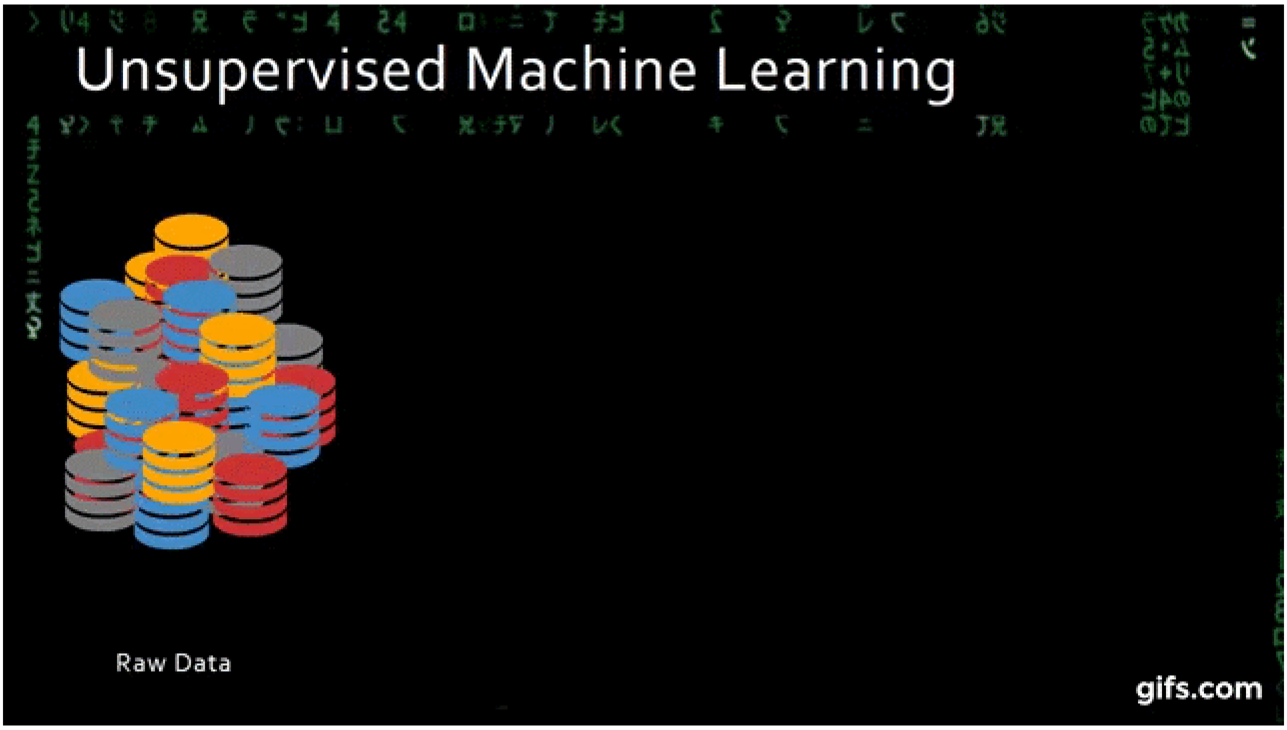
Introduction to Clustering



Clustering is considered to be the most important technique of **unsupervised** learning.

Before we start about clustering, let's understand about cluster

1. **Cluster** is the collection of data objects which are **similar to one another** within the same group **(class or category)** and are different from the objects in the other clusters.
2. It is an unsupervised learning technique in which there is **predefined classes** and prior information which defines how the data should be **grouped or labeled** into separate classes
3. It can also work as a standalone tool to get the insights about the **data distribution** or as a **preprocessing** step in other algorithms.

Why Clustering?

Clustering allows us to find hidden relationship between the data points in the dataset.

Examples:

1. In marketing, customers are segmented according to similarities to carry out targeted marketing.
2. Given a collection of text, we need to organize them, according to the content similarities to create a topic hierarchy

3. Detecting distinct kinds of pattern in image data (Image processing). It's effective in biology research for identifying the underlying patterns.

Classification Vs Clustering

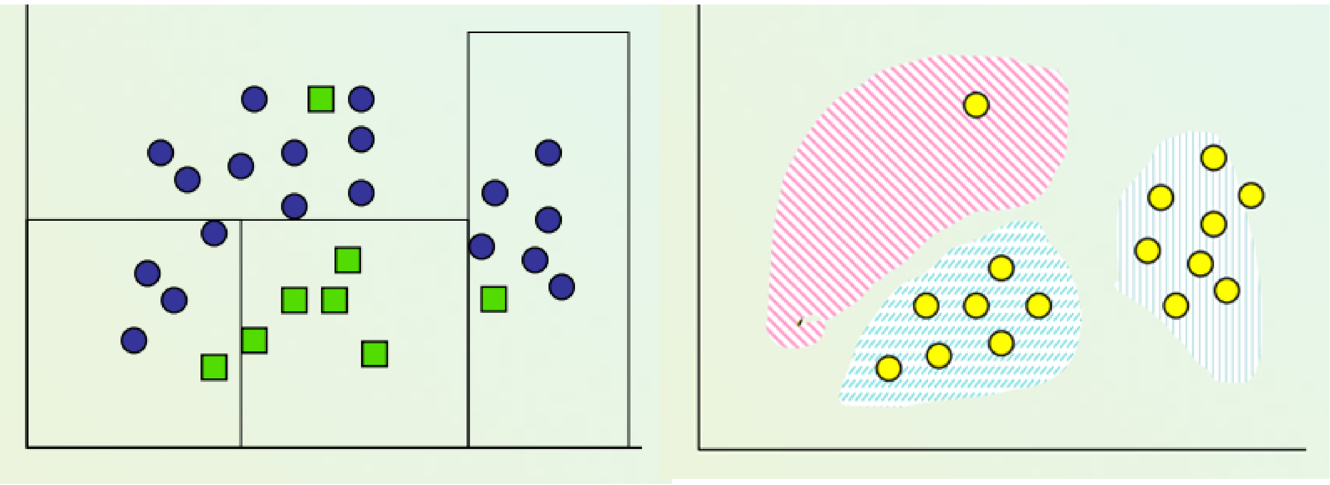
Let's understand how classification in supervised learning is different from clustering in unsupervised learning.

Classification

In Supervised learning our model learns a method for predicting the instance class from a pre-labeled (classified) instances.

Clustering

In unsupervised learning our model tries to find **"natural"** grouping of instances for a given unlabeled data.

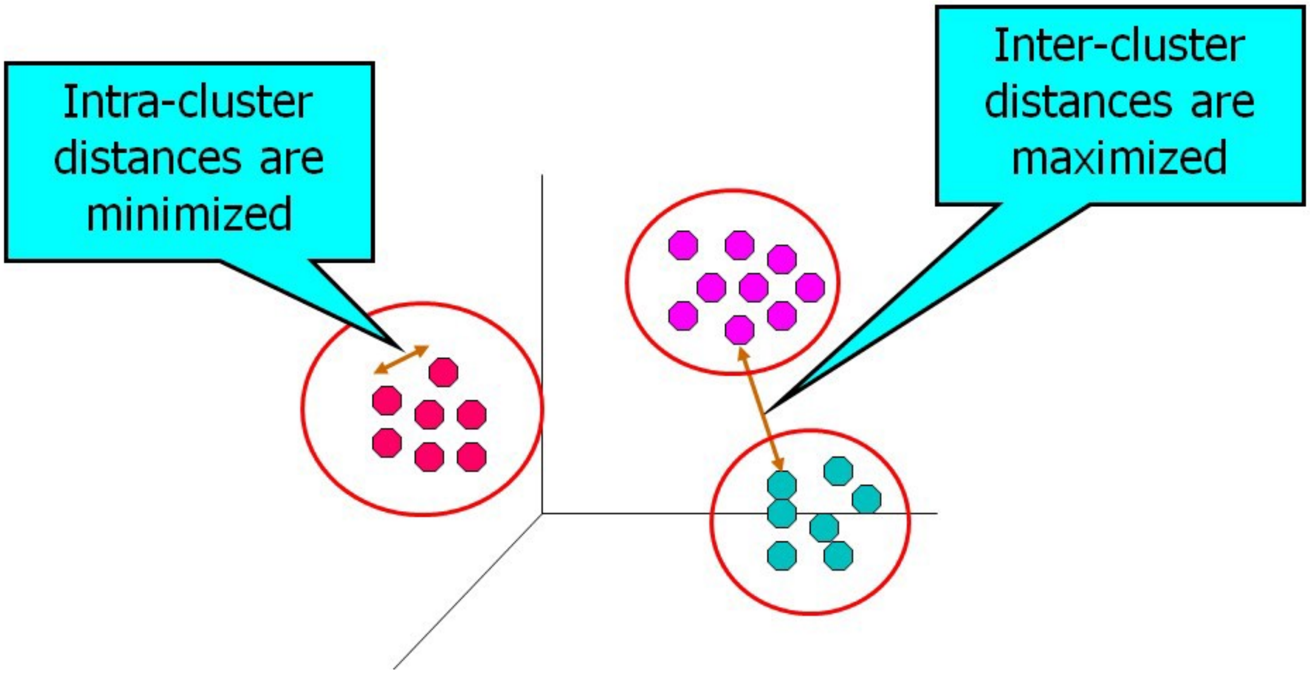


How do we define good Clustering algorithms?

High quality clusters can be created by reducing the distance between the objects in the same cluster known as intra-cluster minimization and increasing the distance with the objects in the other cluster known as inter-cluster maximization.

Intra-cluster minimization: The closer the objects in a cluster, the more likely they belong to the same cluster.

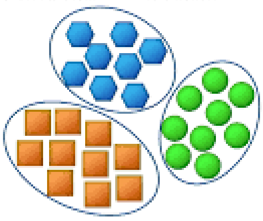
Inter-cluster minimization: This makes the separation between two clusters. The main goal is to maximize the distance between 2 clusters.

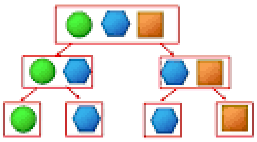

Intra-cluster distances are minimized
Inter-cluster distances are maximized

There are lot of clustering algorithms and they all use different techniques to cluster.

They can be classified into two categories as:

1. Flat or partitioning algorithms
2. Hierarchical algorithms





Flat or Partitioning algorithm

This algorithm try to divide the dataset of interest into predefined number of groups/ clusters.

All the groups/ clusters are independent of each other.

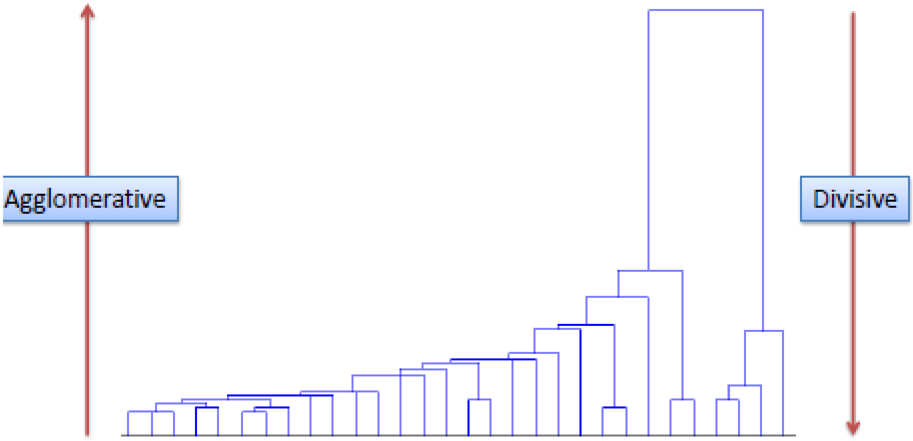
For Example: K-means

Hierarchical Clustering Algorithm

Hierarchical clustering does not partition the dataset into clusters in a single step. Instead it involves multiple steps which run from a single cluster containing all the data points to n clusters containing single data point.

This algorithm is further classified into Divisive and Agglomerative Methods.

Hierarchical clustering can be shown using dendrogram.



Divisive Method

This method is also known as top-down clustering method. It assigns all the data points to a single cluster and then it partitions the cluster to two least similar clusters. Then the same method is applied recursively on both the clusters until we get the cluster of each data point.

Agglomerative method

It is also known as bottom up clustering method. Here it assigns n data points to n clusters and joins the most similar clusters by computing the similarity i.e the distance between each of the clusters. This process is continued until we get a single cluster.