Clustering in Machine Learning

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset.

It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

It does it by finding some similar patterns in the unlabelled dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns.

It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.

After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets.

Example:

Let's understand the clustering technique with the real-world example of Mall: When we visit any shopping mall, we can observe that the things with similar usage are grouped together.

Such as the t-shirts are grouped in one section, and trousers are at other sections, similarly, at vegetable sections, apples, bananas, Mangoes, etc., are grouped in separate sections, so that we can easily find out the things.

The clustering technique can be widely used in various tasks. Some common uses of clustering are grouping documents according to the topic.

The clustering technique can be widely used in various tasks. Some most common uses of this technique are:

1.Market Segmentation

2.Statistical data analysis

3.Social network analysis

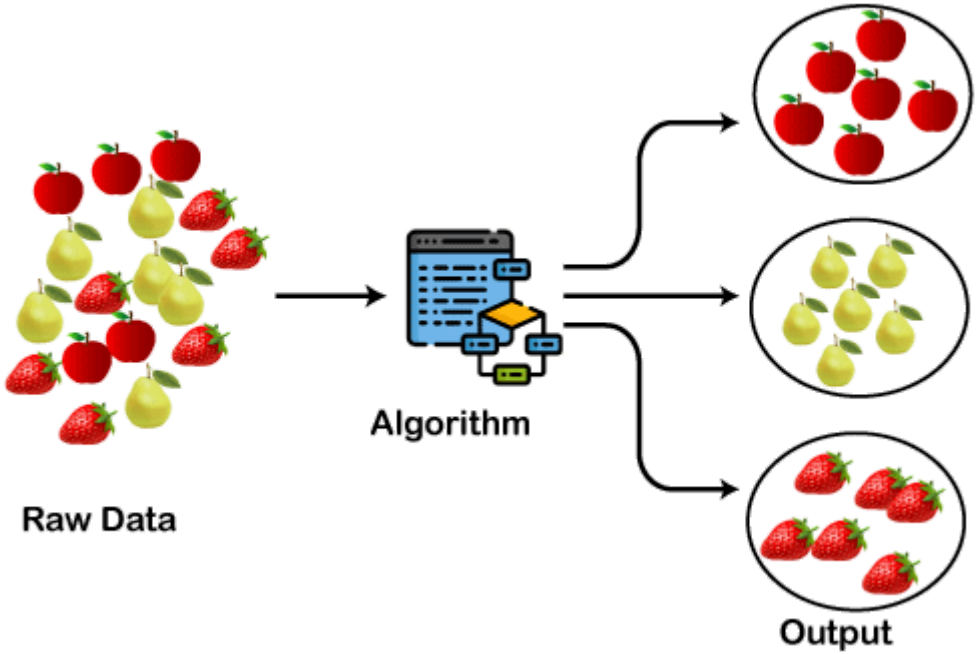
4.Image segmentation

5.Anomaly detection, etc.

Apart from these general usages, it is used by the Amazon in its recommendation system to provide the recommendations as per the past search of products.

Netflix also uses this technique to recommend the movies and web-series to its users as per the watch history.

The below diagram explains the working of the clustering algorithm. We can see the different fruits are divided into several groups with similar properties.



Types of Clustering Methods

The clustering methods are broadly divided into Hard clustering (datapoint belongs to only one group) and Soft Clustering (data points can belong to another group also). But there are also other various approaches of Clustering exist. Below are the main clustering methods used in Machine learning:

1.Partitioning Clustering

2.Density-Based Clustering

3.Distribution Model-Based Clustering

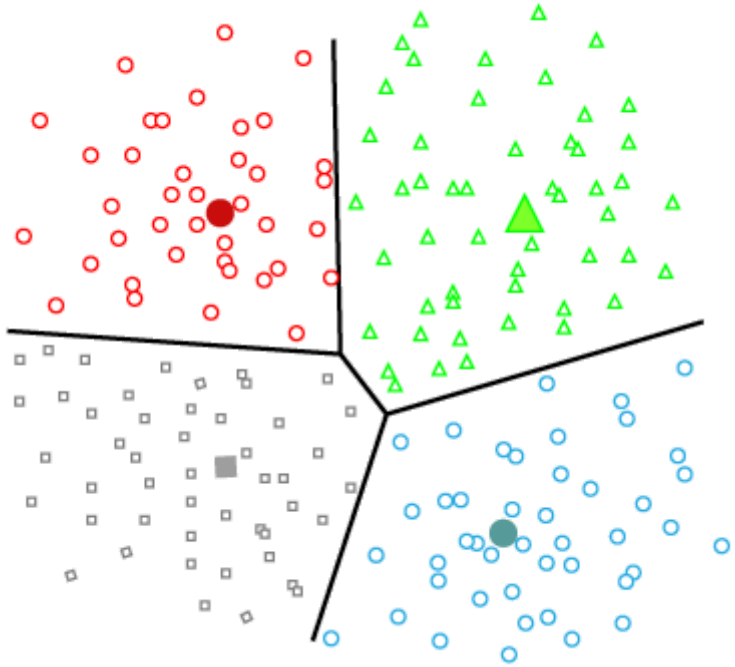
Partitioning Clustering

It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the centroid-based method.

The most common example of partitioning clustering is the K-Means Clustering algorithm.

In this type, the dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups.

The cluster center is created in such a way that the distance between the data points of one cluster is minimum as compared to another cluster centroid.



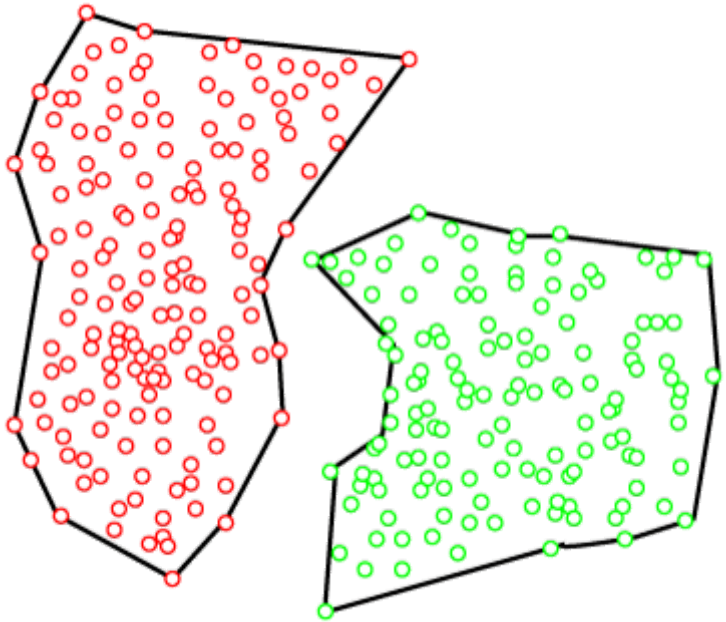
Density-Based Clustering

as into clusters, and the arbitrarily shaped distributions are formed as long

This algorithm does it by identifying different clusters in the dataset and connects the areas of high densities into clusters.

These algorithms can face difficulty in clustering the data points if the

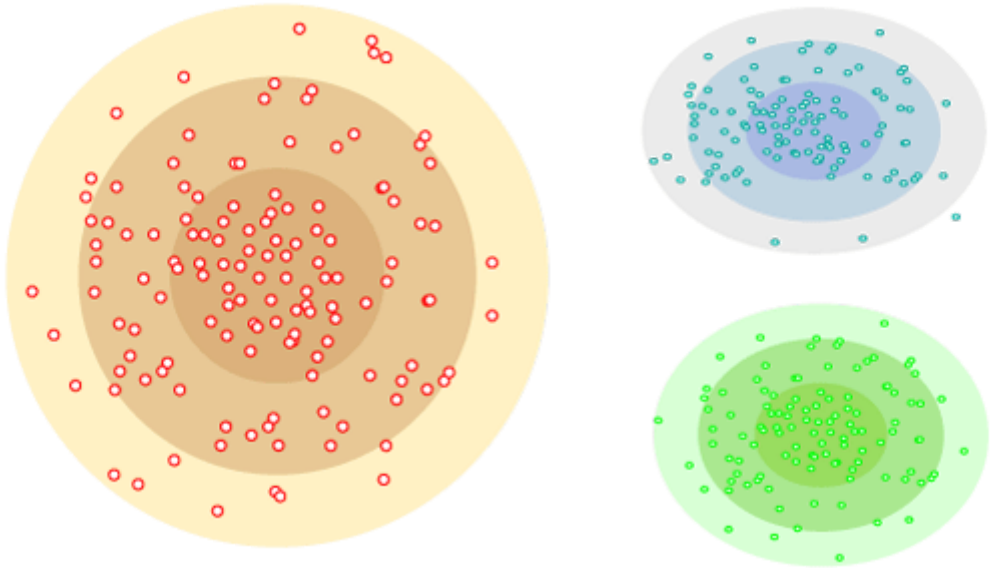
dataset has varying densities and high dimensions.



Distribution Model-Based Clustering

In the distribution model-based clustering method, the data is divided based on the probability of how a dataset belongs to a particular distribution. The grouping is done by assuming some distributions commonly Gaussian Distribution.

The example of this type is the Expectation-Maximization Clustering algorithm that uses Gaussian Mixture Models (GMM).



Double-click (or enter) to edit

popular Clustering algorithms that are widely used in machine learning:

1.K-Means algorithm: The k-means algorithm is one of the most popular clustering algorithms. It classifies the dataset by dividing the samples into different clusters of equal variances. The number of clusters must be specified in this algorithm. It is fast with fewer computations required, with the linear complexity of O(n).

2.Mean-shift algorithm: Mean-shift algorithm tries to find the dense areas in the smooth density of data points. It is an example of a centroid-based model, that works on updating the candidates for centroid to be the center of the points within a given region.

of Applications with Noise. It is an example of a density-based model his algorithm, the areas of high density are separated by the areas of low density. Because of this, the clusters can be found in any arbitrary shape.

Applications of Clustering

Below are some commonly known applications of clustering technique in Machine Learning:

1. In Identification of Cancer Cells: The clustering algorithms are widely used for the identification of cancerous cells. It divides the cancerous and non-cancerous data sets into different groups.

2. In Search Engines: Search engines also work on the clustering technique. The search result appears based on the closest object to the search query. It does it by grouping similar data objects in one group that is far from the other dissimilar objects. The accurate result of a query depends on the quality of the clustering algorithm used.

3. Customer Segmentation: It is used in market research to segment the customers based on their choice and preferences.

4. In Biology: It is used in the biology stream to classify different species of plants and animals using the image recognition technique.

5. In Land Use: The clustering technique is used in identifying the area of similar lands use in the GIS database. This can be very useful to find that for what purpose the particular land should be used, that means for which purpose it is more suitable.

Colab paid products - Cancel contracts here

To undo cell deletion use Ctrl+M Z or the Undo option in the Edit menu