Mall Customer Segmentation: A Comparative Study of K-Means Initialization Methods

CHRIST SAGOMBAYE

Master in Machine Intelligence, African Institute for Mathematical Sciences

Abstract

This report explores the performance of three K-Means initialization methods:Random, K-Means++, and Maximin on the Mall Customer Segmentation dataset. By analyzing convergence speed, cluster compactness (inertia), and separation (Silhouette Score), we determine the optimal number of clusters ($k = 5$) using the Elbow Method. Mathematical formulations of the algorithms and practical insights into their trade-offs are discussed.

1 Introduction & Context

Customer segmentation is critical for personalized marketing strategies. This project applies K-Means clustering to the Mall Customer Segmentation dataset, focusing on two features: *Annual Income (k$)* and *Spending Score (1–100)*. The goal is to compare initialization methods and validate the optimal number of clusters mathematically.

2 Methodology

2 Data Preparation

The dataset contains 200 customer records with features: Customer ID, Age, Gender, Annual Income, and Spending Score. After extracting *Income* and *Spending Score*, we standardized both variables using Z-score normalization:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

This ensures equal contribution of both features to the Euclidean distance metric used in K-Means.

2 K-Means Algorithm

K-Means minimizes the Within-Cluster Sum of Squares (WCSS):

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

where μ_k is the centroid of cluster k. The algorithm iterates between two steps until convergence:

(a) **Assignment**: Assign each point to the nearest centroid.

(b) **Update**: Recompute centroids as the mean of assigned points.

2 Initialization Methods

2.3.1 Random Initialization

Centroids are chosen uniformly at random from the data points. While computationally efficient, this method risks suboptimal convergence due to poor initial centroid placement.

2.3.2 K-Means++

This method selects centroids sequentially, each new centroid being far from existing ones. The probability of selecting a new centroid x is proportional to the squared distance to the nearest existing centroid:

$$P(x) \propto D(x)^2$$

This reduces the risk of poor cluster formation compared to random initialization.

2.3.3 Maximin Initialization

Maximin selects centroids by maximizing the minimum distance between them. The first centroid is chosen randomly, and each subsequent centroid is the point farthest from all existing centroids. This balances compactness and separation.

3 Results & Analysis

3 Performance Comparison

Method	Time (s)	Iterations	Inertia	Silhouette
Random	0.0132	10	44,454.48	0.554
K-Means++	0.0248	5	44,448.46	0.562
Maximin	0.0185	7	**44,320.12**	**0.580**

Table 1: Performance comparison of initialization methods

Key Observations:

- **Random Initialization:** Fastest but requires more iterations and yields the highest inertia.

- **K-Means++:** Halves iterations but sacrifices slight compactness compared to Maximin.

- **Maximin:** Balances speed and quality, achieving the lowest inertia and highest Silhouette Score.

3 Mathematical Insights

- **Inertia:** Measures cluster compactness. Lower values indicate tighter clusters.
- **Silhouette Score:** Evaluates cohesion ($a(i)$) and separation ($b(i)$) per point:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

A score of 0.554 (for $k = 5$) suggests well-defined clusters.

4 Choosing the Optimal k

4 Elbow Method

Plotting inertia vs. k reveals diminishing returns at $k = 5$ (Figure 1). Beyond this, marginal gains in inertia reduction are minimal.

4 Silhouette Validation

The Silhouette Score peaks at $k = 5$, confirming optimal cluster separation.

Why $k = 5$?

- **Business Logic:** Aligns with intuitive segments (e.g., VIP customers, budget-conscious shoppers).
- **Mathematical Justification:** Balances model complexity and interpretability.

5 Interpretation & Actionable Segments

Visualizing clusters (Figure 2) reveals distinct customer profiles:

Cluster	Income	Spending	Strategy
1	Low	Low	Target with discounts
2	Low	High	Loyalty programs
3	Medium	Medium	Standard promotions
4	High	Low	Premium product offers
5	High	High	Exclusive VIP experiences

Table 2: Customer segments and business strategies

6 Mathematical Foundations of Clustering

6 Optimization Landscape

K-Means solves a non-convex optimization problem. The objective function J is minimized via alternating optimization:

(a) **Assignment:** NP-hard combinatorial problem (solved greedily).

(b) **Update:** Convex subproblem (closed-form solution).

6 Dimensionality and Distance Metrics

By focusing on 2D features (*Income* and *Spending Score*), we avoid the "curse of dimensionality." However, higher-dimensional data may benefit from PCA preprocessing:

$$\text{PCA}(X) = \arg\max_{W} \text{Var}(XW)$$

This projects data onto orthogonal axes capturing maximal variance.

7 Algorithmic Complexity

- **Time Complexity:** $O(n \cdot k \cdot d \cdot t)$, where n = samples, d = dimensions, t = iterations.
- **Space Complexity:** $O(k + n)$ for centroids and labels.

8 Limitations and Future Work

8 Limitations

- **Bias from 2D Focus:** Excluding *Age* or *Gender* may oversimplify customer behavior.
- **Fixed k:** The Elbow Method's "knee" is subjective without rigorous criteria like Gap Statistics.

8 Future Directions

- **DBSCAN:** For non-convex clusters (e.g., outliers in high-income brackets).
- **Hierarchical Clustering:** To explore nested customer segments.
- **MiniBatch K-Means:** For scalability on larger datasets.

9 Conclusion

This project demonstrates that thoughtful initialization (e.g., Maximin) improves K-Means performance, reducing iterations and enhancing cluster quality. The choice of $k = 5$ balances mathematical rigor and business relevance.

References

References

[1] Scikit-learn documentation. (2023). *Clustering.* https://scikit-learn.org/stable/modules/clustering.html

[2] Ravi Kanth, K. V., et al. (1998). A Fast Algorithm for Clustering Large Datasets. *Proceedings of the 1998 International Conference on Knowledge Discovery and Data Mining (KDD).*

[3] Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1027–1035).

[4] Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). Pearson.

[5] Jain, A. K. (2010). Data clustering: 50 years beyond K-Means. *Pattern Recognition Letters*, 31(8), 651–666.

[6] Mint, R. (n.d.). *Algorithme K-Means*. https://mrmint.fr/algorithme-k-means

[7] Data-Transition Numérique. (n.d.). *K-Means Clustering Explained*. https://www.data-transitionnumerique.com/k-means/