Clustering

Methods and Application

What is Clustering ?

The task of grouping data points based on their similarity with each other is called Clustering or Cluster Analysis. This method is defined under the branch of Unsupervised Learning, which aims at gaining insights from unlabelled data points, that is, unlike supervised learning we don’t have a target variable.

Clustering aims at forming groups of homogeneous data points from a heterogeneous dataset. It evaluates the similarity based on a metric like Euclidean distance, Cosine similarity, Manhattan distance, etc.

Types of Clustering Algorithms

Clustering is the process of determining how related the objects are based on a metric called the similarity measure.

Similarity metrics are easier to locate in smaller sets of features. It gets harder to create similarity measures as the number of features increases. Depending on the type of clustering algorithm being utilized in data mining, several techniques are employed to group the data from the datasets.

What is the best clustering method?

The top 10 clustering algorithms are:

- K-means Clustering
- Hierarchical Clustering
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)
- Gaussian Mixture Models (GMM)
- Agglomerative Clustering
- Spectral Clustering
- Mean Shift Clustering
- Affinity Propagation
- OPTICS (Ordering Points To Identify the Clustering Structure)
- Birch (Balanced Iterative Reducing and Clustering using Hierarchies)

In the above Top 10 Algorithms, We discuss few of them.

Affinity Propagation Clustering Algorithm

- Initialize each data point as a potential exemplar
- Measure similarity between data points Euclidean distance, similarity measures etc.
- Iteratively update responsibiliry and availability metrices.
- Excemplar selection
- Cluster asssignments
- Convergence check and finalization

Advantages, Disadvanges & Applications of Affinity Propagation

Applications

Social Network Analysis
Marketing Fields(Analysing)
Medical Image fields
Speech Recognition

Advantages

- Automatic cluster determination
- Discovery of exemplars
- Handles non linear relationships

Disadvantages

- Computationally expensive
- In High dimensional datas Its not suitable