Certificate Course in Machine Learning using Python [6 Weeks]

Clustering Problem
clustering vs classification

- Classification is the process of classifying the data with the help of class labels whereas, in clustering, there are no predefined class labels.
- Classification is supervised learning, while clustering is unsupervised learning.
- In Classification, algorithms like Decision trees, Bayesian classifiers are used whereas, in Clustering, algorithms like K-means, Expectation Maximization is used.
- Classification has prior knowledge of classes but the cluster doesn't have any prior knowledge of classes.
- Mostly, clustering deals with unsupervised data; thus, unlabeled whereas classification works with supervised data; thus, labeled. This is one of the major reasons why clustering does not need training sets while classification does.
- Clustering seeks to verify how data are similar or dissimilar among each other while classification focuses on determining data's "classes" or groups. This makes the clustering process more focused on boundary conditions and the classification analysis more complicated in the sense that it involves more stages.
- **Example of classification**: classification between gender, images.
- **Example of a cluster**: discovery of patterns, grouping.

Types of Clustering

Broadly speaking, clustering can be divided into two subgroups :

- **Hard Clustering:** In hard clustering, each data point either belongs to a cluster completely or not. For example, each customer is put into one group out of the 10 groups.
- **Soft Clustering:** In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, from the above scenario each costumer is assigned a probability to be in either of 10 clusters of the retail store.

Types of Clustering Algorithms

In total, there are five distinct types of clustering algorithms. They are as follows –

• Partitioning Based Clustering

• Hierarchical Clustering

• Model-Based Clustering

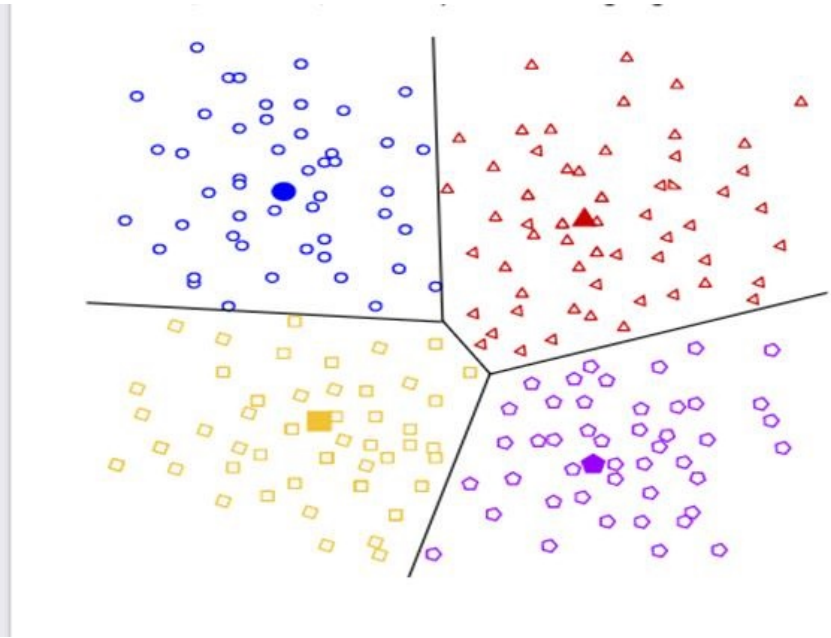
• Density-Based Clustering

• Fuzzy Clustering

Partitioning/ Centroid-based Clustering

- In this type of clustering, the algorithm subdivides the data into a subset of k groups.
- These k groups or clusters are to be pre-defined. It divides the data into clusters by satisfying these two requirements –

1. Firstly, Each group should consist of at least one point.
2. Secondly, each point must belong to exactly one group. K-Means Clustering is the most popular type of partitioning clustering method.

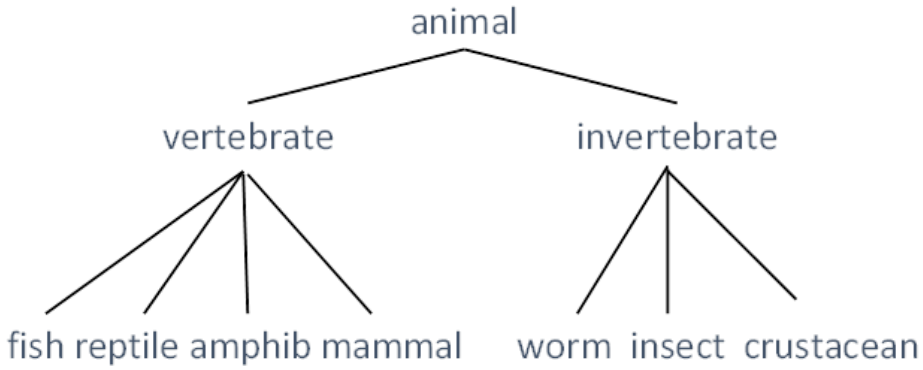


Hierarchical Clustering

- The basic notion behind this type of clustering is to create a hierarchy of clusters.
- As opposed to Partitioning Clustering, it does not require pre-definition of clusters upon which the model is to be built.

There are two ways to perform Hierarchical Clustering.

- The first approach is a bottom-up approach, also known as Agglomerative Approach.
- The second approach is the Divisive Approach which moves hierarchy of clusters in a top-down approach. As a result of this type of clustering, we obtain a tree-like representation known as a dendogram.



Density-Based Models

- In these type of clusters, there are dense areas present in the data space that are separated from each other by sparser areas.
- These type of clustering algorithms play a crucial role in evaluating and finding non-linear shape structures based on density.
- The most popular density-based algorithm is DBSCAn which allows spatial clustering of data with noise.
- It makes use of two concepts – Data Reachability and Data Connectivity.
- These algorithms have difficulty with data of varying densities and high dimensions. Further, by design, these algorithms do not assign outliers to clusters.

Model-Based Clustering

- In this type of clustering technique, the data observed arises from a distribution consisting of a mixture of two or more cluster components.
- Furthermore, each component cluster has a density function having an associated probability or weight in this mixture.

Fuzzy Clustering

- In this type of clustering, the data points can belong to more than one cluster.
- Each component present in the cluster has a membership coefficient that corresponds to a degree of being present in that cluster.
- Fuzzy Clustering method is also known as a soft method of clustering.

Previous Next

PREVIOUS ACTIVITY

◄ Cross Validation: Python code

Jump to...

NEXT ACTIVITY

Feature Importance & Correlation Matrix ►

Stay in touch

Contact Us

- http://nielit.gov.in/gorakhpur/
- abhinav@nielit.gov.in or ajay.verma@nielit.gov.in