Introduction to Machine Learning and Artificial Neural Networks CS 554: HOMEWORK 2 REPORT

Amin Deldari Alamdari
S033174

I. INTRODUCTION

Clustering is a fundamental task in unsupervised machine learning, where the goal is to group similar data points without predefined labels. Among various clustering algorithms, k-*means* is one of the most popular due to its simplicity, efficiency, and effectiveness on low-dimensional datasets. In this homework assignment, we explore the behavior and performance of k-means clustering on a 2D dataset. We implement the algorithm from scratch and evaluate its reconstruction loss for varying numbers of clusters. Through repeated trials and visualizations, we aim to identify how cluster count affects performance and gain intuition about the data's natural groupings. The objective of this assignment was to implement the k-means clustering algorithm from scratch and analyze its performance on a 2D dataset. The clustering process was evaluated across different values of

II. METHODOLOGY

A. Data

The dataset `data.csv` contained 2D points and was loaded using pandas, as shown in Figure 1. The data was then processed using `numpy` and visualized using `matplotlib`.

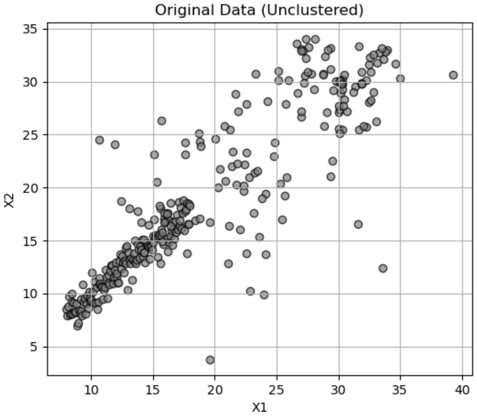

Fig. 1. Visualization of the original 2D dataset before clustering. Each point represents a data instance with no prior label or grouping. This plot serves as a baseline reference to compare with k-means clustering results.

B. k-Means Implementation

The k-means algorithm was implemented from scratch and included the following steps:

- Initialization: Randomly select k points as initial centroids.
- Assignment: Assign each point to the closest centroid using Euclidean distance.
- Update: Recalculate centroids as the mean of the points assigned to each cluster.
- Convergence Check: Repeat assignment and update steps until centroids stabilize or maximum iterations (100) are reached.

C. Reconstruction Loss

The **reconstruction loss** (also known as the within-cluster sum of squared errors) was calculated as:

$$Loss = \sum_{i=1}^{N} ||x_i - \mu_{c_i}||^2 \tag{1}$$

where x_i is a data point and $\mu_{c_i}$is the centroid of its assigned cluster.

D. Experimentation

For each value of k:

- 10 independent trials of k-means were conducted (random initialization).
- The mean reconstruction loss was computed over the 10 trials.
- The trial with the lowest reconstruction loss was saved for visualization.

III. RESULTS

A. Mean Reconstruction Loss Plot

The following plot shows the mean reconstruction loss versus number of clusters k:
As expected, the reconstruction loss decreases with increasing k, as more clusters allow better fitting to the data. The elbow in the curve (if present) can indicate a suitable value for k.

B. Best Clustering Visualizations

The plots below show the best clustering (lowest loss) for each value of k. Points are colored based on their assigned clusters, and centroids are marked with a black "**X**".

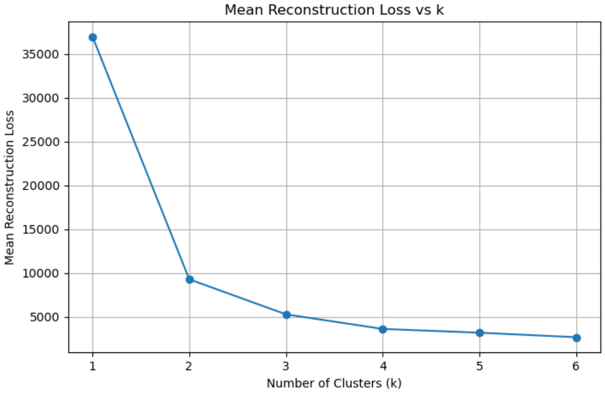

Fig. 2. Mean reconstruction loss for each value of k (number of clusters) across 10 k-means trials. The loss decreases as k increases, reflecting better data fitting. The ”elbow” in the curve can help indicate an optimal k value.

1) $k = 1$:

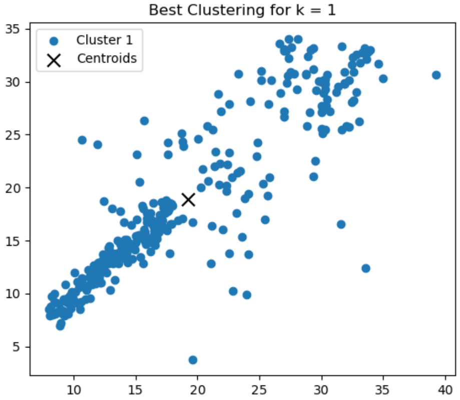

Fig. 3. Best k-means clustering result for $k = 1$. All data points are assigned to a single cluster, leading to the highest reconstruction loss.

2) $k = 2$:

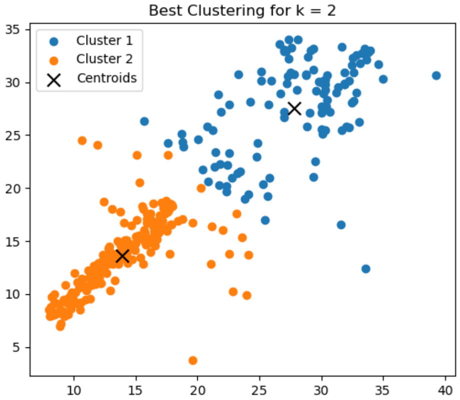

Fig. 4. Best k-means clustering result for $k = 2$. The data is split into two clusters, reducing reconstruction loss compared to $k = 1$.

3) $k = 3$:

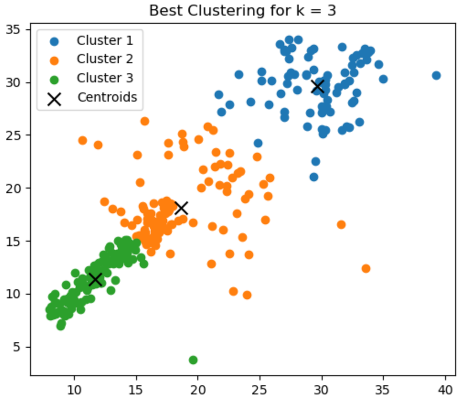

Fig. 5. Best k-means clustering result for $k = 3$. The clustering begins to reflect more distinct groupings within the data.

4) $k = 4$:

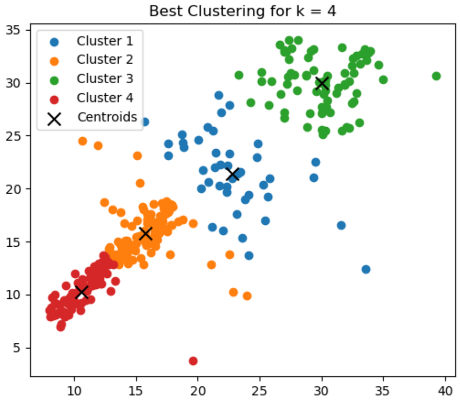

Fig. 6. Best k-means clustering result for $k = 4$. Finer separation is achieved, with reduced intra-cluster variance.

5) $k = 5$:

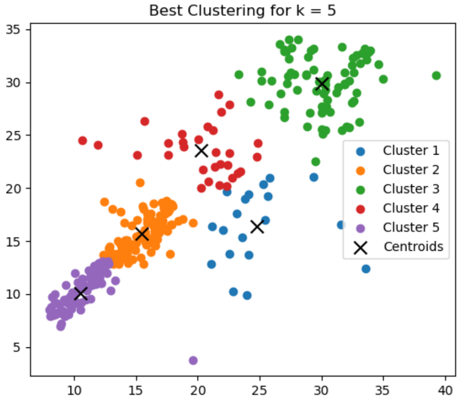

Fig. 7. Best k-means clustering result for $k = 5$. The data is further partitioned, and the centroids capture smaller groups.

6) k = 6:

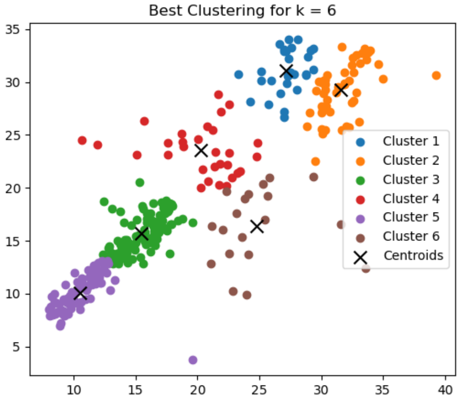

Fig. 8. Best k-means clustering result for k = 6. Each cluster represents a smaller, tighter group of points, yielding the lowest reconstruction loss among all tested k values.

IV. Discussion and Conclusion

- The reconstruction loss decreased consistently as k increased, which is a typical behavior for k-means clustering.
- From visual inspection and the loss curve, an ”elbow” may be visible around k = 3 or k = 4, suggesting these might be optimal cluster counts.
- The cluster shapes and separations appear well-defined for k = 3 and above, while lower k values result in more generalized groupings.
- Random initialization sometimes led to local minima; multiple trials ensured a fairer assessment.