Part 1: K-means Strategy Project

Shachi Shah
ID#: 1217121828
School of Computing and Augmented Intelligence
Arizona State University Online
Azusa, California, United States of America
spshah22@asu.edu

I. Introduction

Clustering is a fundamental unsupervised machine learning technique used for data segmentation. In this project, we implemented K-Means clustering on a given dataset and analyzed the results using different values of *k* as the number of clusters. The project aimed to identify optimal cluster centroids, evaluate the loss function, and visualize the clustering process. The *Cluster Plots* and *Loss Function Plot* were used to assess cluster quality, while scatter plots helped visualize data segmentation.

II. Implementation

A. *Data Loading and Preprocessing*

a. The dataset was loaded from ‘AllSamples.npy’

b. The clustering process was initialized using function from ‘precode.py’

B. *K-Means Clustering Execution*

a. The clustering algorithm was executed for k=2 to k=10.

b. Initial cluster centers were assigned using ‘initial_S1’.

c. The final cluster centroids were computed using ‘KMeans’ from ‘sklearn.cluster’.

C. *Loss Function Computation*

a. The loss function (inertia) was calculated for each *k* value.

b. The results were plotted to determine the optimal number of clusters.

D. *Visualization*

a. An *Elbow Method Plot* was generated to analyze the loss function trend.

b. Cluster Scatter Plots were created for each *k* value to visually examine data segmentation.

III. Results

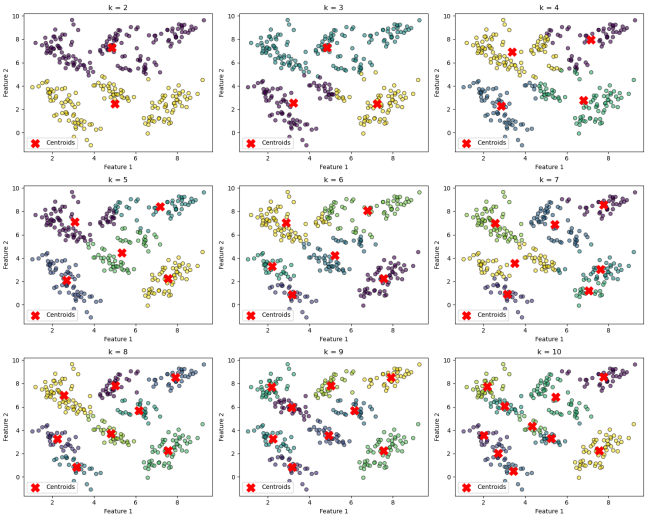

Figure 1: Cluster Plots

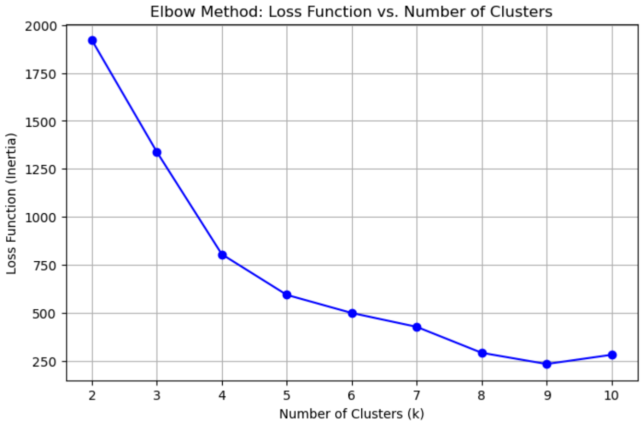

Figure 2: Loss Function Plot

The Elbow Method plot showed a significant drop in loss at a specific , indicating the optimal number of clusters (Figure 1). This sharp decline suggests that beyond this point, adding more clusters does not result in substantial improvement in clustering efficiency. The scatter plots provided further insight into the clustering structure, showcasing distinct formations that became increasingly refined as they increased. At lower values of , the clusters were broad and encompassed a variety of data points, sometimes grouping together observations that were visibly distinct. As increased, the clustering granularity improved, allowing for a more precise separation of data points, thereby capturing finer details of the data distribution (Figure 2). However, beyond a certain , the improvements in cluster separation became marginal, reinforcing the findings from the Elbow Method. This indicates that while increasing can provide more specific clusters, it reaches a point where additional clusters do not contribute significantly to better segmentation, leading to unnecessary complexity without meaningful gains.

IV. Observations

A. As *k* increased, the loss function value (inertia) decreased, but with diminishing returns.
B. The Elbow Method suggested an optimal *k* where the loss reduction rate slowed down.
C. Some clusters had overlapping points, requiring further tuning of cluster initialization.
D. Higher *k* values resulted in smaller, more refined clusters but at the cost of higher complexity.

V. Challenges and Solutions

A. *Improper Initialization Leading to Suboptimal Clustering:* Used predefined initialization from `precode.py` to ensure reproducibility.
B. *Formatting Issues in Test Cases:* Carefully followed the expected output format and verified using test functions.
C. *Selecting the Best k:* Used the Elbow Method to confirm the optimal number of clusters.

VI. Conclusion

The project successfully implemented K-Means clustering on the dataset, identifying meaningful clusters. The Elbow Method helped determine the best , while scatter plots visually confirmed the clustering results. The loss function plot provided additional validation by illustrating the diminishing returns of adding more clusters. Despite some initialization and formatting challenges, structured testing and visualization techniques allowed for a robust analysis. This project provided hands-on experience in clustering techniques and performance evaluation, crucial for real-world machine learning applications.

References

[1] Yiran, Luo. “K-means Strategy Project” *CSE 575 - Statistical Machine Learning*, Ira A. Fulton Schools of Engineering, 13 January 2025.