SNS COLLEGE OF TECHNOLOGY

AN AUTONOMOUS INSTITUTION

Approved by AICTE New Delhi & Affiliated to Anna University Chennai
Accredited by NBA & Accredited by NAAC with "A+" Grade, Recognized by UGC

COIMBATORE

DEPARTMENT OF CIVIL ENGINEERING

MACHINE LEARNING FOR CIVIL ENGINEERS

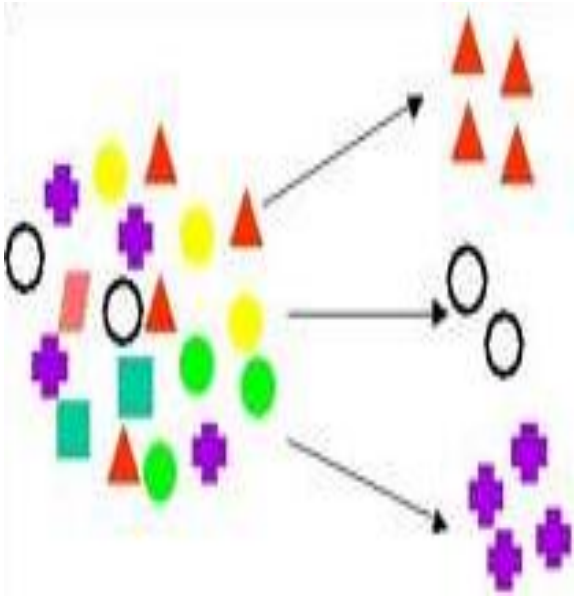
II YEAR / IV SEMESTER

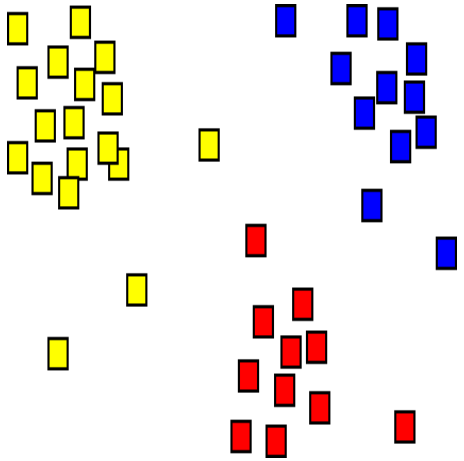
Unit 3 : Unsupervised Learning

Topic 1: Clustering

Introduction

- Defined as extracting the information from the huge set of data.
- Extracting set of patterns from the data set





Clustering

- Clustering means grouping the objects based on the information found in the data describing the objects or their relationships.
- The goal is that the objects in a group will be similar (or related) to one other and different from (or unrelated to) the objects in other groups.
- grouped according to logical relationships or consumer preferences.
- unsupervised learning no target field,
- bottom-up approach.
- originated in anthropology by Driver and Kroeber in 1932 and introduced to psychology by Zubin in 1938 and Robert
- tryon in 1939 and famously used by Cattell beginning in 1943
- for trait theory classification in personality psychology.

Why Clustering?

- 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.
- Scalability - We need highly scalable clustering algorithms to deal with large databases.
- Ability to deal with different kind of attributes - Algorithms should be capable to be applied on any kind of data such as interval based (numerical) data, categorical, binary data.
- Discovery of clusters with attribute shape - The clustering algorithm should be capable of detect cluster of arbitrary shape. It should not be bounded to only distance measures that tend to find spherical cluster of small size.

Working Definitions of Clustering?

- Well separated cluster definition
- Centre based cluster definition
- Contiguous cluster definition(Nearest neighbour or transitive clustering)
- Similarity based cluster definition
- Density based cluster definition

Methods in Clustering

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Partitioning Method

- **Suppose we are given a database of n objects, the partitioning method construct k partition of data. Each partition will represents a cluster and k≤n. It means that it will classify the data into k groups,**
  - Each group contain at least one object.
  - Each object must belong to exactly one group.
- **For a given number of partitions (say k), the partitioning method will create an initial partitioning.**
- **Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.**

Hierarchical Method

- **2 types of HM's**
  - Agglomerative Approach
  - Divisive Approach
- **Agglomerative Approach**
  - bottom-up approach.
  - we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.
- **Divisive Approach**
  - top-down approach.
  - we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds.
- **DisadvantageThis method is rigid i.e. once merge or split is done, It can never be undone.**

$$O(x1\,x2\,\ldots\,xn) = \sigma(WX)$$

where: $\sigma(WX) = 1/1 + e^{-wx}$

Back propagation Algorithm

The back propagation learning algorithm can be divided into two phases:

Phase 1: Propagation

- Forward propagation of a training pattern's input through the neural network in order to generate the propagation's output activations.
- Backward propagation of the propagation's output activations through the neural network using the training pattern target in order to generate the deltas (the difference between the input and output values) of all output and hidden neurons.

Phase 2: Weight update

- Multiply its output delta and input activation to get the gradient of the weight.
- Subtract a ratio (percentage) of the gradient from the weight.

Density Based Method

- This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold i.e. for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

Grid based Method

- In this the objects together from a grid. The object space is quantized into finite number of cells that form a grid structure.
- AdvantageThe major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

Model Based Method

- In this method a model is hypothesize for each cluster and find the best fit of data to the given model. This method locate the clusters by clustering the density function. This reflects spatial distribution of the data points.
- This method also serve a way of automatically determining number of clusters based on standard statistics , taking outlier or noise into account. It therefore yield robust clustering methods.

Constraint Based Method

In this method the clustering is performed by incorporation of user or application oriented constraints. The constraint refers to the user expectation or the properties of desired clustering results. The constraint give us the interactive way of communication with the clustering process. The constraint can be specified by the user or the application requirement.

Applications

- **Pattern Recognition**
- **Spatial Data Analysis:**
- **Image Processing**
- **Economic Science (especially market research)**
- **Crime analysis**
- **Bio informatics**
- **Medical Imaging**
- **Robotics**
- **Climatology**

Thank You!!