Experiment 8 - Comparison of Clustering Algorithms

May 1, 2023

1 Experiment Details

1.1 Submitted By

Desh Iyer, 500081889, Year III, AI/ML(H), B5

1.2 Problem Statement

Apply the Expectation-Maximization (EM) algorithm to cluster a set of data stored in a .CSV file. Use the same dataset for clustering using the k-Means algorithm. Compare the results of these two algorithms and comment on the quality of clustering.

1.3 Theory

Clustering is a technique used in unsupervised learning to group together similar data points. There are many different clustering algorithms, including k-Means and EM. The k-Means algorithm is a simple and widely-used clustering algorithm that partitions data into k clusters based on distance. The EM algorithm is a more complex clustering algorithm that involves estimating the probability distribution of the data.

The EM algorithm consists of two main steps: the E-step and the M-step. In the E-step, the algorithm estimates the probabilities of each data point belonging to each cluster. In the M-step, the algorithm updates the parameters of the probability distribution based on these probabilities. These two steps are repeated until convergence.

The k-Means algorithm also has two main steps: the assignment step and the update step. In the assignment step, each data point is assigned to the nearest centroid. In the update step, the centroids are updated based on the mean of the data points assigned to them. These two steps are repeated until convergence.

1.4 Steps

Here are the steps to apply the EM and k-Means algorithms for clustering:

1.4.1 EM Algorithm

1. Load the data from the .CSV file.
2. Initialize the parameters of the probability distribution.
3. Repeat until convergence:
    1. E-step: Calculate the probability of each data point belonging to each cluster.

2. M-step: Update the parameters of the probability distribution based on these probabilities.
4. Assign each data point to the cluster with the highest probability.

1.4.2 k-Means Algorithm

1. Load the data from the .CSV file.
2. Initialize k centroids randomly.
3. Repeat until convergence:
    1. Assignment step: Assign each data point to the nearest centroid.
    2. Update step: Update the centroids based on the mean of the data points assigned to them.
4. Assign each data point to the cluster with the nearest centroid.

2 Import Required Libraries

[ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn.mixture import GaussianMixture
```

[ ]:
```
# Load the IRIS dataset
iris = datasets.load_iris()
X = iris.data
```

[ ]:
```
# Define the number of clusters
k = 3
```

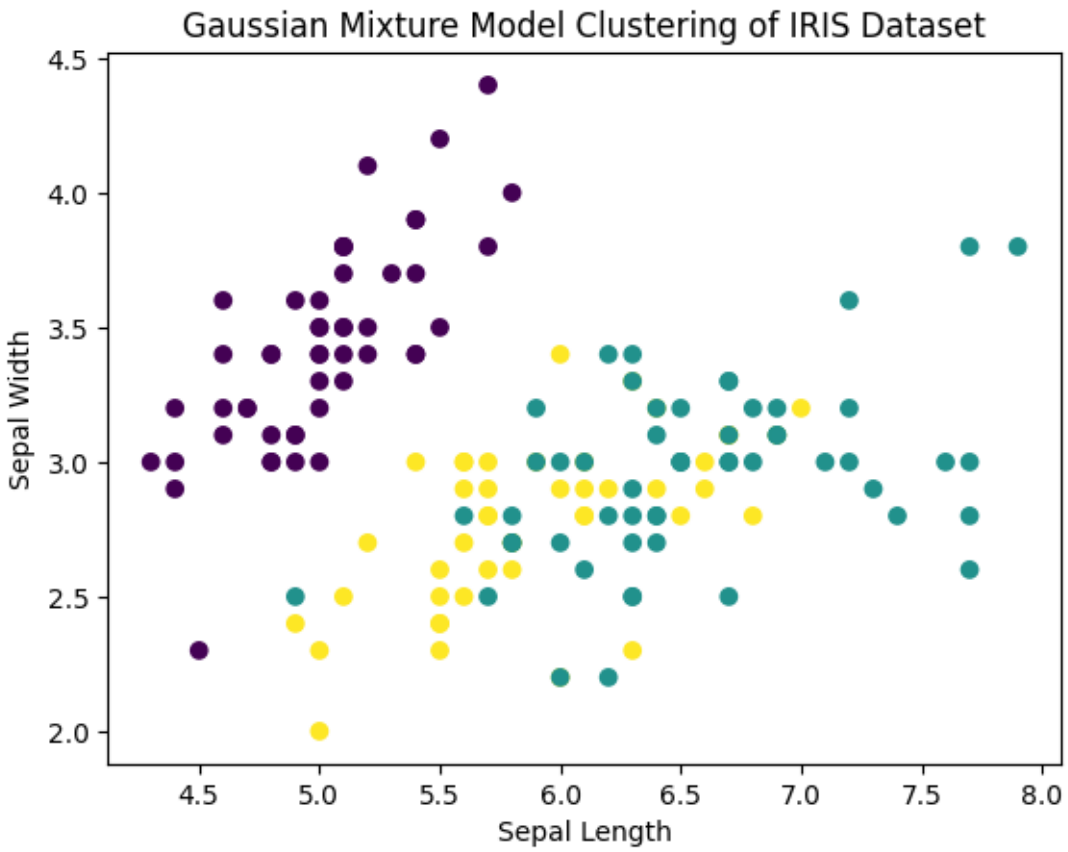
3 Clustering using a Gaussian Mixture and EM Algorithm

[ ]:
```
# Fit the Gaussian mixture model using the EM algorithm
gmm = GaussianMixture(n_components=k, covariance_type='full', random_state=0)
gmm.fit(X)
```

[ ]:
```
GaussianMixture(n_components=3, random_state=0)
```

[ ]:
```
# Get the predicted cluster labels
labels = gmm.predict(X)
```

[ ]:
```
# Create a scatter plot of the IRIS dataset with the predicted cluster labels
plt.scatter(X[:, 0], X[:, 1], c=labels)
plt.xlabel('Sepal Length')
plt.ylabel('Sepal Width')
plt.title('Gaussian Mixture Model Clustering of IRIS Dataset')
plt.show()
```



4 Clustering using KNN

[ ]:
```
# Define the number of clusters
k = 3
```

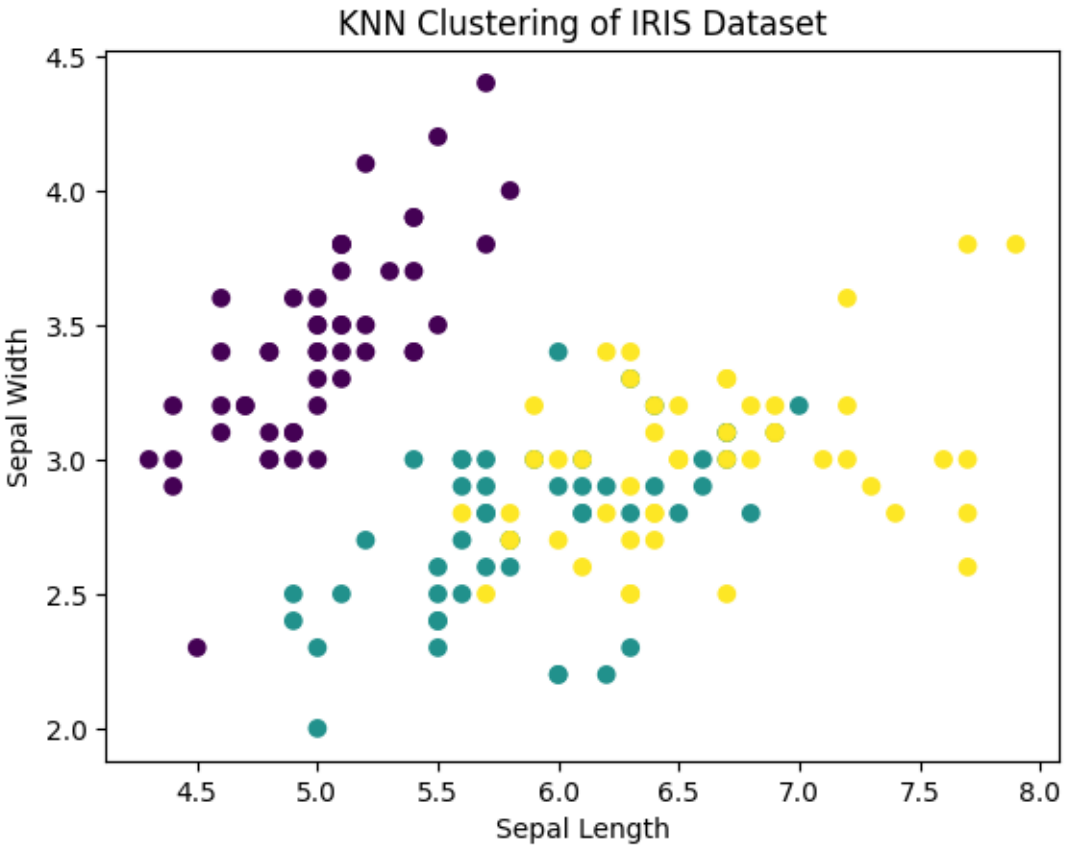
[ ]:
```
# Fit the KNN model
knn = KNeighborsClassifier(n_neighbors=k)
knn.fit(X, iris.target)
```

[ ]:
```
KNeighborsClassifier(n_neighbors=3)
```

[ ]:
```
# Get the predicted cluster labels
labels = knn.predict(X)
```

[ ]:
```
# Create a scatter plot of the IRIS dataset with the predicted cluster labels
plt.scatter(X[:, 0], X[:, 1], c=labels)
plt.xlabel('Sepal Length')
plt.ylabel('Sepal Width')
plt.title('KNN Clustering of IRIS Dataset')
```

```
plt.show()
```



5 Conclusion

5.1 Quality of Clustering:

- GMM/EM: The GMM/EM algorithm is a probabilistic clustering method that models each cluster as a Gaussian distribution. It can capture complex nonlinear relationships between variables and can work well with high-dimensional data. It also provides a measure of uncertainty in the form of posterior probabilities, which can be used to identify ambiguous points. However, it assumes that the data is generated from a mixture of Gaussian distributions, which may not always be true, and the results can be sensitive to the choice of initialization and the number of clusters.
- KNN: The KNN algorithm is a distance-based clustering method that assigns each data point to the nearest cluster center. It is simple and easy to implement and can work well with small datasets and simple patterns. However, it can be sensitive to the choice of distance metric, the number of neighbors, and the distribution of the data. It also does not provide a measure of uncertainty or the underlying probability distribution.

5.2 Pros and Cons:

- GMM/EM:
  - Pros:
    - Provides a measure of uncertainty and posterior probabilities.
    - Can capture complex nonlinear relationships between variables.
    - Can work well with high-dimensional data.
  - Cons:
    - Assumes that the data is generated from a mixture of Gaussian distributions.
    - Results can be sensitive to the choice of initialization and the number of clusters.
    - Can be computationally expensive for large datasets.
- KNN:
  - Pros:
    - Simple and easy to implement.
    - Can work well with small datasets and simple patterns.
    - Does not assume any underlying distribution of the data.
  - Cons:
    - Can be sensitive to the choice of distance metric and the number of neighbors.
    - Does not provide a measure of uncertainty or the underlying probability distribution.
    - Can be computationally expensive for large datasets with many features.

Overall, the choice between GMM/EM and KNN depends on the specific characteristics of the data and the problem at hand. If the data is high-dimensional and has complex nonlinear relationships, GMM/EM may be a better choice. On the other hand, if the data is low-dimensional and has simple patterns, KNN may be a better choice.
