CS554 K-Means Clustering for Unsupervised Learning Homework Report

Mert Erkol

March 26, 2024

1 Introduction

In this study multiple K-means models have been fitted by given multiple data and calculated errors. According to error plots, the best performing model is decided and explained why.

2 Background

The goal is to fit a K-Means model that minimizes the mean reconstruction error. To achieve this we tune the hyperparameter 'k' that decides the number of centroids in the k-means model. K-Means is an unsupersived learning method to optimize the given number of centroids location in the given data so that it can represent the data without providing any given label. The structure and the main algorithm can seen below [1]. We also calculate the mean reconstruction loss over 10 runs with different initial random cluster assignment.

Algorithm 1 K-means Algorithm

```
Input: Data points $X = \{x_1, x_2, \ldots, x_n\}$, number of clusters $k$
Output: Cluster centroids $C = \{c_1, c_2, \ldots, c_k\}$
Initialize centroids randomly: $c_1, c_2, \ldots, c_k \in X$
repeat
    for each data point $x_i \in X$ do
        Assign $x_i$ to the nearest centroid:
        $c_{\min} = \arg\min_{c_j} \|x_i - c_j\|$
        Update cluster assignment: $x_i \in C_{\min}$
    end for
    for each cluster $C_j$ do
        Update centroid $c_j$ as the mean of points in $C_j$:
        Calculate reconstruction loss of the cluster $C_j$
        $c_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$
    end for
until Centroids updated between iterations
```

The reconstruction loss can be described as:

$$\text{Reconstruction Loss}(k) = \sum_j \left\| X_{\text{cluster}(j)} - C_j \right\|^2$$

where:

- $X_{\text{cluster}(j)}$ represents the data points belonging to the j-th cluster.
- C_j represents the centroid of the j-th cluster.

The mean reconstruction loss for each k value over ten runs can be calculated using the formula:

$$\text{Mean Reconstruction Loss}(k) = \frac{1}{10} \sum_{i=1}^{10} \text{Reconstruction Loss}_i(k)$$

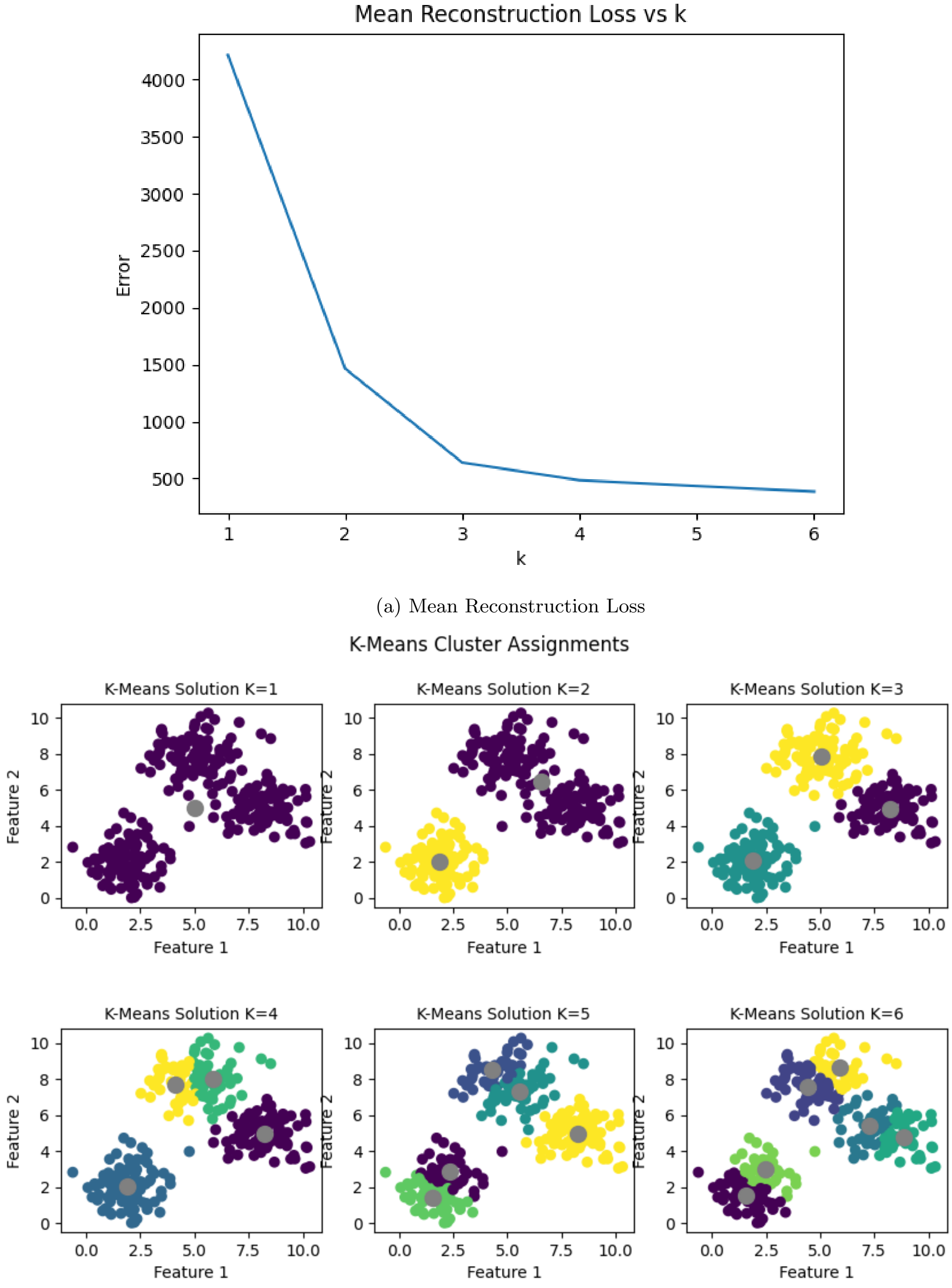
where:

- k is the number of clusters,
- $\text{Reconstruction Loss}_i(k)$ is the reconstruction loss for the i^{th} run with k clusters.

3 Methodology

To complete the project Python version 3.11 has been used as a programming language and two libraries: NumPy and Matplotlib. The structure is follows, Read the given 2 datasets, fit the data using the numpy implemented K-Means class with different k values, evaluate them with mean reconstruction loss over 10 iterations. At last the mean errors plotted as a function of k and best models over 10 visualized their cluster assignments. For MNIST cluster centroids visualized as gray scale images.

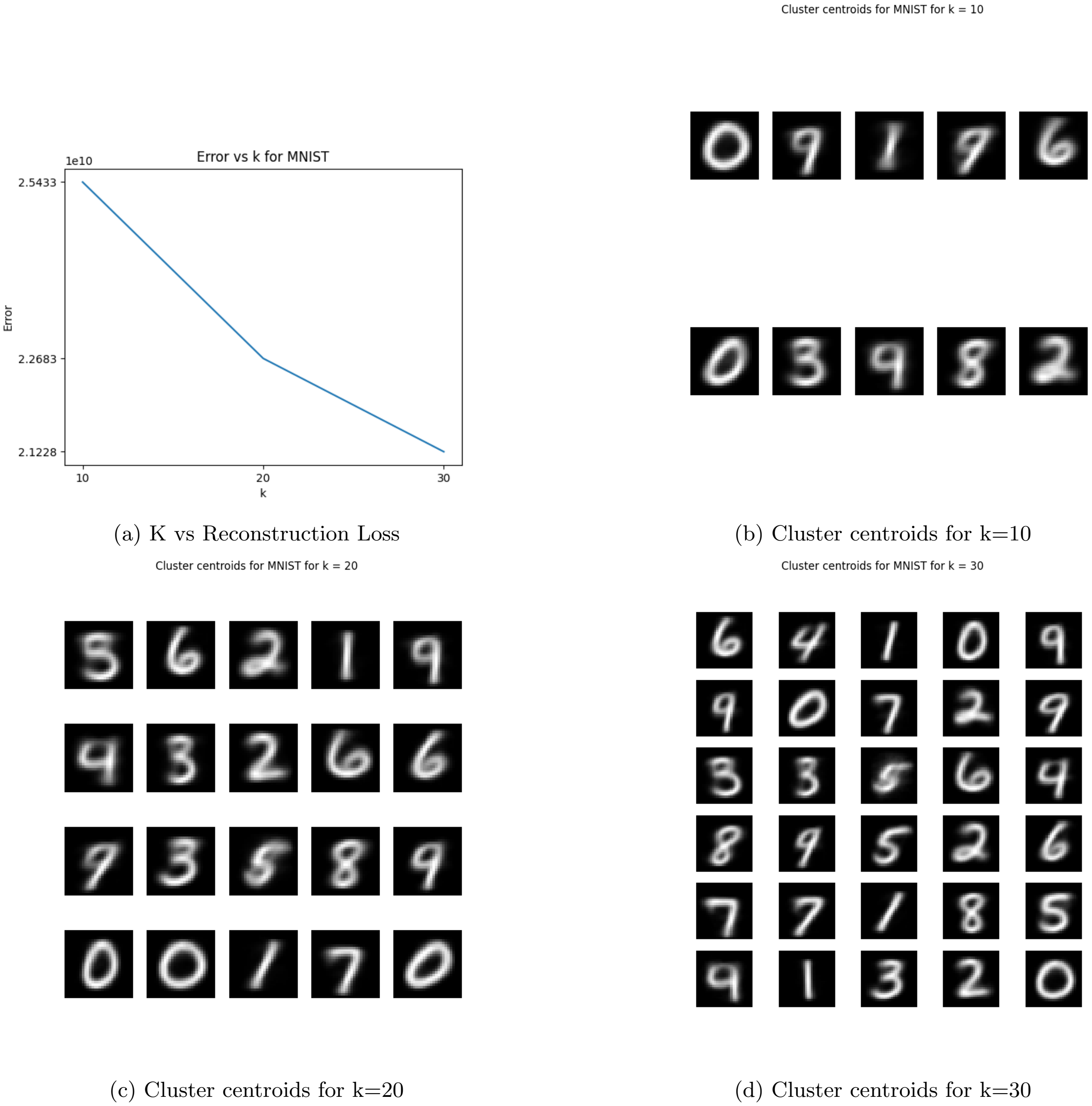
4 Results



(a) Mean Reconstruction Loss

(b) Clustering Assignments

Figure 1: data.csv results



(a) K vs Reconstruction Loss

(b) Cluster centroids for k=10

(c) Cluster centroids for k=20

(d) Cluster centroids for k=30

Figure 2: MNIST dataset results

5 Discussion

According to the figures 1 and 2 in the results section, For dataset 1 We have less dimensional dataset with only 2 features and when we observe the dataset in 2D it seems it can be distributed among 3 classes. When we fit our k-means according to the loss figure 1a our loss is decreasing as we increase the number of clusters but after cluster count 3 the decrease starts slowing down and converges. This method of selecting k is called ”Elbow Method” our goal is to find the elbow point of the plot which is $k = 3$. We may also select $k = 4$ but we gain very little performance also we increase the complexity of the model unnecessarily. For MNIST dataset results 2 reconstruction loss decreasing as we increase the number of clusters k but it does not seem that it converge enough we need more centroids to represent the dataset. Cluster centroid images 2 tells us even though we know there are 10 classes in the MNIST dataset $k = 10$ can not represent whole numbers because in the high dimensonal space like images there are different types of same numbers with different features so we need more clusters to represent whole of them. As we increase k we gain more performance in the mnist dataset.

6 Conclusion

Overall, the discussion highlights the importance of selecting an appropriate number of clusters based on the characteristics of the dataset, with considerations for both performance and model complexity.

References

[1] Wikipedia. *k-means clustering — Wikipedia, The Free Encyclopedia*. 2024. URL: `https://en.wikipedia.org/wiki/K-means_clustering`.