Segmenting customers sector with k-means

García Manuel, ISC A01701414, Tecnológico de Monterrey Campus Querétaro

Abstract - This document presents the implementation and explanation of a clustering analysis employing a K-means algorithm, as well as its change in performance due to the amount of clusters.

I. INTRODUCTION

Nowadays the amount of information caused by social networks, surveys, internet and many others, created an enormous opportunity to understand the behaviour and relationship between previously unrecognized features. This fact is commonly applied to the marketing sector, the quantity of clients and data they generate around certain products have given a competitive advantage to the ones that are able to process it and figure out the keys to create a precise targeting to them. The concept employed is known as customer segmentation, and is defined as the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately. Normally through features like demographics or personal characteristics.
Within this background a clustering problem is identified and one of the approaches to handle its K-means.

II. STATE OF THE ART

K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The standard algorithm was first proposed by Stuart Lloyd of Bell Labs in 1957 as a technique for pulse-code modulation.
This method has shown a constant practicity due to the ease of computation and simplicity to the math involved.

III. DATASET

The dataset comes from Kaggle, it contains 200 registries of customers belonging to a Mall, which are separated in different categories such as gender, age, annual income and spending score. Figure 1 shows an example of the data from the dataset.

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Figure 1. Dataset example.

The distribution of the features employed for this classification are age and spending score and a representation of the distribution of them can be shown in Figure 2.

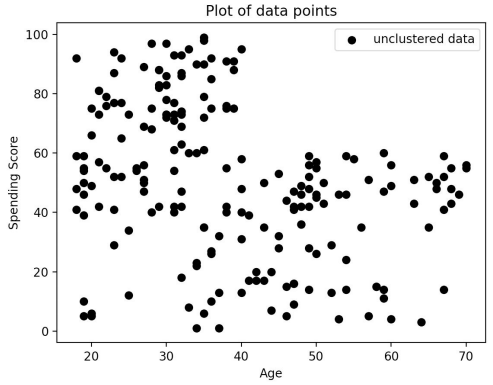

Figure 2. Plot of features age and spending score.

IV. MODEL PROPOSAL

The purpose of this implementation is to define different groups so a campaign could be targeted to each of them.

The model consists of K centroids which would categorise the data in K sectors. In order to create a relation between centroids and data points we compute the euclidean distance between each of the centroids and it will belong to the one with the minimum distances.

This procedure is iterated a 100 times and on each of it, the proximity between the data and the centroids arrange them to be classified in a different cluster, after this a new centroid is computed within the mean of all the points belonging to each cluster.

The result of an iteration for k clusters is observable in Figure 3.

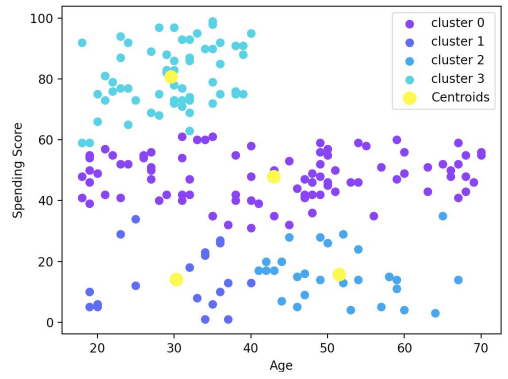

Figure 3. Classification in 4 clusters

V. TEST AND OPTIMIZATION

In order to acquire a good and accurate result several tests were performed with the presented dataset. In this particular algorithm the optimization spins around the number of algorithms needed for the dataset employed. A common way to address this issue is by applying Within-cluster Sum of Squares (WCSS) to the final result given by the model. The core of this solution regards that the final result would have a minimum Sum of Squared distances between data point and centroids.

The optimal cluster number is given by the “elbow” also called inflexion in the trend, observable in Figure 4.

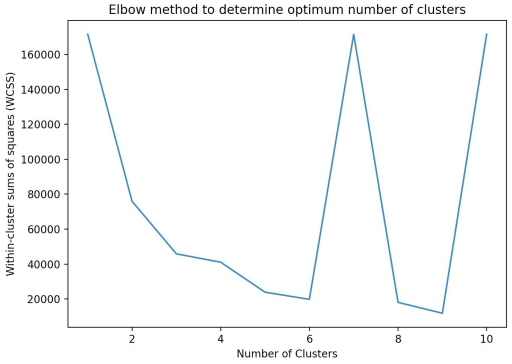

Figure 4. Graphical representation of the WCSS and the inflexion point observable in cluster 5.

The different outputs based on the number of k clusters are observable in the following images.

Based on these results the optimal number of clusters for the given dataset is 6.

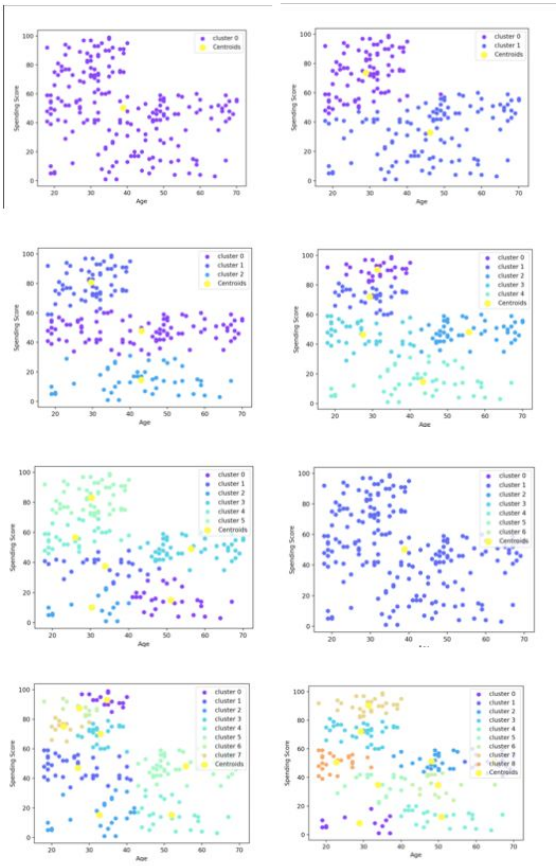

Figure 5. Collection of clustering results.

VI. CONCLUSION

Based on the set of customers from the dataset is assurable that the optimal number of segments for classifying the market of the mall is 6. With this information the marketing team of the institution will be able to acquire the best profit.

On the other hand, the simplicity of the algorithm and model presented could satisfy common unsupervised and classification problems. As well, the optimization method applied is not feasible for large datasets so other methods should be applied.

Finally, the use of these tools in real life problems are definitely an aid to solve enterprise issues.