Report on K-means Clustering and Principal Component Analysis of the Iris Dataset

Youri HALMAERT

June 12, 2025

1 Introduction

This report details an exploratory data analysis conducted on the Iris dataset, focusing on two key unsupervised learning techniques: K-means clustering and Principal Component Analysis (PCA). The primary objective was to understand how these methods can be applied to discover inherent structures within the data, particularly when true labels are unknown, and to evaluate the effectiveness of combining PCA with K-means for improved clustering.

2 Part 1: K-means Clustering

The first part of the analysis focused on applying the K-means algorithm to the Iris dataset to identify optimal clustering, assuming the true labels were unknown. Two common methods for determining the optimal number of clusters (K) were employed: the Elbow method and Silhouette analysis.

2.1 Methodology

- **Data Loading:** The Iris dataset was loaded using `sklearn.datasets.load_iris`.
- **K-means Application:** The `KMeans` algorithm from `sklearn.cluster` was used to perform clustering for various values of K.
- **Elbow Method:** This method involves plotting the Within-Cluster Sum of Squares (WCSS) against different values of K. The "elbow" point, where the rate of decrease in WCSS significantly slows down, suggests an optimal K.
- **Silhouette Analysis:** This technique measures how similar an object is to its own cluster compared to other clusters. A silhouette score close to +1 indicates that the object is well-matched to its own cluster and poorly matched to neighboring clusters, while a score near -1 indicates the opposite. The average silhouette score for various K values helps identify the best clustering.

2.2 Results and Discussion

The Elbow method typically suggested K=2 or K=3 as potential optimal values, with K=2 showing the sharpest reduction in WCSS. This was consistent with the Silhouette analysis, which also indicated K=2 as yielding the best-defined clustering structure.

Comparing these findings with the true labels of the Iris dataset (which has three distinct species, implying K=3), it was observed that K-means partially aligns with the true labels. While the Iris dataset inherently has three classes, the clustering metrics (Elbow and Silhouette) leaned towards K=2. This discrepancy highlights a known characteristic of the Iris dataset: the Iris-versicolor and Iris-virginica species exhibit some overlap in their feature space, making a clear separation into three distinct clusters challenging based purely on feature proximity. Therefore, K=2 better reflected the underlying separable structure based on the clustering performance metrics.

3 Part 2: Principal Component Analysis (PCA) and K-means

The second part of the analysis explored the combination of PCA with K-means, aiming to see if dimensionality reduction could lead to a better clustering outcome for the Iris dataset.

3.1 Methodology

- **Data Loading:** The Iris dataset was reloaded.
- **PCA Implementation:** A custom PCA algorithm was implemented based on the covariance method. This involved:
    1. Centering the data by subtracting the mean of each feature.
    2. Calculating the covariance matrix of the centered data.
    3. Computing eigenvalues and eigenvectors of the covariance matrix.
    4. Sorting eigenvalues in descending order and arranging eigenvectors accordingly.
    5. Selecting the top `n_components` eigenvectors (principal components).
    6. Transforming the centered data into the new principal component space by dot product with the selected eigenvectors.
- **PCA Application:** PCA was applied to the Iris dataset, reducing its dimensionality to the top 2 principal components. The resulting transformed data had a shape of (150, 2).
- **K-means on PCA-transformed Data:** The `KMeans` algorithm (with K=3) was then applied to the 2-component PCA-transformed data to obtain cluster labels.

- **Visualization:** The K-means clustering results in the latent space were compared visually with the true labels.

3.2 Results and Discussion

When K-means (with K=3) was applied to the PCA-transformed data, the clustering revealed distinct separations among the three clusters in the reduced latent space. However, some overlaps between clusters were still observed.

A comparison with the true labels showed that one cluster (often visually represented as purple in plots) was well-matched to its corresponding true species. In contrast, the clustering struggled to perfectly align with the other two true species (often yellow and green in plots), which exhibited more inter-class overlap. This suggests that while K-means effectively captures the global structure present in the data, it still faces challenges in distinguishing between classes that are not clearly separable in the feature space, even after dimensionality reduction.

The use of the latent space (obtained via PCA) successfully simplified the data for clustering purposes. PCA effectively reduced the noise and redundancy in the original features, allowing K-means to operate on a more compact representation. However, the inherent ambiguity or overlap between certain Iris species in their feature distributions persisted even in the projected space, limiting the ability of K-means to achieve a perfect separation that fully matches the true labels for all classes.

4 Conclusion

This lab demonstrated the application of K-means clustering and Principal Component Analysis on the Iris dataset. The K-means analysis, utilizing the Elbow method and Silhouette analysis, indicated that K=2 provided a more optimal clustering based on internal validation metrics, despite the dataset having three true classes. This highlighted the inherent overlap between two of the Iris species.

The subsequent combination of PCA with K-means showed that dimensionality reduction can effectively simplify the data, making it more amenable to clustering. While PCA helped in visualizing and processing the data in a lower-dimensional space, the K-means algorithm still reflected the challenges in achieving a perfect separation for all three true classes due to existing feature overlaps. Overall, the analysis provided valuable insights into the strengths and limitations of K-means and PCA for uncovering data structures in an unsupervised learning context.