calculated responsibilities.

$$\pi_k = \frac{N_k}{N} \tag{2}$$

$$\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} x_i \tag{3}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T \tag{4}$$

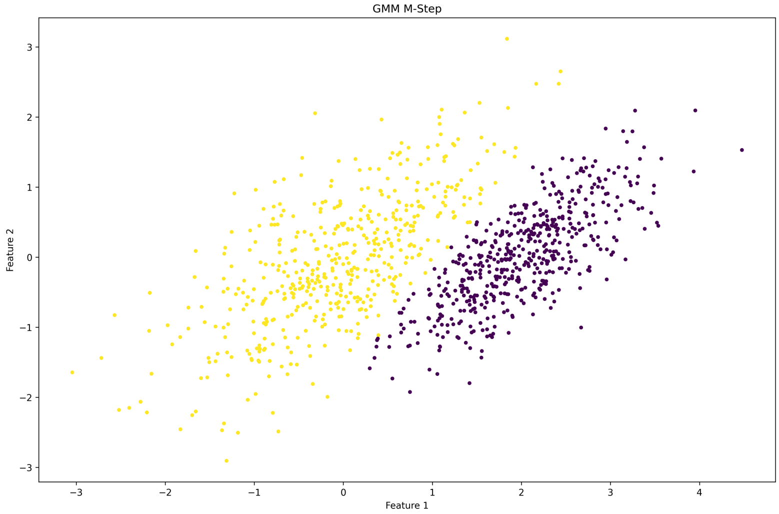


Figure 3: GMM M-Step

3 Comparison with KMeans

KMeans is a popular clustering algorithm but has some limitations:

- Assumes clusters are spherical and equally sized.
- Assigns each data point to a single cluster (hard clustering).

GMM, on the other hand, offers several advantages:

3.1 Flexibility in Cluster Shape

KMeans assumes clusters are spherical and equally sized, which is often not true. GMM allows clusters to have different shapes and sizes.

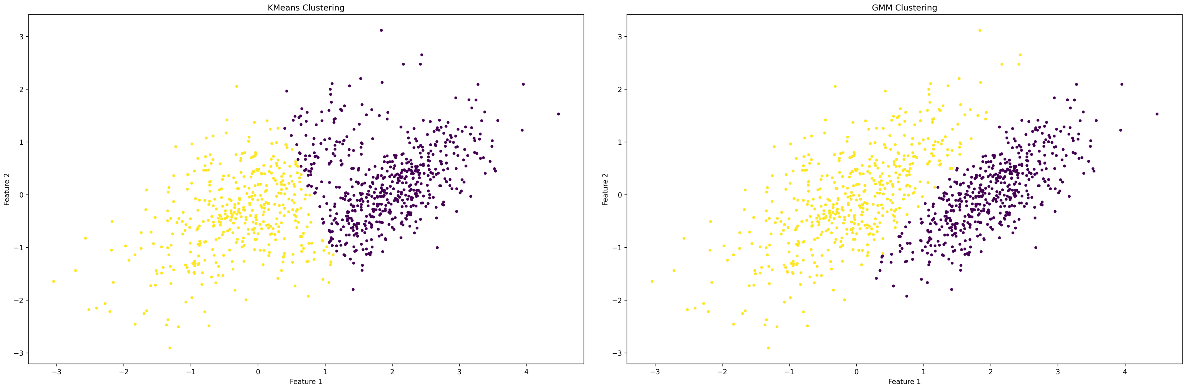


Figure 4: Comparison of KMeans (left) and GMM (right) clustering. GMM captures the true cluster shapes better.

3.2 Probabilistic Assignment

KMeans assigns each point to exactly one cluster, while GMM assigns probabilities, allowing for more nuanced clustering.

3.3 Better Handling of Outliers

GMM can better handle outliers by giving them lower probabilities, whereas KMeans is more sensitive to outliers.

4 Conclusion

The GMM, combined with the EM algorithm, provides a powerful tool for clustering and density estimation. Compared to KMeans, GMM is more flexible and can capture complex cluster shapes better.

End of Report