K-Means Clustering Assignment

Problem Statement

The aim of this assignment is to utilize K-Means Clustering algorithms to analyze the provided dataset, extracting meaningful insights from its underlying structures. By exploring the datasets and employing K-Means Clustering techniques, students are expected to categorize the data into distinct clusters. This assignment emphasizes parameter tuning for optimal clustering results and requires interpretation of the clustering outcomes to derive valuable insights.

Guidelines

1. Foundational Knowledge
- Understand the principles of clustering and how data can be segmented into clusters using K-Means.
- Familiarize yourself with the K-Means Clustering algorithm and comprehend its principles and advantages.
- Recognize the importance of choosing an appropriate number of clusters (K) and initialization methods.

2. Data Exploration
- Analyze the dataset's structure and characteristics using various exploratory techniques such as scatter plots, boxplots, heatmaps, etc.
- Gain insights into the dataset's attributes to guide the clustering process.

3. Preprocessing and Parameter Selection
- Standardize features if required, as K-Means Clustering can be sensitive to feature scales.
- Choose an appropriate number of clusters (K) based on the dataset's characteristics.
- Select appropriate initialization methods (e.g., random, k-means++).

4. K-Means Clustering
- Implement the K-Means Clustering algorithm using chosen parameters and methods.
- Evaluate the clustering quality using metrics like inertia, silhouette score, etc.
- Iterate through different values of K to find the optimal number of clusters.

5. Cluster Analysis
- Analyze resulting clusters to understand their attributes and characteristics.
- Evaluate unclustered data points to derive conclusions about the dataset.
- Compare findings with initial exploratory analysis to reinforce insights.

Step-by-Step Approach to K-Means Clustering

1. Setup and Data Preparation
- Import necessary libraries: pandas, matplotlib, and Scikit-Learn.
- Load the dataset for clustering.
- Preprocess the data, ensuring standardized features if necessary.

2. K-Means Clustering Parameters
- Choose an appropriate number of clusters (K).
- Define initialization methods suitable for the dataset.

3. Performing K-Means Clustering
- Initialize the K-Means Clustering model with selected parameters.
- Apply the model on the prepared data.

4. Result Analysis
- Examine cluster labels and interpret the clusters formed.
- Visualize clusters to differentiate them using appropriate markers/colors.

5. Evaluation and Iteration
- Evaluate clustering quality using metrics like inertia, silhouette score, etc.
- Adjust the number of clusters and initialization methods if clustering results are unsatisfactory.

6. Interpretation and Conclusion
- Understand cluster patterns and distinctions.
- Decide on handling noise or outliers, if any.

Link to Dataset for the Assignment

- Credit Card Dataset for Clustering
https://www.kaggle.com/datasets/arjunbhasin2013/ccdata
- Shop Customer Data
https://www.kaggle.com/datasets/datascientistanna/customers-dataset
- EastWestAirlines Dataset
https://www.kaggle.com/datasets/singhnproud77/eastwestairlines-heirarchical-clustering