

Gaussian Mixture Model - Two-dimensional dataset

Aim: To apply the Gaussian Mixture Model (GMM) clustering algorithm to a two-dimensional dataset, to identify clusters within the data.

Algorithm:

The Gaussian Mixture Model (GMM) is a probabilistic clustering algorithm that assumes the data is generated from a mixture of several Gaussian distributions with unknown parameters. Unlike K-Means, which assigns points to the nearest cluster centroid, GMM assigns probabilities to each data point belonging to a particular cluster.

The probability of a data point x_i belonging to cluster k is given by:

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

Where:

- $\mathcal{N}(x_i|\mu_k, \Sigma_k)$ is the Gaussian distribution with mean μ_k and covariance Σ_k .
- π_k is the mixing coefficient for cluster k .
- K is the total number of clusters.

Step 1: Import Libraries

- Import necessary Python libraries such as NumPy, Matplotlib, and Scikit-learn.

Step 2: Load the Dataset

- Load the Iris dataset and extract only the sepal length and sepal width features (bivariate dataset).

Step 3: Scale the Data

- Normalize the dataset using StandardScaler to ensure that features contribute equally to clustering.

Step 4: Apply GMM

- Initialize the Gaussian Mixture Model with n_components=3 to find three clusters.
- Fit the model to the scaled data and predict cluster labels.

Step 5: Visualize the Clusters

- Plot the clustered data points with distinct colors.
- Display a color bar to indicate the assigned clusters.
- Label the axes and provide a title for the graph.

Import the libraries

```
In [19]: import numpy as np
import matplotlib.pyplot as plt
from sklearn import datasets
import warnings
from sklearn.mixture import GaussianMixture
from sklearn.preprocessing import StandardScaler
warnings.filterwarnings("ignore")
```

Load the Dataset

```
In [20]: iris = datasets.load_iris()
X = iris.data[:, :2]
```

Scale the data

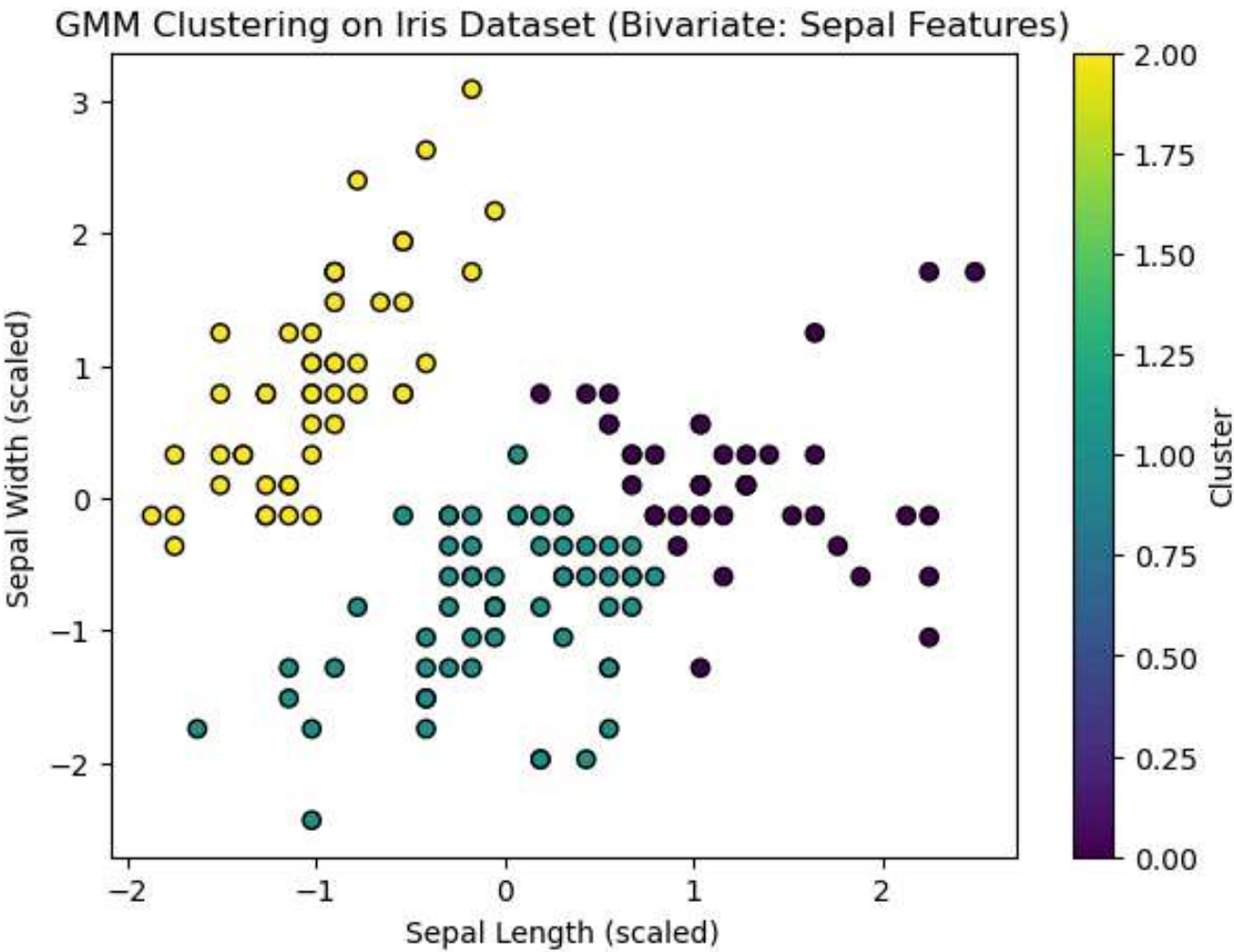
```
In [21]: scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Apply the Gaussian Mixture Model

```
In [22]: gmm = GaussianMixture(n_components=3, random_state=42)
gmm_labels = gmm.fit_predict(X_scaled)
```

Plot the graph

```
In [23]: plt.figure(figsize=(7,5))
plt.scatter(X_scaled[:, 0], X_scaled[:, 1], c=gmm_labels, cmap='viridis', edgecolors='k')
plt.colorbar(label='Cluster')
plt.xlabel('Sepal Length (scaled)')
plt.ylabel('Sepal Width (scaled)')
plt.title('GMM Clustering on Iris Dataset (Bivariate: Sepal Features)')
plt.show()
```



Result

The Gaussian Mixture Model successfully clustered the two-dimensional dataset into three clusters based on probability distributions.
