TECHKRITI IIT KANPUR ML/AI SUMMER INTERNSHIP PROJECT SYNOPSIS

Myself Harshvardhan Pardeshi, B. Tech Information Technology 3rd Year 6th Semester student from Guru Ghasidas University, Bilaspur (C.G.).

This file is the documentation for my project titled **“Mall Customers Segmentor”**. I created this project as it was assigned to me during my internship period at ***TECHKRITI’21 IIT KANPUR*** through my trainer-in-charge Mr. Chintoo Kumar.

PROBLEM STATEMENT

Provided the dataset of mall customers, you have to create partitions of customers in such a way that the members of a cluster share some similar characteristics. First, use the elbow method to find the optimal number of clusters and then use K-means clustering.

OBJECTIVE

To perform customer segmentation (also known as market basket analysis) using unsupervised ML technique. You are responsible for building an unsupervised machine learning model for the mall.

BACKGROUND

Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things. It can be defined as:

*“Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.”*

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.

1. Elbow Method for finding optimal k:

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use. Find the average distance of each point in a cluster to its centroid, and represent it in a plot. Pick the value of k, where the average distance falls suddenly.

*“To find the optimal number of clusters (k),* ***observe the plot and find the value of k for which there is a sharp and steep fall of the distance****. This is will be an optimal point of k where an elbow occurs.”*

1. Silhouette Method for finding optimal k:

The silhouette Method is also a method to find the optimal number of clusters and interpretation and validation of consistency within clusters of data. The silhouette method computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters. by providing a succinct graphical representation of how well each object has been classified.

*“Compute silhouette coefficients for each of point, and average it out for all the samples to get the silhouette score.”*

1. K-means Clustering:

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science.

This Algorithm groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

*“It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.”*

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

* Determines the best value for K center points or centroids by an iterative process.
* Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

1. Fuzzy-c-means Clustering:

The unsupervised k-means clustering algorithm gives the values of any point lying in some particular cluster to be either as 0 or 1 i.e., either true or false. But the fuzzy logic gives the fuzzy values of any particular data point to be lying in either of the clusters. Here, in fuzzy c-means clustering, we find out the centroid of the data points and then calculate the distance of each data point from the given centroids until the clusters formed becomes constant.

SYNOPSIS

Firstly, I observed the dataset which was provided to me. Then as per the given instructions I had to create an unsupervised machine learning model. I found out the optimal number of clusters using the Elbow Method and Silhouette method. After finding the optimal number of clusters I performed clustering using K-means clustering method.

After some research I found out another approach for performing clustering named fuzzy-c-means clustering. Furthermore, I discovered that fuzzy-c-means is a better approach than K-means clustering. So I created an unsupervised machine learning model using fuzzy-c-means clustering as well.

CONCLUSION

So, through this project, I have successfully made clusters of the customers of the mall using Unsupervised Machine Learning Techniques (K-means clustering and fuzzy-c-means clustering)

REFERENCES AND BIBLIOGRAPHY

1. Notebooks provided by Chintoo Sir
2. An Introduction to Machine Learning by Miroslav Kubat (provided by Chintoo Sir)
3. https://towardsdatascience.com/fuzzy-c-means-clustering-is-it-better-than-k-means-clustering-448a0aba1ee7