Analysis of Clustering Algorithms

Ivan Seslija

November 18, 2023

Abstract

In this report, we analyze the performance of different versions of Lloyd's algorithm and hierarchical agglomerative clustering across various datasets. Our objective is to determine the optimal number of clusters by examining cost metrics for Lloyd's algorithm and interpreting dendrograms for hierarchical clustering. This analysis aims to provide insights into the effectiveness of each method under different conditions and to recommend best practices for clustering large datasets.

1 Introduction

Clustering algorithms play a crucial role in data analysis, enabling the grouping of similar data points. This report delves into two key clustering techniques: Lloyd's algorithm, commonly linked with k-means clustering, and hierarchical agglomerative clustering (HAC). The former requires specifying the number of clusters, k, and iteratively optimizes cluster assignments. In contrast, HAC incrementally builds up clusters without a predefined cluster count, using a dendrogram to represent the process.

The aim of this study is to evaluate the performance of different implementations of these algorithms across various datasets. For Lloyd's algorithm, the focus is on how different values of k influence the clustering outcomes. For HAC, the emphasis is on understanding and analyzing the dendrogram to determine the most appropriate cluster formation. This comparative analysis seeks to elucidate the strengths and limitations of each method, particularly in the context of large datasets.

2 Methodology

This section outlines the methodologies used for evaluating Lloyd's algorithm and hierarchical agglomerative clustering (HAC) across two datasets.

2.1 Datasets Description

- **Dataset 1 (2D):** A two-dimensional dataset, used to assess algorithm performance in a simpler spatial context.
- **Dataset 2 (3D):** A three-dimensional dataset, providing a more complex scenario to test the adaptability and efficiency of the algorithms.

2.2 Lloyd's Algorithm

- **Cluster Initialization:** The algorithm was tested with two different initialization techniques:
  - Uniform Randomization: Clusters are initialized with randomly selected points from the dataset.
  - K-Means++: An advanced initialization technique that spreads out the initial cluster centroids.
- **Range of K:** Values of K from 2 to 10 were tested to determine the optimal number of clusters.
- **Repeated Runs:** For each value of K and each initialization method, the algorithm was executed 5 times. This repetition helps to minimize the effects of random initialization, and the results were averaged for more reliable metrics.
- **Application on Datasets:** The above steps were conducted for both Dataset 1 (2D) and Dataset 2 (3D), enabling a comparison of algorithm performance in different dimensional spaces.

2.3 Hierarchical Agglomerative Clustering (HAC)

- **Linkage Criteria:** Two approaches were used:
  - Single Linkage: Focusing on the minimum distance between clusters.
  - Average Linkage: Considering the average distance between all pairs in two clusters.
- **Dendrogram Construction:** For each linkage criterion, a dendrogram was built to visually represent the clustering process and to assist in determining the optimal clusters.
- **Application on Datasets:** Both linkage criteria were applied to Dataset 1 (2D) and Dataset 2 (3D), enabling an analysis across varied data structures.

This methodology provides a comprehensive approach to comparing the effectiveness of Lloyd's algorithm and HAC. By employing multiple runs and averaging results for Lloyd's algorithm, and exploring different linkage criteria for HAC, a thorough evaluation of these clustering techniques is achieved.

3 Experiments and Results

3.1 Lloyd's Algorithm

3.1.1 Experiment Setup

- The experiments were conducted using Lloyd's algorithm for cluster analysis.
- Two datasets were tested: a 2D points dataset (Dataset 1) and a 3D points dataset (Dataset 2).
- A range of values for k (number of clusters) were tested to determine the optimal clustering.
- Both datasets were tested with uniform random initial k and k-means++ initialization.

3.1.2 Results and Discussion

Dataset 1 (2D Points)

- The cost vs. k plot showed the classic elbow graph, indicating diminishing returns for the cost with increased k.
- Optimal k for Dataset 1 was found to be 5.
- No significant difference was observed between uniform random and k-means++ initialization methods.

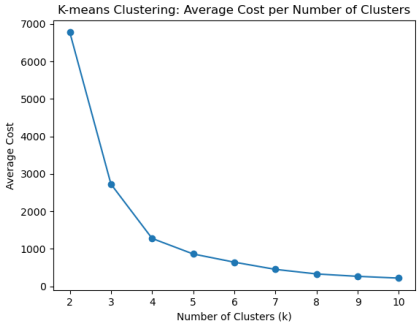

Figure 1: Cost vs. k for Dataset 1 with Uniform Initialization

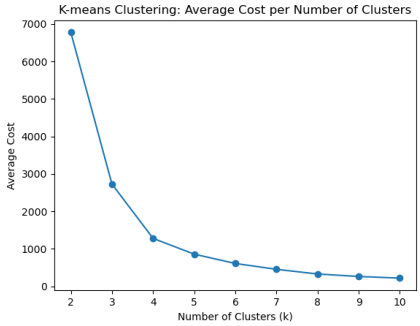

Figure 2: Cost vs. k for Dataset 1 with k-means++ Initialization

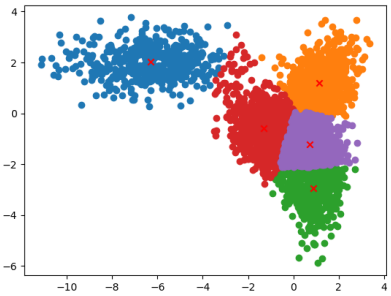

Figure 3: Dataset 1 with Uniform Initialization (K = 5)

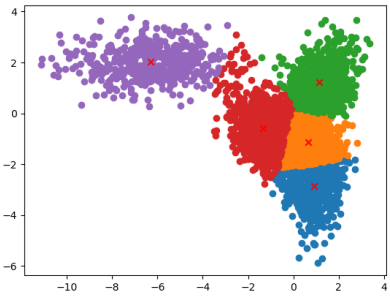

Figure 4: Dataset 1 with k-means++ Initialization (K = 5)

Dataset 2 (3D Points)

- A similar pattern was observed with Dataset 2, with the elbow graph indicating an optimal k.
- For Dataset 2, the optimal k was determined to be 6.
- As with Dataset 1, no significant difference was noted between the two initialization methods.

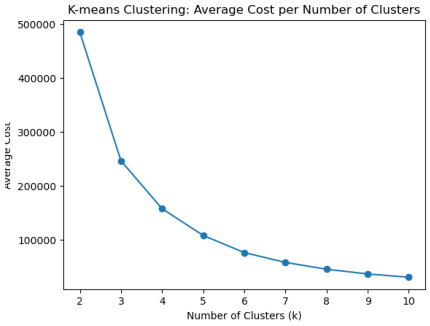

Figure 5: Cost vs. k for Dataset 2 with Uniform Initialization

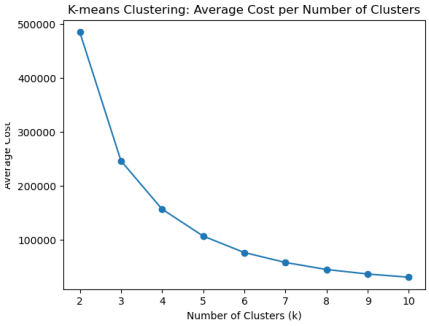

Figure 6: Cost vs. k for Dataset 2 with k-means++ Initialization

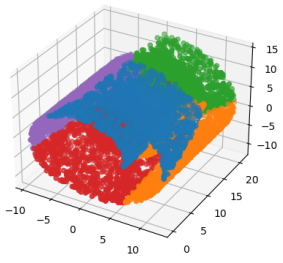

Figure 7: Dataset 2 with k-means Initialization (K = 6)

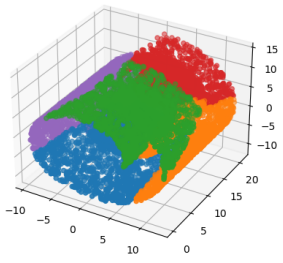

Figure 8: Dataset 2 with k-means++ Initialization (K = 6)

General Observations Our experiments with Lloyd's algorithm on two datasets revealed consistent findings: the method of initial cluster selection (uniform random versus k-means++) showed no significant impact on determining the optimal number of clusters K. This held true even when we moved from simpler 2D data to more complex 3D data, where the only notable change was a slight increase in the optimal K value. These results suggest that in such contexts, the complexity of the data, rather than the initial cluster selection method, plays a more pivotal role in influencing the optimal number of clusters.

3.2 Hierarchical Agglomerative Clustering

3.2.1 Experiment Setup

- The experiments employed Hierarchical Agglomerative Clustering to understand the inherent groupings in the datasets.
- Two linkage criteria were tested: 'average' and 'single' linkage, to assess their impact on cluster formation.
- Dendrogram analysis was utilized as a primary tool to visualize and interpret the clustering results, especially to decide where to cut the dendrogram to determine the number of clusters.

3.2.2 Results and Discussion

Dataset 1 (2D Points)

- The dendrogram analysis for Dataset 1 revealed a clear preference for the 'average' linkage method over 'single' linkage.
- With 'average' linkage, the dendrogram suggested cutting at a height of 3, resulting in four distinct groupings. This height was deemed optimal based on the dataset's characteristics.
- In contrast, 'single' linkage resulted in a highly imbalanced dendrogram, making it impractical to choose a meaningful height for cutting, as too many data points were clustered into a single group.

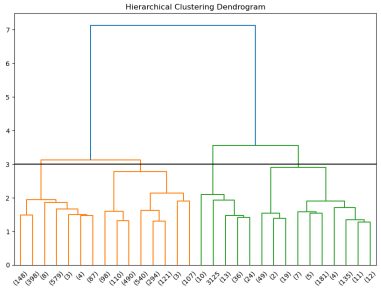

Figure 9: Dendogram for HAC (Average Linkage)

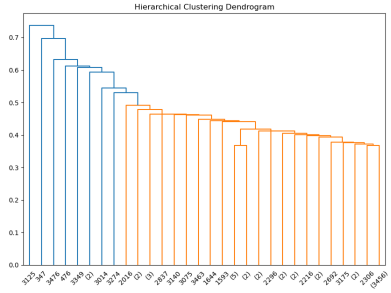

Figure 10: Dendogram for HAC (Single Linkage)

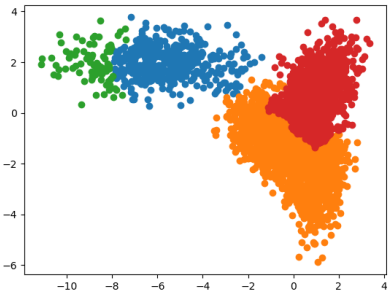

Figure 11: Clusters for HAC (Average Linkage) (K=4)

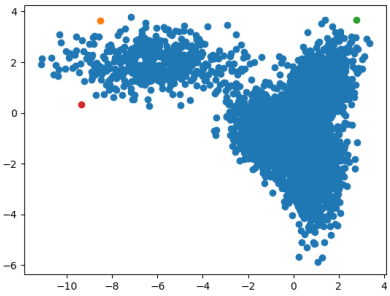

Figure 12: Clusters for HAC (Single Linkage) (K=4)

Dataset 2 (3D Points)

- A similar pattern was observed in Dataset 2, where 'average' linkage outperformed 'single' linkage in terms of producing more meaningful and balanced clusters.
- For 'average' linkage, a cut at a height of 12 was found to be optimal, resulting in a reasonable number of clusters.
- As with Dataset 1, the 'single' linkage dendrogram was less useful for cluster determination, leading to the decision that selecting a specific height for cutting was pointless due to the clustering of many points into a single group.

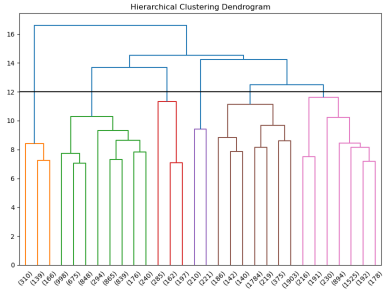

Figure 13: Dendogram for HAC (Average Linkage)

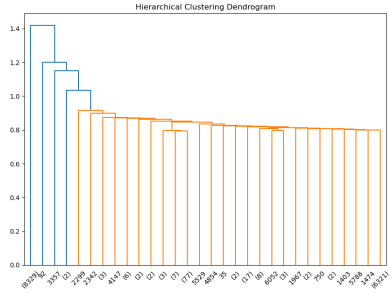

Figure 14: Dendogram for HAC (Single Linkage)

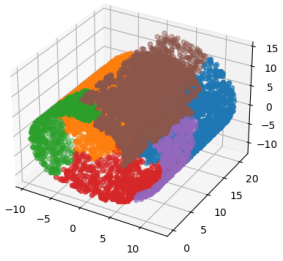

Figure 15: Clusters for HAC (Average Linkage) (K=6)

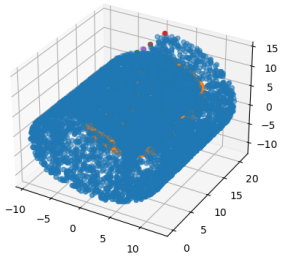

Figure 16: Clusters for HAC (Single Linkage) (K=6)

General Observations These results highlight the effectiveness of 'average' linkage in Hierarchical Agglomerative Clustering for the datasets tested, presenting a more balanced and interpretable division of clusters than what is observed with 'single' linkage. Importantly, the decision of where to cut the dendrogram is key in determining the number of clusters. Our analysis indicates that 'average' linkage offers a more intuitive and practical approach for making this critical decision.

4 Comparative Analysis

4.1 Performance Comparison

Both Lloyd's algorithm and Hierarchical Agglomerative Clustering (HAC) showed similar effectiveness in their respective optimal settings. Despite their methodological differences, they performed comparably across the datasets.

4.2 Determining the Number of Clusters

Lloyd's algorithm provided a straightforward approach to incrementally adjust the number of clusters, with an optimal K being 5 or 6. In contrast, HAC involved a more user-based decision-making process in determining where to cut the dendrogram leading to K being 4 or 6 depending on the dataset. This complexity, however, offered deeper insights into the similarities and relationships between clusters.

4.3 Insights on Cluster Similarity

HAC's dendrogram-based approach allowed for a detailed understanding of cluster structures and similarities, an aspect less emphasized in Lloyd's algorithm, which focuses more on optimal data partitioning.

4.4 Cluster Differences for Same K in Lloyd's Algorithm and HAC

When Lloyd's algorithm (K-means) and Hierarchical Agglomerative Clustering (HAC) are used to create the same number of clusters (K), their results differ notably:

Lloyd's Algorithm: Tends to produce evenly distributed, intuitive clusters. It's effective in datasets with clear separations, but can be sensitive to outliers.

HAC: Generates clusters that can vary greatly in size and shape, capturing more nuanced data relationships. These clusters are often less intuitive but can be more insightful, especially in datasets with complex structures.

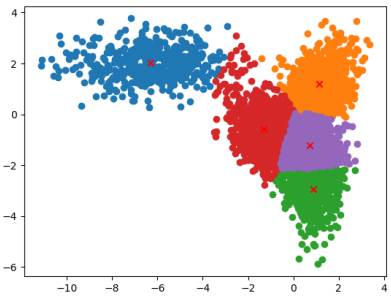

Figure 17: K-Means clusters (K=5)

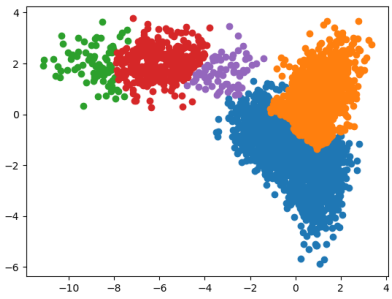

Figure 18: HAC clusters (K=5)

5 Conclusion

5.1 Summary of Key Findings

Our study demonstrated that both Lloyd's algorithm and Hierarchical Agglomerative Clustering (HAC) are effective in segregating data when configured with their optimal settings. Each method, with its unique approach, provided valuable insights into the clustering process and proved equally adept at separating the datasets under investigation.

5.2 Reflections on the Efficacy of Different Clustering Methods

The comparison of these two methods highlighted their respective strengths. HAC, with its dendrogram-based analysis, excels in identifying natural groupings without a predetermined number of clusters. On the other hand, Lloyd's algorithm, with its straightforward approach to incrementing clusters, offers a clear pathway to fine-tune the number of clusters.

5.3 Suggestions for Future Work

A notable realization from our analysis is the potential synergy between these two methods. Future work could explore using HAC to initially identify a range of potential values for kk (the number of clusters), which can then be

further refined using Lloyd's algorithm. This integrated approach could harness the strengths of both methods, leveraging the intuitive groupings from HAC and the precision of Lloyd's algorithm to achieve even more accurate clustering results.