## horizontal line



Clustering

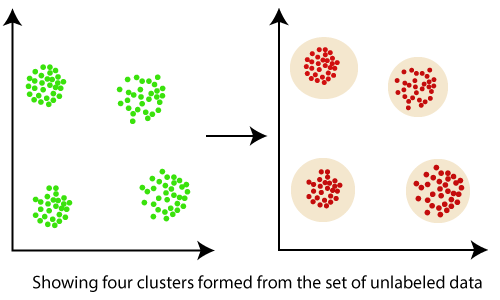
-Definition

-Clustering overview

.

Definition

The process of making a group of abstract objects into classes of similar objects is known as clustering.

Clustering overview

Clustering, falling under the category of **unsupervised machine learning**, is one of the problems that machine learning algorithms solve.

Clustering only utilizes input data, to determine patterns, anomalies, or similarities in its input data.

A good clustering algorithm aims to obtain clusters whose:

* The intra-cluster similarities are high, It implies that the data present inside the cluster is similar to one another.
* The inter-cluster similarity is low, and it means each cluster holds data that is not similar to other data.
* Data objects of a cluster can be considered as one group.
* We first partition the information set into groups while doing cluster analysis. It is based on data similarities and then assigns the levels to the groups.
* The over-classification main advantage is that it is adaptable to modifications, and it helps single out important characteristics that differentiate between distinct groups.

## Applications of cluster analysis

* In many applications, clustering analysis is widely used, such as data analysis, market research, pattern recognition, and image processing.
* It assists marketers to find different groups in their client base and based on the purchasing patterns. They can characterize their customer groups.
* It helps in allocating documents on the internet for data discovery.
* Clustering is also used in tracking applications such as detection of credit card fraud.
* As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to analyze the characteristics of each cluster.
* In terms of biology, It can be used to determine plant and animal taxonomies, categorization of genes with the same functionalities and gain insight into structure inherent to populations.
* It helps in the identification of areas of similar land that are used in an earth observation database and the identification of house groups in a city according to house type, value, and geographical location.

## Requirements of Clustering

The following points throw light on why clustering is required in data mining −

* Scalability − We need highly scalable clustering algorithms to deal with large databases.
* Ability to deal with different kinds of attributes − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
* Discovery of clusters with attribute shape − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
* High dimensionality − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
* Ability to deal with noisy data − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
* Interpretability − The clustering results should be interpretable, comprehensible, and usable.