Applying K-means Clustering for Mall Customer Segmentation

Aradhya Shukla

May 5, 2024

Abstract

Customer segmentation is crucial for targeted marketing strategies, especially in retail environments like malls. This dissertation explores the application of K-means clustering for segmenting mall customers based on their demographic attributes and spending behavior. The study utilizes the elbow method to determine the optimal number of clusters and employs K-means clustering to categorize customers into distinct segments. The findings provide valuable insights for mall management in devising personalized marketing initiatives.

1 Introduction

Customer segmentation is a fundamental strategy in retail marketing, enabling businesses to tailor their approaches to different customer groups. K-means clustering, an unsupervised learning algorithm, offers a robust method for segmenting customers based on shared characteristics. This dissertation investigates the effectiveness of K-means clustering in segmenting mall customers using a dataset comprising demographic attributes and spending behavior.

2 Methodology

The methodology involves several steps:

1. Data Preprocessing: The dataset containing customer attributes is cleaned and standardized.
2. Determining Optimal Cluster Number: The elbow method is used to identify the optimal number of clusters.
3. K-means Clustering: The K-means algorithm is applied to segment customers into distinct clusters.

3 Elbow Method

The elbow method is a heuristic used to determine the optimal number of clusters (k) in a dataset for K-means clustering. It works by plotting the within-cluster sum of squares (WCSS) against the number of clusters. The WCSS measures the compactness of the clusters; smaller values indicate tighter clusters.

When plotted, the graph typically resembles an arm, and the "elbow" point represents the optimal number of clusters. This point is where the rate of decrease in WCSS starts to slow down significantly. Choosing the number of clusters at this point balances the trade-off between maximizing the number of clusters to capture intricate patterns and minimizing the number of clusters to prevent overfitting.

4 K-means Clustering

K-means clustering is an unsupervised machine learning algorithm used for partitioning a dataset into K clusters. It operates by iteratively assigning each data point to the nearest cluster centroid and then updating the centroids based on the mean of the points assigned to each cluster. This process continues until the centroids stabilize or a specified number of iterations is reached.

The algorithm aims to minimize the within-cluster sum of squares, effectively partitioning the data into clusters that are compact and well-separated. It is computationally efficient and widely used for clustering tasks, especially when the number of clusters is known or can be estimated.

5 Mall Customer Segmentation

Mall customer segmentation refers to the process of dividing the mall's customers into distinct groups based on their characteristics, behaviors, and preferences. This segmentation allows marketers and mall management to better understand their customer base, tailor marketing strategies, and improve customer experience.

Segmentation can be achieved through various techniques, including demographic segmentation (age, gender, income), psychographic segmentation (lifestyle, values, interests), and behavioral segmentation (purchase history, frequency of visits). Machine learning algorithms, such as K-means clustering, can also be employed to automatically identify meaningful segments within the data.

6 Selection of K-means for Mall Customer Segmentation

K-means clustering is often chosen for mall customer segmentation due to its simplicity, efficiency, and interpretability. Here are some reasons why K-means might be preferred over other algorithms:

- **Ease of Interpretation:** K-means produces clusters with clear boundaries, making interpreting and understanding the resulting segments easy.
- **Scalability:** K-means is computationally efficient and can handle large datasets, which is advantageous for analyzing extensive customer data typically found in malls.
- **Assumption of Spherical Clusters:** K-means assumes that clusters are spherical and isotropic, which may be suitable for certain types of data, such as customer demographics or purchase behavior.
- **Suitability for Numeric Data:** K-means works well with numeric data, which is common in customer segmentation tasks involving attributes like age, income, and spending habits.

However, it's essential to consider the dataset's specific characteristics and the segmentation task's goals when selecting an algorithm. Other clustering algorithms, such as hierarchical clustering or DBSCAN, may be more appropriate depending on the nature of the data and the desired segmentation outcome.

7 K-means Steps

K-means clustering involves the following steps:

Step 1: Select Initial Cluster Centers

- Select k data points as the initial cluster centers (randomly chosen).

Step 2: Calculate Euclidean Distance

- Find the Euclidean distance of each data point towards each cluster center.

Step 3: Assign Data Points to Nearest Cluster

- Assign each data point to the nearest cluster based on the calculated distances.

Step 4: Update Cluster Centers

- Recompute the new cluster centers by taking the mean of the data points belonging to that cluster.

Step 5: Repeat Steps 2 to 4

- Repeat steps 2 to 4 until convergence is achieved.

Step 6: Stop the Process

- Stop the process when zero convergence is reached.

End Result

- The data points are clustered into k clusters.

8 Python Libraries Used

The following Python libraries were utilized in the analysis:

- **Numpy**: Used for numerical operations and array manipulation.
- **pandas**: Utilized for data manipulation and analysis.
- **Seaborn**: Employed for data visualization and statistical graphics.
- **Matplotlib**: Used for creating plots and visualizations.

9 Dataset Description

The dataset used in this study is titled *Customer Segmentation of Mall datasets with K-means*, provided by Anup Mondal on kaggle.com. It contains information on mall customers, including their gender, age, annual income, and spending score.

9.1 Parameters Used

The parameters utilized for customer segmentation are:

- Gender
- Age
- Annual Income
- Spending Score

9.2 Cluster Descriptions based on Priority Order

The dataset is segmented into the following clusters, each with its respective meaning and priority:

- **Cluster 2**: Annual Income is high and spending is also more, potential customers (Priority 1)

- **Cluster 0**: Income is less but spending is more, careless people (Priority 2)
- **Cluster 5**: Mediocre customers (Priority 3)
- **Cluster 4**: Earns more but spends less (Priority 4)
- **Cluster 3**: Earning low, spending also low (Priority 5)

9.3 Cluster Labels for Customer Segmentation

Based on the cluster descriptions, the following labels are assigned:

```
df['Cluster'] = df.Cluster.replace({
    0: "Careless",
    1: "Standard",
    2: "Target",
    3: "Sensible",
    4: "Careful"
})
```

10 Results

The analysis reveals the optimal number of clusters, indicating distinct customer segments within the mall dataset. These segments encompass diverse demographic and spending characteristics, providing valuable insights for targeted marketing strategies.

11 Discussion

The findings highlight the efficacy of K-means clustering in customer segmentation for mall environments. By identifying homogeneous customer groups, malls can tailor their marketing strategies to meet the needs and preferences of different segments.

12 Conclusion

This dissertation demonstrates the utility of K-means clustering for mall customer segmentation, providing actionable insights for mall management in optimizing marketing efforts and enhancing customer satisfaction.

References

[1] J. Han, J. Pei, and H. Tong, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2022.

[2] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Inc., 1988.

[3] V. Lakshman Narayana, S. Sirisha, G. Divya, N. Lakshmi Sri Pooja, and Sk. Afraa Nouf, *Mall Customer Segmentation Using Machine Learning*, 2022 International Conference on Electronics and Renewable Systems (ICEARS), pp. 1280–1288, doi: 10.1109/ICEARS53579.2022.9752447.