Part 2: 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

K-Means clustering is a widely used unsupervised learning algorithm for partitioning data into distinct clusters. This project aims to analyze and visualize different cluster formations by varying the number of clusters *k* from 2 to 10. The objective is to understand the effect of different k-values on clustering and loss function behavior. This study applies K-Means clustering on a given dataset, evaluates cluster centroids, calculates the loss function for each k-value, and visualizes the clustering results.

II. Implementation

A. *Data Loading and Preprocessing*

a. The dataset 'AllSamples.npy' was loaded, containing two-dimensional points.
b. Initial centroids for each k-value (from 2 to 10) were generated using the 'initial_S2' function from 'precode.py', ensuring consistent initialization.

B. *Clustering Algorithm*

a. The K-Means clustering algorithm was applied using the computed initial centroids.
b. Cluster assignments were made by associating each data point with its closest centroid.
c. Centroids were recomputed iteratively until convergence was achieved.

C. *Loss Function Computation*

a. The loss function was computed for each k-value using the sum of squared Euclidean distances between data points and their respective centroids.
b. The computed loss values were printed in the required format.

D. *Visualization*

a. The clustering results were visualized using scatter plots for k-values ranging from 2 to 10.
b. A combined image displaying all cluster plots in a 3x3 grid format was generated.
c. The loss function values were plotted to analyze the elbow effect and determine the optimal k-value.

III. Results

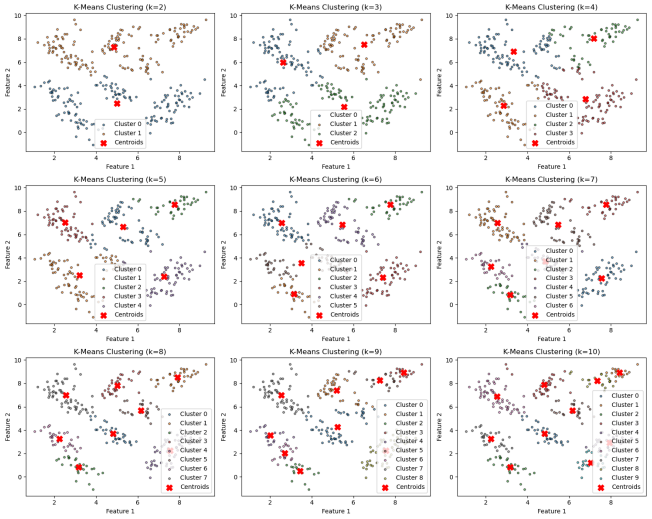

Figure 1: Cluster Plots

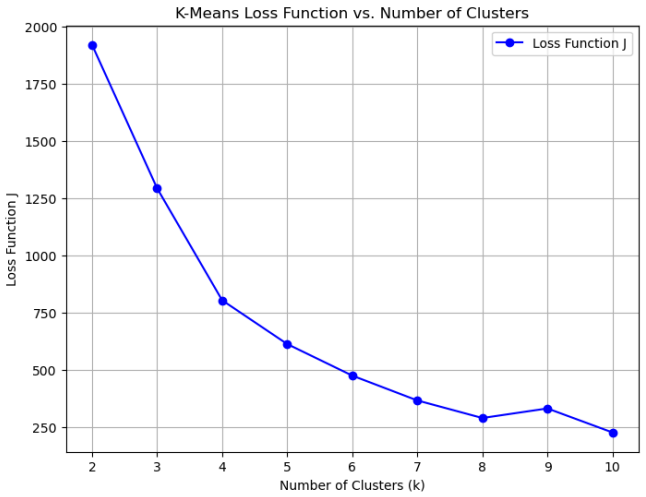

Figure 2: Loss Function Plot

The clustering results demonstrated how different k-values influenced the formation of clusters. As expected, the loss function consistently decreased with increasing k-values, indicating improved clustering performance. The elbow method was evident in the loss function plot, highlighting diminishing returns in loss reduction beyond a certain k-value. The centroids were appropriately repositioned throughout iterations to minimize intra-cluster variance, ensuring that the final clusters were well-separated and accurately represented the dataset structure.

IV. Observations

A. The loss function plot provided insight into the optimal number of clusters, revealing that after a certain point, increasing k-values yielded diminishing improvements in clustering quality.
B. The combined cluster plots made it easier to analyze clustering behavior across different k-values. The importance of proper centroid initialization was evident, as incorrect initialization led to suboptimal clustering results. Additionally, ensuring a consistent loss function computation was crucial in accurately determining the optimal k-value.

V. Challenges and Solutions

A. *Incorrect Centroid Initialization*
   a. Initial attempts resulted in incorrect centroid computations, leading to poor clustering performance.
   b. Solution: Used `initial_point_idx2` and `init_point` functions to generate proper initial centroids.
B. *Loss Function Discrepancies*
   a. The loss values varied due to inconsistent centroid initialization.
   b. Solution: Ensured fixed random seed usage and verified loss function computations.
C. *Combining Cluster Plots into a Single Image*
   a. Initially, separate plots were generated for each k-value, making comparisons difficult.
   b. Solution: Used Matplotlib's 'subplots()' function to create a 3x3 grid displaying all cluster plots in one image.

VI. Conclusion

This project successfully applied K-Means clustering to analyze how different k-values impact clustering results. The loss function plot provided insight into the optimal number of clusters using the elbow method. The final combined cluster plots allowed for a clear comparison of clustering performance across different k-values. The challenges faced during implementation were effectively resolved, ensuring accurate results and meaningful visualizations. This study highlights the importance of proper centroid initialization and loss function computation in achieving optimal clustering outcomes.

References

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