**Machine Learning**

**M.Janarthanan**

**15MIS1060**

**Clustering Techniques**

**Digital Assingnment - 3**

**Introduction to Clustering**

It is basically a type of [unsupervised learning method](https://www.geeksforgeeks.org/supervised-unsupervised-learning/) . An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labelled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.  
Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

There are many types of clustering algorithms.

Many algorithms use similarity or distance measures between examples in the feature space in an effort to discover dense regions of observations. As such, it is often good practice to scale data prior to using clustering algorithms

Some clustering algorithms require you to specify or guess at the number of clusters to discover in the data, whereas others require the specification of some minimum distance between observations in which examples may be considered “close” or “connected.”

As such, cluster analysis is an iterative process where subjective evaluation of the identified clusters is fed back into changes to algorithm configuration until a desired or appropriate result is achieved.

The scikit-learn library provides a suite of different clustering algorithms to choose from.

A list of 10 of the more popular algorithms is as follows:

* Affinity Propagation
* Agglomerative Clustering
* BIRCH
* DBSCAN
* K-Means
* Mini-Batch K-Means
* Mean Shift
* OPTICS
* Spectral Clustering
* Mixture of Gaussians

**Affinity Propagation**

Affinity Propagation involves finding a set of exemplars that best summarize the data

**Agglomerative Clustering**

Agglomerative clustering involves merging examples until the desired number of clusters is achieved

**BIRCH**

BIRCH Clustering (BIRCH is short for Balanced Iterative Reducing and Clustering using  
Hierarchies) involves constructing a tree structure from which cluster centroids are extracted.

**DBSCAN**

DBSCAN Clustering (where DBSCAN is short for Density-Based Spatial Clustering of Applications with Noise) involves finding high-density areas in the domain and expanding those areas of the feature space around them as clusters

**K-Means**

[K-Means Clustering](https://en.wikipedia.org/wiki/K-means_clustering) may be the most widely known clustering algorithm and involves assigning examples to clusters in an effort to minimize the variance within each cluster

**Mini-Batch K-Means**

Mini-Batch K-Means is a modified version of k-means that makes updates to the cluster centroids using mini-batches of samples rather than the entire dataset, which can make it faster for large datasets, and perhaps more robust to statistical noise

**Mean Shift**

Mean shift clustering involves finding and adapting centroids based on the density of examples in the feature space.

**OPTICS**

OPTICS clustering (where OPTICS is short for Ordering Points to Identify the Clustering Structure) is a modified version of DBSCAN.

**Spectral Clustering**

Spectral Clustering is a general class of clustering methods, drawn from [linear algebra](https://machinelearningmastery.com/linear-algebra-machine-learning-7-day-mini-course/)

**Gaussian Mixture Model**

A Gaussian mixture model summarizes a multivariate probability density function with a mixture of Gaussian probability distributions as its name suggest.

**Why Clustering ?**  
Clustering is very much important as it determines the intrinsic grouping among the un labeled data present. There are no criteria for a good clustering. It depends on the user, what is the criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions which constitute the similarity of points and each assumption make different and equally valid clusters.

**Clustering Methods :**

* **Density-Based Methods :** These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and ability to merge two clusters. Example DBSCAN (Density-Based Spatial Clustering of Applications with Noise) , OPTICS (Ordering Points to Identify Clustering Structure) etc.
* **Hierarchical Based Methods :** The clusters formed in this method forms a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category
  + **Agglomerative** (bottom up approach)
  + **Divisive** (top down approach)

examples CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies) etc.

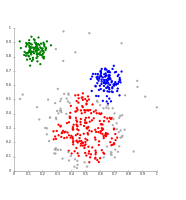
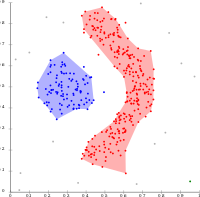
* **Partitioning Methods :** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example K-means, CLARANS (Clustering Large Applications based upon Randomized Search) etc.
* **Grid-based Methods :** In this method the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operation done on these grids are fast and independent of the number of data objects example STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest) etc.

**Applications of Clustering in different fields**

* **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.

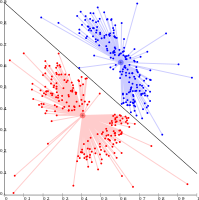
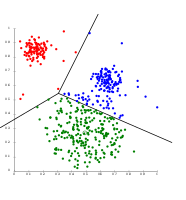
### **Density-based Clustering**

Density-based clustering connects areas of high example density into clusters. This allows for arbitrary-shaped distributions as long as dense areas can be connected. These algorithms have difficulty with data of varying densities and high dimensions. Further, by design, these algorithms do not assign outliers to clusters.



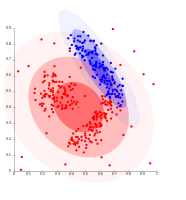
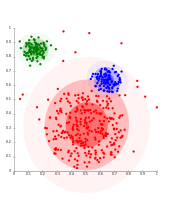
### **Centroid-based Clustering**

**Centroid-based clustering** organizes the data into non-hierarchical clusters, in contrast to hierarchical clustering defined below. k-means is the most widely-used centroid-based clustering algorithm. Centroid-based algorithms are efficient but sensitive to initial conditions and outliers. This course focuses on k-means because it is an efficient, effective, and simple clustering algorithm.



### **Distribution-based Clustering**

This clustering approach assumes data is composed of distributions, such as **[Gaussian distributions](https://wikipedia.org/wiki/Normal_distribution)**. the distribution-based algorithm clusters data into three Gaussian distributions. As distance from the distribution's center increases, the probability that a point belongs to the distribution decreases. The bands show that decrease in probability. When you do not know the type of distribution in your data, you should use a different algorithm.



### **Hierarchical Clustering**

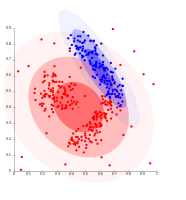
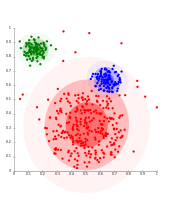
**Hierarchical clustering** creates a tree of clusters. Hierarchical clustering, not surprisingly, is well suited to hierarchical data, such as taxonomies. See [Comparison of 61 Sequenced Escherichia coli Genomes](https://www.researchgate.net/figure/Pan-genome-clustering-of-E-coli-black-and-related-species-colored-based-on-the_fig1_45152238) by Oksana Lukjancenko, Trudy Wassenaar & Dave Ussery for an example. In addition, another advantage is that any number of clusters can be chosen by cutting the tree at the right level.

### **Distribution-based clustering**

The clustering model most closely related to statistics is based on [distribution models](https://en.wikipedia.org/wiki/Probability_distribution" \o "Probability distribution). Clusters can then easily be defined as objects belonging most likely to the same distribution. A convenient property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution.

While the theoretical foundation of these methods is excellent, they suffer from one key problem known as [overfitting](https://en.wikipedia.org/wiki/Overfitting" \o "Overfitting), unless constraints are put on the model complexity. A more complex model will usually be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult.

One prominent method is known as Gaussian mixture models (using the [expectation-maximization algorithm](https://en.wikipedia.org/wiki/Expectation-maximization_algorithm" \o "Expectation-maximization algorithm)). Here, the data set is usually modeled with a fixed (to avoid overfitting) number of [Gaussian distributions](https://en.wikipedia.org/wiki/Gaussian_distribution" \o "Gaussian distribution) that are initialized randomly and whose parameters are iteratively optimized to better fit the data set. This will converge to a [local optimum](https://en.wikipedia.org/wiki/Local_optimum" \o "Local optimum), so multiple runs may produce different results. In order to obtain a hard clustering, objects are often then assigned to the Gaussian distribution they most likely belong to; for soft clustering, this is not necessary.



Connectivity-based clustering (hierarchical clustering)

Main article: [Hierarchical clustering](https://en.wikipedia.org/wiki/Hierarchical_clustering" \o "Hierarchical clustering)

Connectivity-based clustering, also known as [hierarchical clustering](https://en.wikipedia.org/wiki/Hierarchical_clustering" \o "Hierarchical clustering), is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a [dendrogram](https://en.wikipedia.org/wiki/Dendrogram" \o "Dendrogram), which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

Connectivity-based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of [distance functions](https://en.wikipedia.org/wiki/Distance_function" \o "Distance function), the user also needs to decide on the linkage criterion (since a cluster consists of multiple objects, there are multiple candidates to compute the distance) to use. Popular choices are known as [single-linkage clustering](https://en.wikipedia.org/wiki/Single-linkage_clustering" \o "Single-linkage clustering) (the minimum of object distances), [complete linkage clustering](https://en.wikipedia.org/wiki/Complete_linkage_clustering" \o "Complete linkage clustering) (the maximum of object distances), and [UPGMA](https://en.wikipedia.org/wiki/UPGMA" \o "UPGMA) or [WPGMA](https://en.wikipedia.org/wiki/WPGMA" \o "WPGMA) ("Unweighted or Weighted Pair Group Method with Arithmetic Mean", also known as average linkage clustering). Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions).

