CSE 4/574 Assignment3

Due date: 22nd April 11:59PM (Submit on Brightspace)

Apply clustering algorithms to the Iris dataset to explore patterns and potentially identify natural groupings of iris flowers based on their measurements. (100 points)

Any function/library can be used.

Report in pdf format.

1. Load and Explore the Dataset: (1 point)
    a. Load the Iris dataset from the `sklearn.datasets` module.

    from sklearn.datasets import load_iris
    iris = load_iris()

    b. Display basic information about the dataset (e.g., number of samples, features, target classes).

2. Data Preprocessing: (**assignment 2 code can be used**) (1 point)
    c. Check for missing values and handle them if necessary (not typically needed for the Iris dataset).
    d. Scale or normalize the features for better clustering performance.

3. Clustering Algorithms: (18 points for report))
    In the report explain the assumptions, advantages and disadvantages of each algorithm.
    e. Clustering algorithms to apply:
        i. K-means clustering
        ii. Hierarchical clustering (Agglomerative clustering) – **self-learn**
        iii. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) – **self-learn**

4. Clustering Experimentation: (20 + 10 (report) points)
    f. Apply each clustering algorithm to the preprocessed dataset.
    g. Explore different parameter settings (e.g., number of clusters for K-means, linkage method for hierarchical clustering, epsilon and min_samples for DBSCAN). Write your observations in report

5. Evaluation: (10 + 10 points)
    h. Evaluate the quality of clustering using metrics like silhouette score (**self-learn, you can also use inertia and Dunn index taught in class**) and visual inspection:
    i. Silhouette score: a higher score indicates better-defined clusters.
    j. Visualize clusters using scatter plots or other appropriate visualizations. Write observations along with visualization in report.

6. Interpretation and Analysis: (10 + 10 points)
    k. Analyze the clustering results:
    l. Interpret the identified clusters and discuss their characteristics.
    m. Compare and contrast the effectiveness of different clustering algorithms for this dataset. Write observations along with visualization in report.

7. Visualization: (10 points)
    n. Visualize the clustered data:
    o. Plot clusters using scatter plots, with each cluster represented by a different color.
    p. Optionally, visualize centroids or dendrograms (for hierarchical clustering).