

CUSTOMER SEGMENTATION

(using K-Means)

PROJECT REPORT

Submitted

By

Name: Tejas Saiprasad Havaladar

Email: durgeshhavaladar.123@gmail.com

For

EXPOSYS DATA LABS

Bengaluru-India

ABSTRACT

Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

Customer segmentation enables a company to customize its relationships with the customers, as we do in our daily lives.

Companies employing customer segmentation operate under the fact that every customer is different and that their marketing efforts would be better served if they target specific, smaller groups with messages that those consumers would find relevant and lead them to buy something. Companies also hope to gain a deeper understanding of their customers' preferences and needs with the idea of discovering what each segment finds most valuable to more accurately tailor marketing materials toward that segment.

Benefits of customer segmentation include personalization, customer retention, better ROI for marketing, and revealing new opportunities.

The use of machine learning can be seen almost everywhere around us, be it Facebook recognizing you or your friends, or YouTube recommending you a video or two based on your history — Machine Learning is everywhere! However, the 'magic' of machine learning is not just limited to only these areas.

Machine Learning is broadly categorized as Supervised and Unsupervised Learning.

Supervised Learning is one in which we teach the machine by providing both independent and *dependent variables*, for example, Classifying or predicting values.

Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with no pre-existing labels

and with a minimum of human supervision. In contrast to supervised learning that usually makes use of human-labeled data, unsupervised learning, also known as self-organization allows for modeling of probability densities over inputs.

Two of the main methods used in unsupervised learning are principal component and cluster analysis. Cluster analysis is used in unsupervised learning to group, or segment, datasets with shared attributes in order to extrapolate algorithmic relationships. Cluster analysis is a branch of machine learning that groups the data that has not been labelled, classified or categorized. Instead of responding to feedback, cluster analysis identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. This approach helps detect anomalous data points that do not fit into either group.

Some of the most common algorithms used in unsupervised learning include: (1) Clustering, (2) Anomaly detection, (3) Neural Networks, and (4) Approaches for learning latent variable models. Each approach uses several methods.

TABLE OF CONTENTS

1. Introduction
 - 1.1 The Problem
 - 1.2 K-means Clustering
 - 1.3 Final Goal
2. Dataset
3. Methodology
 - 3.1 Tools used for implementing and visualizing
 - 3.2 K-means
4. Implementation
 - 4.1 What is Clustering
 - 4.2 K-Means Clustering
 - 4.3 The Elbow Method
 - 4.4 Clustering
5. Analysis
 - 5.1 Cluster Analysis
6. Conclusion

INTRODUCTION

Clustering is the task of dividing the population or data points into several groups, so that the data points in the same groups are more similar to other data points in the same group than those of other groups. In simple words, the goal is to segregate groups with similar traits and assign them to clusters.

Cluster analysis can be done based on the resources in which we try to find subgroups of samples based on resources or based on samples in which we try to find subgroups of resources based on samples. We will address resource-based clustering here. Clustering is used in market segmentation; where we try to find customers, who are similar to each other, whether in terms of behaviors or attributes, segmentation / compression of images; where we try to group similar regions, group documents based on topics, etc.

The Problem:

Companies employing customer segmentation operate under the fact that every customer is different and that their marketing efforts would be better served if they target specific, smaller groups with messages that those consumers would find relevant and lead them to buy something. Companies also hope to gain a deeper understanding of their customers' preferences and needs with the idea of discovering what each segment finds most valuable to more accurately tailor marketing materials toward that segment.

Malls or shopping complexes are often indulged in the race to increase their customers and hence making huge profits. To achieve this task machine learning is being applied by many stores already.

It is amazing to realize the fact that how machine learning can aid in such ambitions. The shopping complexes make use of their customers'

data and develop ML models to target the right ones. This not only increases sales but also makes the complexes efficient.

K-means Clustering:

Among many clustering algorithms, the K-means clustering algorithm is widely used because of its simple algorithm and fast convergence. However, the K-value of clustering needs to be given in advance and the choice of K-value directly affects the convergence result. To solve this problem, we mainly analyze four K-value selection algorithms, namely Elbow Method, Gap Statistic, give the pseudo code of the algorithm; and use the standard dataset Iris for experimental verification. K-means clustering is a method used for clustering analysis, especially in data mining and statistics. It aims to partition a set of observations into a number of clusters (k), resulting in the partitioning of the data into Voronoi cells. It can be considered a method of finding out which group a certain object really belongs to. It is used mainly in statistics and can be applied to almost any branch of study.

The approach K-means follows to solve the problem is called Expectation Maximization:

1. Specify number of clusters K .
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
4. Compute the sum of the squared distance between data points and all centroids.
5. Assign each data point to the closest cluster (centroid).
6. Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

Stopping Criteria for K-Means Clustering:

There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

- Centroids of newly formed clusters do not change
- Points remain in the same cluster
- Maximum number of iterations are reached

When to use Cluster Analysis?

If we are using a labeled data, we can use classification technique, whereas in case when the data is not labeled, we can cluster the data based on certain feature and try to label it on our own. So, when we use cluster analysis, we don't have labels (i.e.. data is not labeled) in the context of machine learning this is called as unsupervised learning.

Final Goal:

The goal of clustering is to maximize the similarity of observation within the cluster and maximize the dissimilarity between the clusters.

DATASET

Companies employing customer segmentation operate under the fact that every customer is different and that their marketing efforts would be better served if they target specific, smaller groups with messages that those consumers would find relevant and lead them to buy something. Companies also hope to gain a deeper understanding of their customers' preferences and needs with the idea of discovering what each segment finds most valuable to more accurately tailor marketing materials toward that segment.

1. Customer ID: It is the unique ID given to a customer
2. Gender: Gender of the customer
3. Age: The age of the customer
4. Annual Income (k\$): It is the annual income of the customer
5. Spending Score: It is the score (out of 100) given to a customer by the mall authorities, based on the money spent and the behavior of the customer.

METHODOLOGY

Tools used for implementing and visualizing:

Various python libraries were used for completing this project. Some of them are listed below: -

1) NumPy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

2) Pandas: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

3) Matplotlib: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

4) Seaborn: Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

5) Scikit-learn: Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbors, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

K-means:

K-means clustering algorithm is the most selected technique to cluster data. K-means is a nonhierarchical clustering and use looping to group

data into K groups. The K-means clustering start the iterative process by finding the initial centroid, or central point, of each group by randomly selecting representative data from raw data to be a centroid in each K data groups. Then assign each data to the closest group by calculating the Euclidean distance between each data record to each centroid to allocate the data record to the nearest group. After that each cluster will find new centroid to replace the initial one and repeat steps of Euclidean distance computation to group data members and send each member to group of the nearest centroid. The process will stop when each group has stable centroid and members do not change their groups.

The steps of k-means algorithm can be summarized as the following:

- 1) Specify group number and select initial centroid of each group.
- 2) Calculate Euclidean distance for each data member and centroid to assign members to the nearest centroid.
- 3) Calculate distance's mean of every data member and own centroid to define new centroid in each group.
- 4) Repeat steps 2 and 3 until each group has stable centroid or same centroid.

IMPLEMENTATION

The steps for implementation.

- Various libraries were imported to manipulate, visualize and modeling data.
- Certain components of the data are visualized for better insight.
- Selecting a k value for the model.
- Model implementation and obtaining the labels.
- Plotting the Clusters on as 2D graph.

What is Clustering?

Imagine that you have a group of chocolates and liquorice candies. You are required to separate the two eatables. Intuitively, you are able to separate them based on their appearances. The process of segregating objects into groups based on their respective characteristics is called clustering. In clusters, the features of objects in a group are similar to other objects present in the same group.

Clustering is used in various fields like image recognition, pattern analysis, medical informatics, genomics, data compression etc. It is part of the unsupervised learning algorithm in machine learning. This is because the data-points present are not labelled and there is no explicit mapping of input and outputs. As such, based on the patterns present inside, clustering takes place.

K-Means Clustering:

K-Means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the

intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

We then proceeded to perform K-means Clustering which will create different clusters to group similar spending activity based on their age and annual income. K Means Clustering selects random values from the data and forms clusters assigned. The closest values from the centre of each cluster were taken to update the cluster and reshape the plot (just like k-NN). The closest values are based on Euclidean Distance.

The Elbow Method:

The Elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

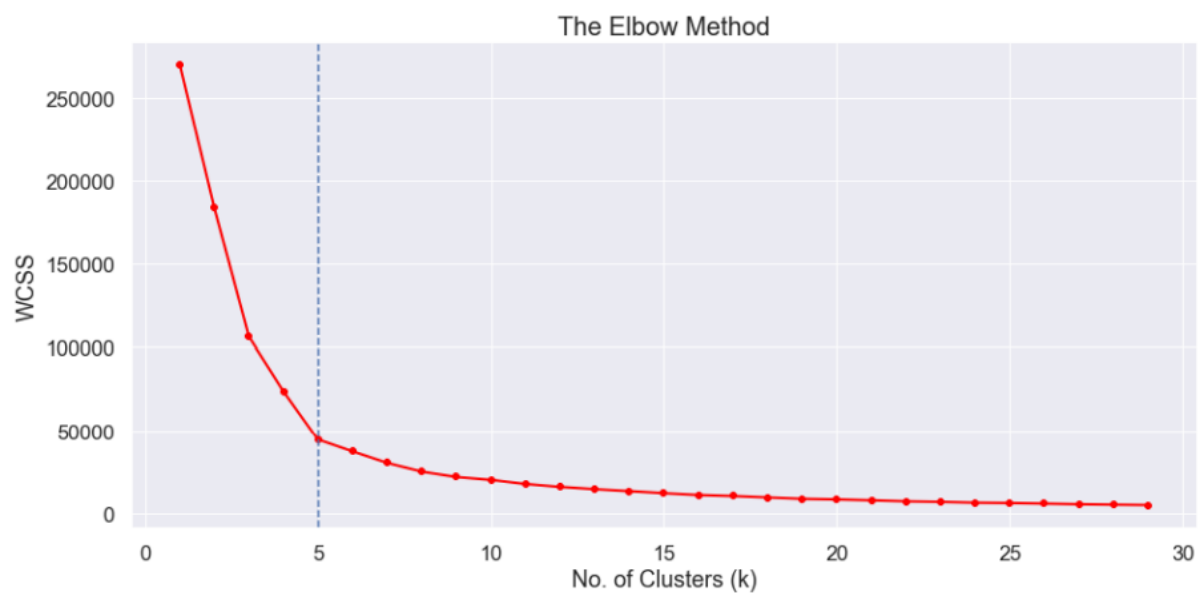
When these overall metrics for each model are plotted, it is possible to visually determine the best value for k. If the line chart looks like an arm, then the "elbow" (the point of inflection on the curve) is the best value of k. The "arm" can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.

We use the Elbow Method which uses Within Cluster Sum Of Squares (WCSS) against the the number of clusters (K Value) to figure out the

optimal number of clusters value. WCSS measures sum of distances of observations from their cluster centroids which is given by the 16 below formula.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

where Y_i is centroid for observation X_i . The main goal is to maximize number of clusters and in limiting case each data point becomes its own cluster centroid.



It is clear, that the optimal number of clusters for our data are 5, as the slope of the curve is not steep enough after it. When we observe this curve, we see that last elbow comes at $k = 5$, it would be difficult to visualize the elbow if we choose the higher range.

Clustering:

We will build the model for creating clusters from the dataset. We will use $n_clusters = 5$ i.e. 5 clusters as we have determined by the elbow method, which would be optimal for our dataset. We also get the centroids of the clusters by the k-means model. Visualizing the clusters will often give the fair idea about the data.



The Clusters are denoted by different colours and the centroids of the clusters is given by yellow colour. By this plot, it is easy to understand that we can divide the customers within 5 clusters of customers.

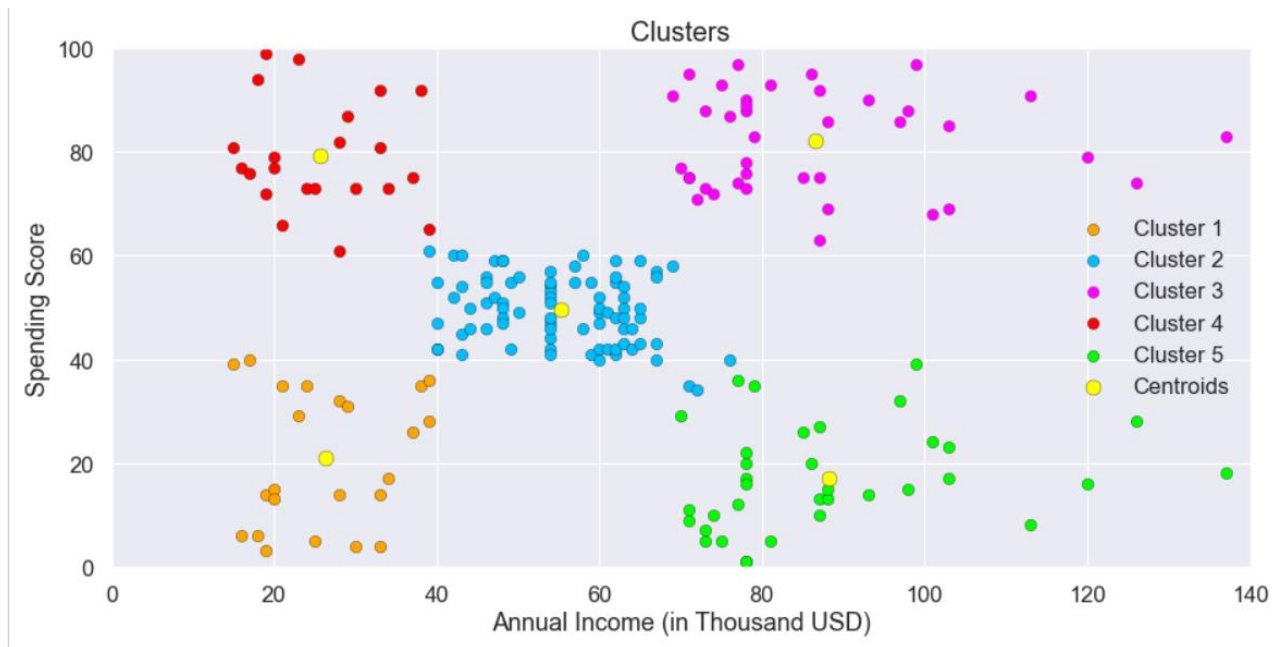
Analyzing Data using the above graph becomes much easier as it gives us a visual aid for better understanding of the data. Kmeans has divided the dataset into 5 clusters based on Annual income and the spending scores of the individual customers.

ANALYSIS

Cluster Analysis:

The following clusters are created by the model,

1. Cluster Orange
2. Cluster Blue
3. Cluster Purple
4. Cluster Red
5. Cluster Green



1. Cluster Orange - Balanced Customers:

They earn less and spend less. We can see people have low annual income and low spending scores, this is quite reasonable as people

having low salaries prefer to buy less, in fact, these are the wise people who know how to spend and save money. The shops/mall will be least interested in people belonging to this cluster.

2. Cluster Blue - Pinch Penny Customers:

Earning high and spending less. We see that people have high income but low spending scores, this is interesting. Maybe these are the people who are unsatisfied or unhappy by the mall's services. These can be the prime targets of the mall, as they have the potential to spend money. So, the mall authorities will try to add new facilities so that they can attract these people and can meet their needs.

3. Cluster Purple - Normal Customer:

Customers are average in terms of earning and spending. An Average consumer in terms of spending and Annual Income we see that people have average income and an average spending score, these people again will not be the prime targets of the shops or mall, but again they will be considered and other data analysis techniques may be used to increase their spending score.

4. Cluster Red - Spenders:

This type of customers earns less but spends more. Annual Income is less but spending high, so can also be treated as potential target customer. We can see that people have low income but higher spending scores, these are those people who for some reason love to buy products more often even though they have a low income. Maybe it's because these people are more than satisfied with the mall services.

The shops/malls might not target these people that effectively but still will not lose them.

5. Cluster Green - Target Customers:

Earning high and also spending high Target Customers. Annual Income High as well as Spending Score is high, so a target consumer. we see that people have high income and high spending scores, this is the ideal case for the mall or shops as these people are the prime sources of profit. These people might be the regular customers of the mall and are convinced by the mall's facilities.

CONCLUSION

K means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of K means is to group data points into distinct non-overlapping subgroups. One of the major applications of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.

Firstly, we tried to visualize the gender distribution from the given dataset. We made a bar plot to check the distribution of male and female population in the dataset. The female population clearly outweighs the male counterpart.

Then we tried to visualize the age distribution. We made a bar plot to check the distribution of number of customers in each age group. Clearly the 26–35 age group outweighs every other age group.

Finally using the K-means algorithm we tried to analyze their annual incomes and spending scores.

So, looking at the final graph we can conclude the following: -

In cluster 4 (light blue colored) we can see people have low annual income and low spending scores, this is quite reasonable as people having low salaries prefer to buy less, in fact, these are the wise people who know how to spend and save money. The shops/mall will be least interested in people belonging to this cluster.

In cluster 2 (blue colored) we can see that people have low income but higher spending scores, these are those people who for some reason love to buy products more often even though they have a low income. Maybe it's because these people are more than satisfied with the mall

services. The shops/malls might not target these people that effectively but still will not lose them.

In cluster 1 (red colored) we see that people have average income and an average spending score, these people again will not be the prime targets of the shops or mall, but again they will be considered and other data analysis techniques may be used to increase their spending score.

In cluster 3 (green-colored) we see that people have high income and high spending scores, this is the ideal case for the mall or shops as these people are the prime sources of profit. These people might be the regular customers of the mall and are convinced by the mall's facilities.

In cluster 5 (pink colored) we see that people have high income but low spending scores, this is interesting. Maybe these are the people who are unsatisfied or unhappy by the mall's services. These can be the prime targets of the mall, as they have the potential to spend money. So, the mall authorities will try to add new facilities so that they can attract these people and can meet their needs.

Finally, based on our machine learning technique we may deduce that to increase the profits of the mall, the mall authorities should target people belonging to cluster 5 and cluster 1 and should also maintain its standards to keep the people belonging to cluster 3 and cluster 2 happy and satisfied.