Clustering in Machine Learning: A Case study

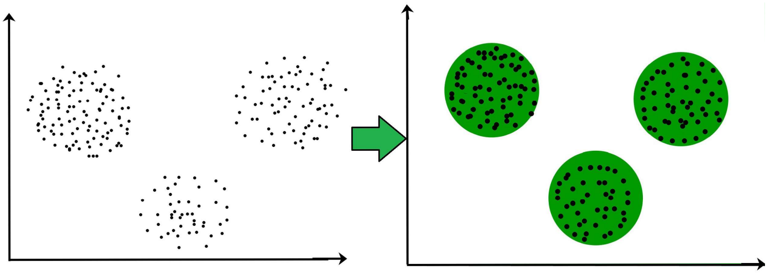
Mr. Dattatraya Bhikaji Raymal
Department of Computer Science ,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad
dattatrayaraymal@gmail.com-

Abstract

The present study carried out in between January 2021 to June 2021. It is basically a type of *unsupervised learning method* . An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples. **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

Introduction:-

Clustering in Machine Learning is one of the main method used in the unsupervised learning technique for statistical data analysis by classifying population or data points of the given dataset into several groups based upon the similar features or properties, while the data point in the different group poses the highly dissimilar property or feature. The clustering methods used in machine learning (i.e. k-mean clustering, Density methods, Grid-based methods, Hierarchical bases method, etc.) performs the collection of the data points based upon the similarity and dissimilarity between them.



How does Clustering Work in Machine Learning?

In clustering, we group unlabeled data set which is known as unsupervised learning. When we first group unlabeled data, we need to find a similar group. When we create a group, we need to understand the features of datasets i.e. similar things. If we create a group by one or two features, it is easy to measure similarity.

- **Example #1:** Movies by the director. Once clustering is done, each cluster is assigned cluster number which is known as ClusterID. Machine learning system like YouTube uses clusterID to represent complex data most easily.

- **Example #2:** YouTube uses our search history or watched history and suggests videos we might like. Feature data set for Facebook contains is people we follow, pages we follow, comments we input, photos or videos we like, pictures or photos we tag at. Clustering Facebook video or photo will replace a set of features with single clusterID due to the compressing of data.

Top 4 Methods of Clustering in Machine Learning

Below are the methods of Clustering in Machine Learning:

1. Hierarchical

The name clustering defines a way of working, this method forms a cluster in a hierarchal way. The new cluster is formed using a previously formed structure. We need to understand the differences between the Divisive approach vs Agglomerative approach. Agglomerative is a bottom-up approach, it starts with individual points in a cluster and combines some arbitrary. Divisive begins with a single cluster, all points in a cluster and divides it into multiple clusters.

2. Density-Based

In this method, a dense region is considered as a cluster who's having some similarities. It is different from the lower dense region of the object space. DBSCAN is known as the Density-based spatial clustering of applications with noise. For data object-orientation, DBSCAN looks for some epsilon we set some radius epsilon and the minimum number of points. Within a radius, if we surpass some minimum number of points then we rank a cluster high density. So, this way we can consider data with a region of high density. DBSCAN differs from the centroid method of clustering as it is not a strict approach. Noise points are points in low-density areas that are left unlabelled or labeled as outliers. That's the reason we don't require specific K. We can specify minimum points for high-density region and radius we want for a region to be or clusters to be.

3. Partitioning

When we have a dataset of N number of objects. This method constructs "K" as the partition of data. This partition is the cluster i.e. construct K, partition (K<=N).

Requirements to be Met:

- Each group or dataset must contain at least one object.
- Each object should belong to one group only.

One of the examples of partitioning is K-means clustering.

4. Grid-based

Object space, a finite number of cells forms a grid structure. This method provides fast cluster processing. These are independent of object space.

Applications of Clustering in Machine Learning

Below are the applications of Clustering in Machine Learning:

1. Medical

The doctor can use a clustering algorithm to find the detection of disease. Let's take an example of thyroid disease. Thyroid disease dataset can be identified using clustering algorithm when we apply unsupervised learning on a dataset which contains thyroid and non-thyroid dataset.

Clustering will identify the cause of the disease and will give a successful result search.

2. Social Network

We are the generation of the internet era, we can meet any person or got to know about any individual identity through the internet. Social networking sites use clustering for content understanding, people face or location of the user. When unsupervised learning is used in social, it is useful for the translation of language. **For example,** Instagram and Facebook provide the feature of translation of language.

3. Marketing

We can see or observe that different technology is growing beside us and people are attracting to use those technologies like cloud, digital marketing. To attract a greater number of customers every company is developing easy to use features and technology. To understand the customer, we can use clustering. Clustering will help the company to understand the user segment and then categorize each customer. This way we can understand the customer and find similarities between customers and group them.

4. Banking

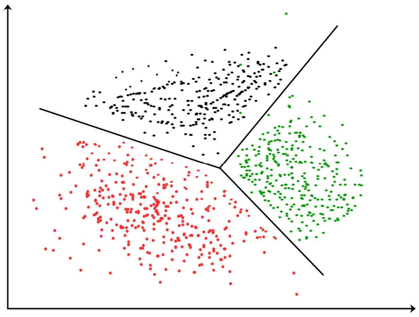
We have observed that fraud of money is happening around us and the company is warning customers about it. With the help of clustering, insurance companies can find fraud, acknowledge customers about it and understand policies brought by the customer.

5. Google

Google is one of the search engine people uses. Let’s take an example when we search for some information like pet store in the area, Google will provide us with different options. This is the result of clustering, clustering of similar result that is provided to you.

Clustering Algorithms :

K-means clustering algorithm – It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partition n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster .



Applications of Clustering in different fields

- **Marketing :** It can be used to characterize & discover customer segments for marketing purposes.

- **Biology :** It can be used for classification among different species of plants and animals.
- **Libraries :** It is used in clustering different books on the basis of topics and information.
- **Insurance :** It is used to acknowledge the customers, their policies and identifying the frauds

Conclusion

We have learned about clustering and machine learning. Way of clustering works in machine learning. Information about unsupervised learning. The real-time usage of unsupervised learning. Methods of clustering and how each method works in machine learning.

Recommended Article

This is a guide to Clustering in Machine Learning. Here we discuss the top 4 methods of clustering in machine learning along with applications. You can also go through our other suggested articles to learn more –

1. Machine Learning Frameworks Top 10
2. K- Means Clustering Algorithm with Advantages
3. Introduction to Machine Learning Techniques
4. Machine Learning Models | Top 5 Types

References :-

❖ https://www.educba.com/kernel-methods/

❖ Deep learning with python by Francois chollet ,and the publication of Manning Publications Co. 20 Baldwin RoadPO Box 761Shelter Island, NY 11964 ISBN 9781617294433

❖ https://www.geeksforgeeks.org/clustering-in-machine-learning/

❖ Master In Data Science: A Practical Understanding Of Data Science: Decision Trees And Random Forests by Merna Laperouse (Author) ISBN-13: 979-8749666083 Publisher : Independently published (May 6, 2021)

❖ Big Data Analytics for Internet of Things by Tausifa Jan Saleem (Editor), Mohammad Ahsan Chishti (Editor) ISBN-13: 978-1119740759 , ISBN-10: 1119740754 Publisher : Wiley; 1st edition (April 20, 2021)

❖ Reverse Clustering: Formulation, Interpretation and Case Studies by Jan W. Owsiński (Author), Jarosław, Stańczak (Author), Karol Opara (Author), Sławomir Zadrożny (Author), Janusz Kacprzyk (Author) ISBN-10 : 3030693589 Publisher : Springer; 1st ed. 2021 edition (March 4, 2021)

❖ Data Clustering: Algorithms and Applications by Charu C. Aggarwal (Editor), Chandan K. Reddy (Editor) ASIN : B00EYROAQU ISBN-13 978-1466558212 Publication date 21 August 2013 Publisher : Chapman and Hall/CRC; 1st edition (21 August 2013