Project Requirements for K-means and DBSCAN Clustering

Introduction

This project aims to apply two clustering algorithms, K-means and DBSCAN, to different datasets and compare their results. The project will use the following datasets:

- file19.txt: a multivariate mammals dataset with 9 attributes and 66 observations, obtained from https://people.sc.fsu.edu/~jburkardt/datasets/hartigan/hartigan.html[1]
- s1.csv: a set of Gaussian clusters with 2 dimensions and 5000 observations, obtained from http://cs.joensuu.fi/sipu/datasets/s1.txt[2]

The project will consist of two phases: Phase 1 for K-means clustering and Phase 2 for DBSCAN clustering.

Objective: To perform K-means and DBSCAN clustering on given datasets and analyze the results.

Datasets:

1. HARTIGAN dataset: file19.txt (Multivariate mammals dataset; 9 columns and 66 rows)
2. s1.csv (Extracted from "s1.txt", Clustering Basic Benchmark; 5,000 observations of two dimensions)

Requirements:

Phase 1: K-means Clustering on file19.txt

1. Perform data cleanup.
2. Determine attributes to omit from the dataset before clustering.
3. Determine the number of clusters needed by running the WSS or Silhouette graph. Plot the graph using fviz_nbclust().
4. Run k-means clustering on the dataset to create the determined number of clusters. Plot the clusters using fviz_cluster().
5. Show the number of observations in each cluster, total SSE of the clusters, and SSE of each cluster.
6. Analyze each cluster to determine how the mammals are grouped in each cluster.

Phase 2: DBSCAN Clustering on s1.csv

1. Plot the dataset and describe what you observe in the plot.
2. Draw the scree plot for the optimal number of clusters using both the "wss" and "silhouette" methods.
3. Determine the appropriate number of clusters for K-Means clustering on this dataset.
4. Perform K-Means clustering on the dataset and plot the results.

5. Comment on how K-Means has clustered the dataset.
6. Determine a reasonable value for MinPts for this dataset.
7. Calculate the average distance of every point to its k nearest neighbors to find the value of ε (eps).
8. Determine the best value of ε that clusters the dataset into the expected number of clusters.
9. Plot the results of the DBSCAN algorithm on the dataset and state how many clusters you see.

Deliverables:

1. Source code files
2. A report detailing the process, findings, and analysis