TP: Application of K-Means Algorithm

Data Mining & Machine Learning

Authors: Brahimi Abdelhak Fares **Date:** November 29, 2025

1 Introduction

The objective of this practical work is to apply the **K-Means** algorithm to group similar individuals into homogeneous clusters. For this study, we selected the domain of **retail marketing**. The goal is to segment customers of a store based on their purchasing behavior to improve marketing strategies.

2 1. Context and Variables

We utilized the "Mall Customers" dataset, which reflects real-world retail distributions. The key variables selected for the clustering process are:

- **Annual Income:** The approximate yearly income of the customer (in thousands of dollars).
- **Spending Score:** A score assigned by the mall (1-100) based on customer behavior and purchasing data.

These two variables allow us to visualize the clusters effectively in a 2D space.

3 2. Data Preparation and Normalization

Before applying the algorithm, preprocessing was required. Since K-Means utilizes Euclidean distance to determine cluster membership, variables with different scales can distort the results.

- We standardized the data using the *StandardScaler* method (Z-score normalization).
- This transformed the data such that the mean is 0 and the standard deviation is 1 for both variables, ensuring equal weight during clustering.

4 3. Determining Optimal Clusters (Elbow Method)

We applied the K-Means algorithm for values of k ranging from 1 to 10. For each k, we calculated the WCSS (Within-Cluster Sum of Squares).

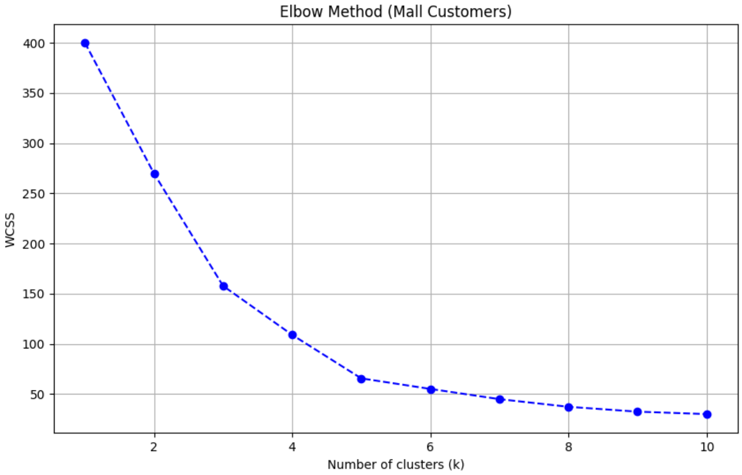

Figure 1: Elbow Method Graph

As observed in Figure 1, the curve forms a distinct "elbow" at $k = 5$. Before this point, the WCSS decreases rapidly; after this point, the decrease becomes linear and marginal. Therefore, we selected $k = 5$ as the optimal number of clusters.

5 4. Clustering Results and Visualization

We trained the K-Means model with $k = 5$. The resulting clusters are visualized below in Figure 2. The yellow stars represent the centroids of each group.

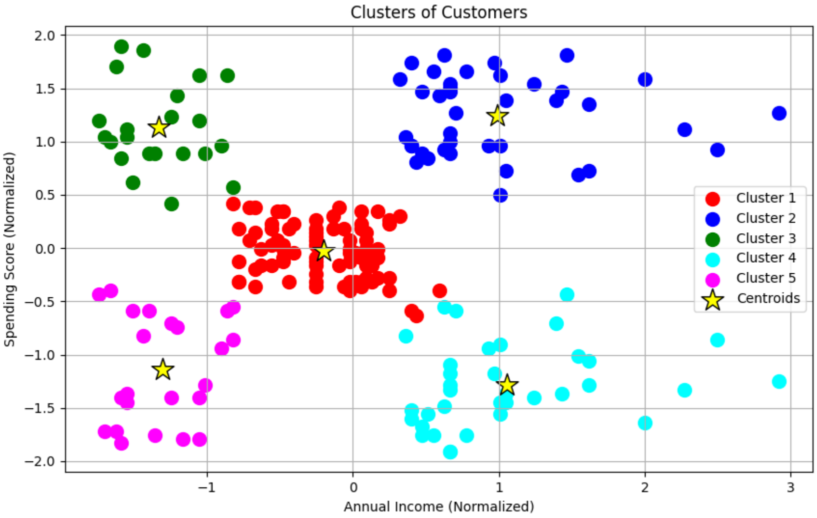

Figure 2: Customer Segments (2D Visualization)

6 5. Interpretation of Profiles

Based on the cluster centers and the visualization, we identified 5 distinct customer profiles. These insights allow for targeted marketing strategies:

1. **Cluster 1 (Red - Standard Customers):** Average income and average spending scores. They represent the middle class of customers.
2. **Cluster 2 (Blue - Target Customers):** High annual income and high spending score. These are the most valuable customers for the store.
3. **Cluster 3 (Green - Impulse Buyers):** Low income but high spending score. These customers spend lavishly despite lower earnings.
4. **Cluster 4 (Cyan - Careful Spenders):** High annual income but low spending score. They have high purchasing power but prefer to save rather than spend.
5. **Cluster 5 (Magenta - Sensible/Economical):** Low income and low spending score. They spend wisely, focusing only on their specific needs.

7 Conclusion

The application of the K-Means algorithm successfully segmented the customer base into interpretable groups. This segmentation is a powerful tool for business intelligence; it allows store management to move away from a ”one size fits all” strategy and instead offer personalized promotions (e.g., discounts for the ”Sensible” group vs. VIP events for the ”Target” group) to maximize revenue and customer loyalty.