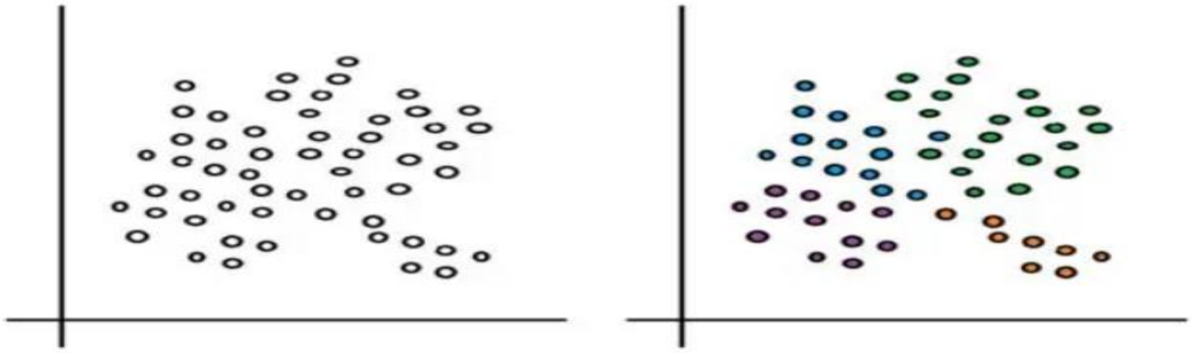
Clustering Methods and Applications

What is Clustering?

Many things around us can be categorized as “this and that” or to be less vague and more specific, we have groupings that could be binary or groups that can be more than two, like a type of pizza base or type of car that you might want to purchase. The choices are always clear – predefined groups and the process predicting that is an important process in the Data Science stack called Classification.

But what if we bring into play a quest where we **don’t have pre-defined choices** initially, rather, we derive those choices! **Choices that are based out of hidden patterns, underlying similarities between the constituent variables, salient features from the data** etc. This process is known as Clustering in Machine Learning or Cluster Analysis, where we group the data together into an unknown number of groups and later use that information for further business processes.

So, to put it in simple words, in machine learning clustering is the process by which we create groups in a data, like customers, products, employees, text documents, in such a way that **objects falling into one group exhibit many similar properties with each other** and are different from objects that fall in the other groups that got created during the process.



Clustering algorithms take the data and using some sort of similarity metrics, they form these groups – later these groups can be used in various business processes like information retrieval, pattern recognition, image processing, data compression, bioinformatics etc. In the Machine Learning process for Clustering, as mentioned above, a distance-based similarity metric plays a pivotal role in deciding the clustering.

Types of Clustering Methods

As we made a point earlier that for a successful grouping, we need to attain two major goals: one, a **similarity between one data point with another** and two, a **distinction of those similar data points with others which most certainly, heuristically differ from those points**. The basis of such divisions begins with our ability to scale large datasets and that's a major beginning point for us. Once we are through it, we are presented with a challenge that our data contains different kinds of attributes – categorical, continuous data, etc., and we should be able to deal with them. Now, we know that our data these days is not limited in terms of dimensions, we have data that is multi-dimensional in nature. The clustering algorithm that we intend to use should successfully cross this hurdle as well.

The clusters that we need, should not only be able to distinguish data points but also they should be inclusive. Sure, a distance metric helps a lot but the cluster shape is often limited to being a geometric shape and many important data points get excluded. This problem too needs to be taken care of.

In our progress, we notice that our data is highly **"noisy"** in nature. Many unwanted features have been residing in the data which makes it rather Herculean task to bring about any similarity between the data points – leading to the creation of improper groups. As we move towards the end of the line, we are faced with a challenge of business interpretation. The outputs from the clustering algorithm should be understandable and should fit the business criteria and address the business problem correctly.

To address the problem points above – **scalability, attributes, dimensional, boundary shape, noise, and interpretation** – we have various types of clustering methods that solve one or many of these problems and of course, many statistical and machine learning clustering algorithms that implement the methodology.

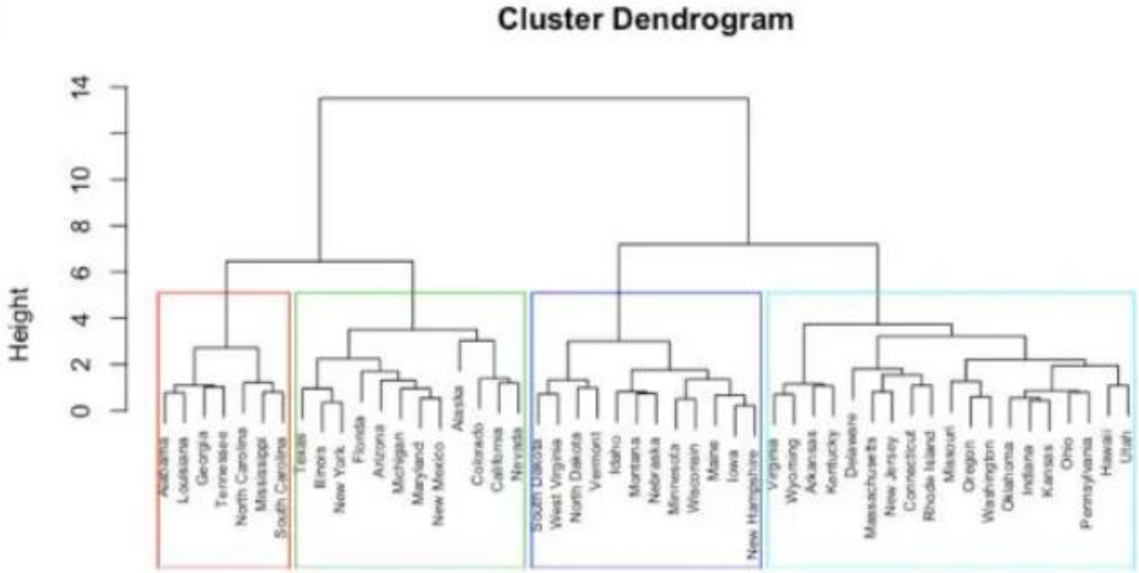
The various types of clustering are:

- Connectivity-based Clustering (Hierarchical clustering)
- Density-based Clustering (Model-based methods)
- Centroids-based Clustering (Partitioning methods)
- Distribution-based Clustering
- Fuzzy Clustering

- ➢ Constraint-based (Supervised Clustering)

Connectivity-Based Clustering (Hierarchical Clustering)

Hierarchical Clustering is a method of unsupervised machine learning clustering where it begins with a pre-defined top to bottom hierarchy of clusters. It then proceeds to perform a decomposition of the data objects based on this hierarchy, hence obtaining the clusters. This method follows two approaches based on the direction of progress, i.e., whether it is the top-down or bottom-up flow of creating clusters. These are Divisive Approach and the Agglomerative Approach respectively.

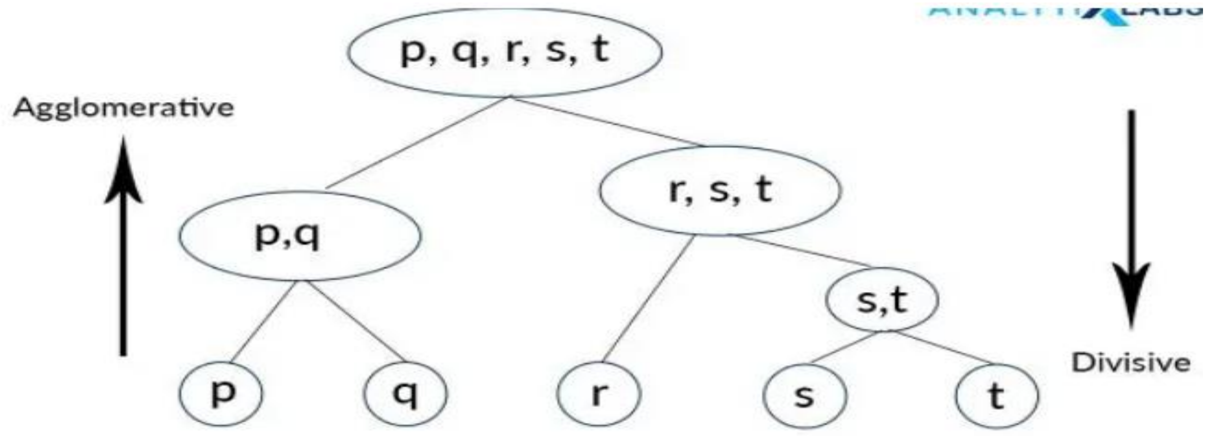


Divisive Approach

This approach of hierarchical clustering follows a top-down approach where we consider that all the data points belong to one large cluster and try to divide the data into smaller groups based on a termination logic or, a point beyond which there will be no further division of data points. This termination logic can be based on the minimum sum of squares of error inside a cluster or for categorical data, the metric can be the GINI coefficient inside a cluster.

Hence, iteratively, we are splitting the data which was once grouped as a single large cluster, to “n” number of smaller clusters in which the data points now belong to.

It must be taken into account that this algorithm is highly “rigid” when splitting the clusters – meaning, one a clustering is done inside a loop, there is no way that the task can be undone.



Agglomerative Approach

Agglomerative is quite the contrary to Divisive, where all the “N” data points are considered to be a single member of “N” clusters that the data is comprised into. We iteratively combine these numerous “N” clusters to fewer number of clusters, let’s say “k” clusters and hence assign the data points to each of these clusters accordingly. This approach is a bottom-up one, and also uses a termination logic in combining the clusters. This logic can be a number based criterion (no more clusters beyond this point) or a distance criterion (clusters should not be too far apart to be merged) or variance criterion (increase in the variance of the cluster being merged should not exceed a threshold, Ward Method)

The Hierarchical clustering Technique can be visualized using a Dendrogram.

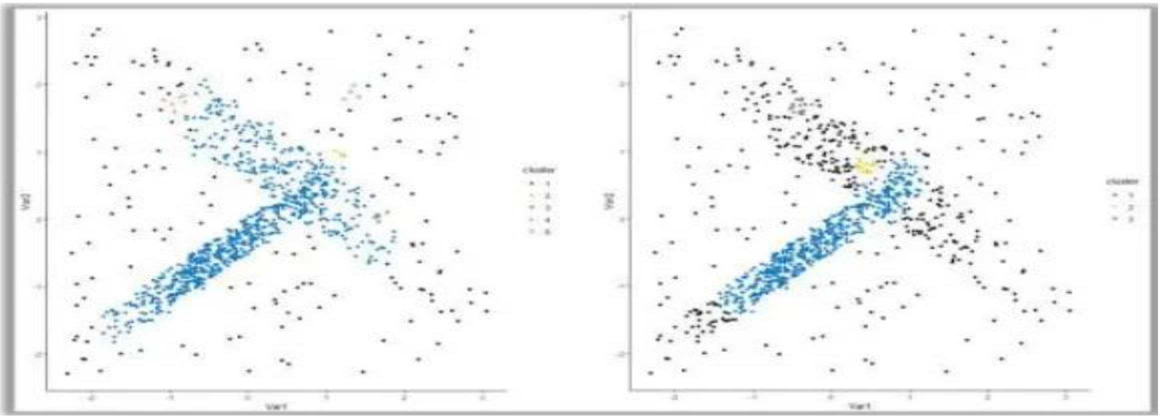
A Dendrogram is a tree-like diagram that records the sequences of merges or splits.

2. Density-based Clustering (Model-based Methods)

If one looks into the previous two methods that we discussed, one would observe that both hierarchical and centroid based algorithms are dependent on a distance (similarity/proximity) metric. The very definition of a cluster is based on this metric. Density-based clustering methods take density into consideration instead of distances. Clusters are considered as the densest region in a data space, which is separated by regions of lower object density and it is defined as a maximal-set of connected points.

When performing most of the clustering, we take two major assumptions, one, the data is devoid of any noise and two, the shape of the cluster so formed is purely geometrical (circular or elliptical). The fact is, data always has some extent of inconsistency (noise) which cannot be ignored. Added to that, we must not limit ourselves to a fixed attribute shape, it is desirable to have arbitrary shapes so as to not to ignore any data points. These are the areas where density based algorithms have proven their worth!

Density-based algorithms can get us clusters with arbitrary shapes, clusters without any limitation in cluster sizes, clusters that contain the maximum level of homogeneity by ensuring the same levels of density within it, and also these clusters are inclusive of outliers or the noisy data.



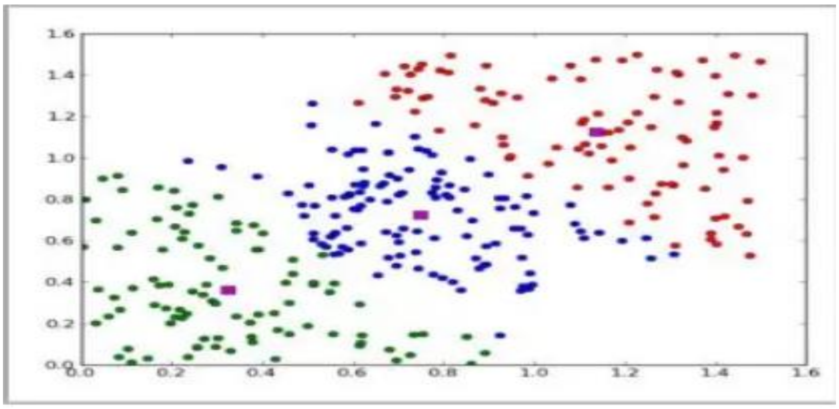
3. Centroid Based Clustering

Centroid based clustering is considered as one of the most simplest clustering algorithms, yet the most effective way of creating clusters and assigning data points to it. The intuition behind centroid based clustering is that a cluster is characterized and represented by a central vector and data points that are in close proximity to these vectors are assigned to the respective clusters.

These groups of clustering methods iteratively measure the distance between the clusters and the characteristic centroids using various distance metrics. These are either of Euclidian distance, Manhattan Distance or Minkowski Distance.

The major setback here is that we should either intuitively or scientifically (Elbow Method) define the number of clusters, “k”, to begin the iteration of any clustering machine learning algorithm to start assigning the data points.

Despite the flaws, Centroid based clustering has proven it’s worth over Hierarchical clustering when working with large datasets. Also, owing to its simplicity in implementation and also interpretation, these algorithms have wide application areas viz., market segmentation, customer segmentation, text topic retrieval, image segmentation etc.


1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6