K-means Clustering and Input Variables

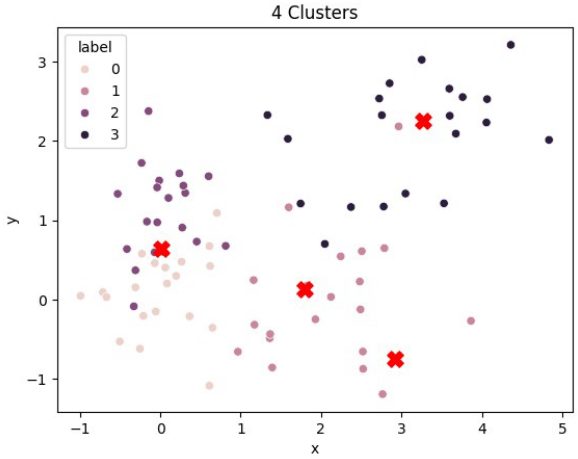
Introduction

This lab focused on implementing the K-means clustering algorithm and applying it to the Olivetti faces dataset and pseudo-randomly generated data. K-means is an unsupervised learning technique used to partition data into clusters based on similarity. The algorithm iteratively assigns data points to the nearest centroid and updates the centroids until convergence. The goal of this lab was to test the K-means algorithm with different input variables and understand the results.

Procedure

Part I: Implementing K-means

The backbone of this experiment is the K-means algorithm which is a somewhat complex algorithm that uses multiple functions, calculations, and code that is designed to group small dimensional data. The graph below shows example points as well as the centroids our algorithm has arrived at.



Part II: Clustering the Olivetti Dataset and testing the algorithm with updated input parameters

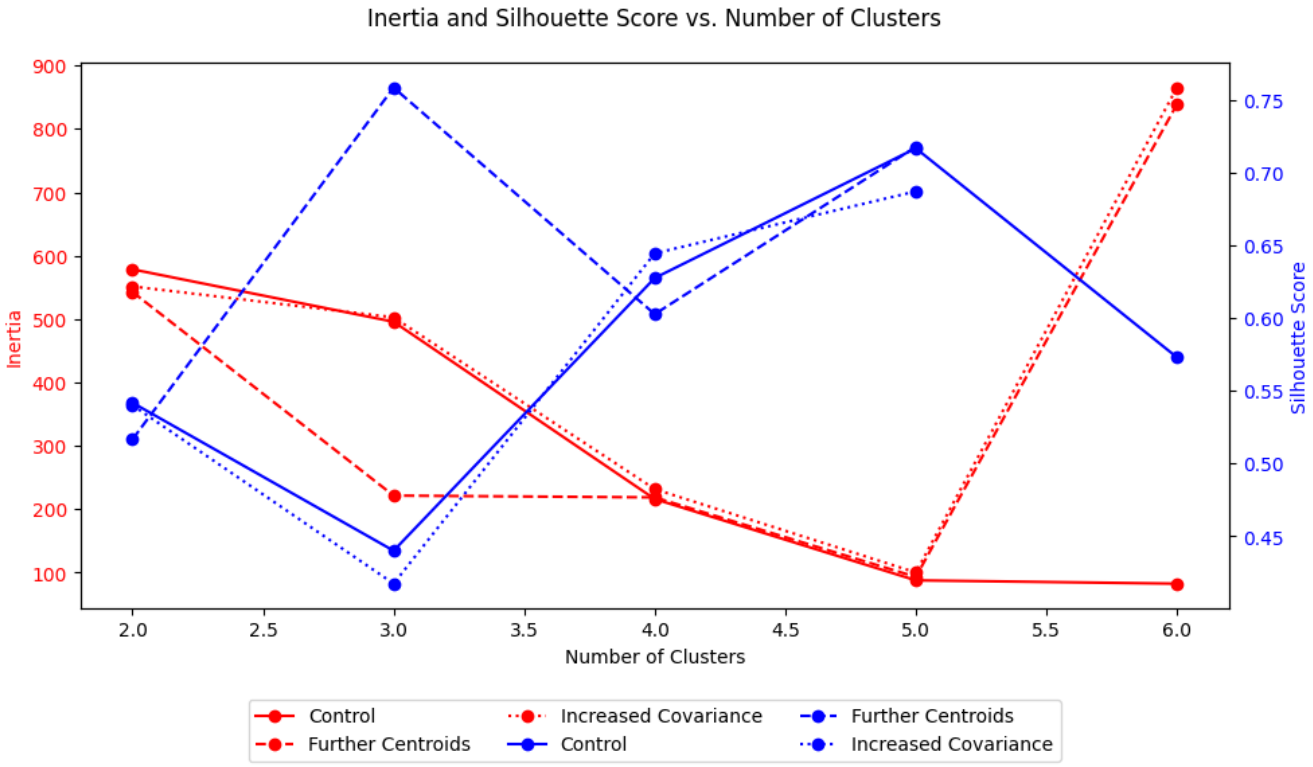
1. **Data Generation:** Pseudo-random data with four clusters is generated for each test
2. **Clustering:** For each test scenario, our k-means code is called as “get_clusters”
3. **Performance Evaluation:** Inertia and silhouette score were used to evaluate the clustering performance across different numbers of clusters and data distributions. These metrics provide insights into the compactness and separation of clusters. The values were stored and plotted for comparison.

Results

Pseudo-Random Data Clustering

The clustering results obtained from applying our K-means implementation to pseudo-random data were inconclusive regarding the impact of cluster centers and covariance matrices on cluster formation. The randomness inherent in the selection of initial centroids appeared to be a more significant factor in determining the final clusters compared to the tested data parameters.

While modifying the cluster centers to increase separation between clusters and changing the covariance matrices to alter the spread of clusters did yield some variations in the inertia and silhouette score, it was difficult to isolate the specific influence of these data parameters. The inherent variability caused by random initialization made it challenging to conclusively attribute these changes to the intended data modifications.



Discussion

Contrary to our initial expectations, the data collected does not strongly support the hypothesis that initial cluster centers and covariance significantly impact clustering outcomes. We can attribute our unimpressive results to testing methodology, our data points should have been held as a control, and our input variables should have been changed incrementally to graph an accurate and gradual correlation. This highlights a limitation in the current experimental setup, where the influence of random centroid selection overshadows the impact of the data parameters we aimed to analyze, and the resulting visual is never the same. To iteratively change the input variables (cluster centers and

covariance) and systematically observe their effect on cluster formation would indeed likely present more informative results. This controlled approach would minimize the confounding effects of randomness and potentially reveal subtle yet meaningful correlations between data parameters and clustering outcomes.

Conclusion

While this lab has helped us implement the K-means algorithm and understand its sensitivity to various parameters, the limitations in our experimental design hindered our ability to conclusively determine the influence of initial cluster centers and covariance on clustering results. Random initial centroid selection introduced a significant source of variability that obscured the relationships we were attempting to study.