Certificate Course in Machine Learning using Python [6 Weeks]

Dashboard   My courses   Certificate Course in Machine Learning using Python [6 Weeks]   Day 26

Clustering Problem

Clustering Problem
Introduction

Have you come across a situation when a Chief Marketing Officer of a company tells you – "Help me understand our customers better so that we can market our products to them in a better manner!"

Or

Election department has data of election and wants to analyse its data.

Or

ICC has data of many matches and wants to analyse its data.

- In all these case analysts was completely clueless what to do!
- As we have studied supervised machine learning and we know that what to find out.
- But I had no clue what to do in this case or we don't know what is the target.
- This is usually the first reaction when you come across an unsupervised learning problem for the first time!
- You are not looking for specific insights for a phenomena, but what you are looking for are structures with in data without them being tied down to a specific outcome.

Clustering

- The method of identifying similar groups of data in a dataset is called clustering. It is one of the most popular techniques in data science.
- Entities in each group are comparatively more similar to entities of that group than those of the other groups.
- Clustering can be easily understand by the given picture-
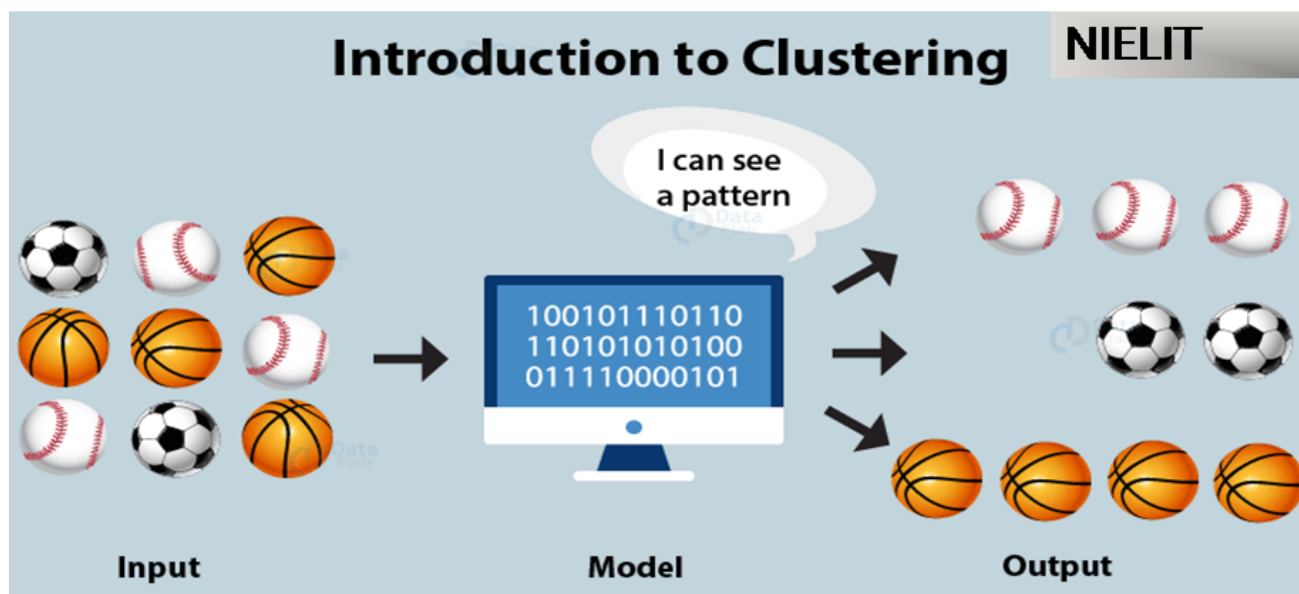
- The basic principle behind cluster is the assignment of a given set of observations into subgroups or clusters such that observations present in the same cluster possess a degree of similarity.
- It is the implementation of the human cognitive ability to discern objects based on their nature.

Example 1:

- when you go out for grocery shopping, you easily distinguish between apples and oranges in a given set containing both of them.
- You distinguish these two objects based on their color, texture and other sensory information that is processed by your brain.
- Clustering is an emulation of this process so that machines are able to distinguish between different objects.

Example 2:

- Let's understand with another example. Suppose, you are the head of a rental store and wish to understand preferences of your costumers to scale up your business.
- Is it possible for you to look at details of each costumer and devise a unique business strategy for each one of them?
- Definitely not. But, what you can do is to cluster all of your costumers into say 5 groups based on their purchasing habits and use a separate strategy for costumers in each of these 5 groups. And this is what we call clustering.

Summary

- It is a method of unsupervised learning since there is no external label attached to the object.
- The machine has to learn the features and patterns all by itself without any given input-output mapping.
- The algorithm is able to extract inferences from the nature of data objects and then create distinct classes to group them appropriately.

Why do we use clustering in ML?

- In basic terms, the objective of clustering is to find different groups within the elements in the data. To do so, clustering algorithms find the structure in the data so that elements of the same cluster (or group) are more similar to each other than to those from different clusters.

- Clustering is an important technique as it performs the determination of the intrinsic grouping among the unlabeled dataset.

- In clustering, there are no standard criteria. All of it depends on the user and the suitable criteria that satisfy their needs and requirements.

- For example, to find the homogeneous groups, one can find the representatives through data reduction and describe their suitable properties.

- One can also find unusual data objects for outlier detection. The algorithm then makes the assumption that constitutes what similarity of points makes valid assumptions.

Applications of clustering

Clustering has a large no. of applications spread across various domains. Some of the most popular applications of clustering are:

- **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.

- **City Planning** : It is used to make groups of houses and to study their values based on their geographical locations and other factors present.

- **Earthquake studies** : By learning the earthquake affected areas we can determine the dangerous zones.

Other applications-

- Recommendation engines

- Social network analysis

- Search result grouping

- Medical imaging

- Image segmentation

- Anomaly detection

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