Week 11 - Clustering Kmeans, GMM and DBSCAN

DS3010 - Introduction to Machine Learning

Instructions

1. Provide commented, indented code. Variables should have meaningful names.
2. Submit one .ipynb file containing all answers. The name should be [student name][roll_number] assignment[number].ipynb
3. Read the questions carefully before answering. If a question asks to follow a particular approach or to use a specific data structure, then it must be followed.
4. Write questions in separate text blocks in Jupyter Notebook before the code blocks containing answers.
5. All plots should have appropriate axis labels, titles, and legends.

---

1. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

(4 points)

Generate a synthetic dataset using Scikit-learn's `make_circles` function for clustering.

You may use the following code to generate the dataset:

```
from sklearn.datasets import make_circles

# Generate concentric circle data
X, y = make_circles(n_samples=500, factor=0.5, noise=0.05, random_state=42)
```

a. Plot the scatter plot of the generated data points.**(1 points)**
b. Implement DBSCAN using Scikit-learn and experiment with different values of epsilon and min_samples to find the best parameters for clustering.**(2 points)**
c. Visualize the clusters along with the noise points. Use different colors for each cluster and a different marker for noise points.**(1 points)**
.

2. K-Means Clustering (11 points)

Use the above generated dataset to solve the given problems.

a. Implement the K-Means clustering algorithm from scratch. Write a function to compute the inertia (sum of squared distances of samples to their closest cluster center).**(7 points)**
b. Use the elbow method to find the optimal number of clusters by plotting inertia values against different values of K.**(1 points)**
c. Perform K-Means clustering on the dataset using the optimal K value obtained.**(1 points)**
d. Visualize the clusters along with their centroids. Use distinct colors and markers for each cluster and label the axes. Include a legend.**(1 points)**

e. Explain how DBSCAN handles clusters of varying shapes and sizes compared to K-Means . Include the advantages and limitations of DBSCAN for the dataset you used**(1 points)**

---

3. Gaussian Mixture Model (GMM) (Optional)

Use the above generated dataset to solve the given problems.

a. Implement a Gaussian Mixture Model (GMM) clustering algorithm using Scikit-learn. Use the Bayesian Information Criterion (BIC) to determine the optimal number of components.
b. Plot the BIC values against the number of components and identify the optimal number of components.
c. Perform GMM clustering with the optimal number of components. Visualize the clusters with ellipses representing the covariance of each cluster. Also, show the cluster centers.
d. Write a short explanation of the difference between K-Means and GMM clustering.
