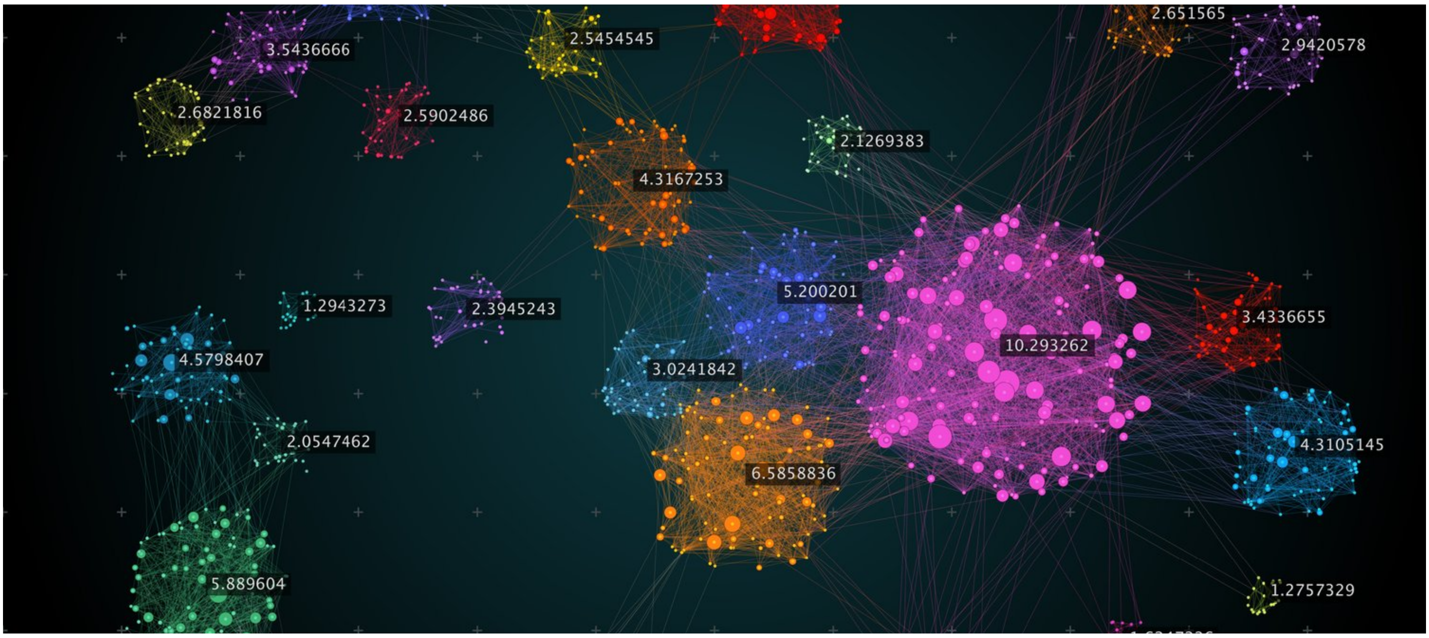
10 INCREDIBLY USEFUL CLUSTERING ALGORITHMS YOU NEED TO KNOW

GAVITA REGUNATH (/BLOG?AUTHOR=60CC50CFDE1E610ACE5F4E1D) · JUNE 14, 2022

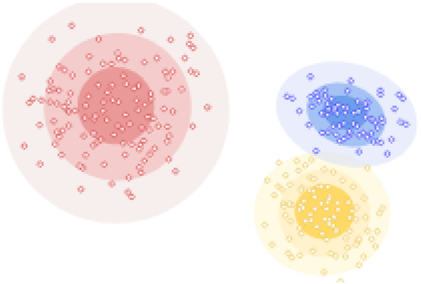
In the previous article (https://www.advancinganalytics.co.uk/blog/2022/6/13/get-started-with-clustering-the-easy-way), we explained how clustering, an unsupervised machine learning method, is different from the supervised machine learning method. We also described the concept of clustering analysis using a simple example and showed how it could be used throughout many businesses to solve various problems.

In this article, we will delve a bit deeper into clustering algorithms.



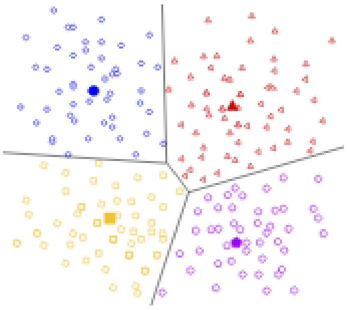
DIFFERENT TYPES OF CLUSTERING ALGORITHMS

There are many clustering algorithms. In fact, there are more than 100 clustering algorithms that have been published so far. However, despite the various types of clustering algorithms, they can generally be categorised into four methods. Let's look at these briefly:



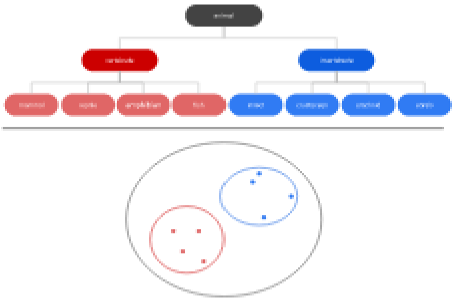
1.Distribution models - Clusters in this model belong to a distribution. Data points are grouped based on the probability of belonging to either a normal or a gaussian distribution. The expectation-maximisation algorithm, which uses multivariate normal distributions, is one of the popular examples of this algorithm.

Example of distribution-based clustering.



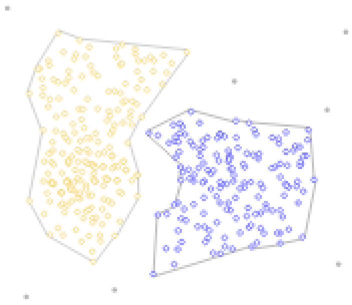
Example of centroid-based clustering

2. Centroid models – This is an iterative algorithm in which data is organised into clusters based on how close data points are to the centre of clusters also known as centroid. An example of centroid models is the K-means algorithm.



Example of a hierarchical tree clustering.

3. Connectivity models – This is similar to the centroid model and is also known as Hierarchical clustering. It is a cluster analysis method that seeks to build a hierarchy of clusters. An example of a connectivity model is the hierarchical algorithm.



Example of density-based clustering.

4. Density models - Clusters are defined by areas of concentrated density. It searches areas of dense data points and assigns those areas to the same clusters. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and OPTICS are two popular density-based clustering models.

Besides the above-mentioned clustering algorithms, below is a list of some of the top clustering algorithms that are often used to solve machine learning problems.

TOP 10 CLUSTERING ALGORITHMS (IN ALPHABETICAL ORDER):

1. **Affinity Propagation**
2. **Agglomerative Hierarchical Clustering**
3. **BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)**
4. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**
5. **Gaussian Mixture Models (GMM)**
6. **K-Means**
7. **Mean Shift Clustering**
8. **Mini-Batch K-Means**
9. **OPTICS**
10. **Spectral Clustering**

Affinity Propagation: Affinity Propagation was first published in 2007 by Brendan Frey and Delbert Dueck in the renowned Science journal. It considers all data points as input measures of similarity between pairs of data points and simultaneously considers them as potential exemplars. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges.

Agglomerative Hierarchical Clustering: This clustering technique uses a hierarchical "bottom-up" approach. This implies that the algorithm begins with all data points as clusters and begins merging them depending on the distance between clusters. This will continue until we establish one large cluster.

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies): This technique is very useful when clustering large datasets as it begins by first generating a more compact summary that retains as much distribution information as possible and then clustering the data summary instead of the original large dataset.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): DBSCAN is a really well known density-based clustering algorithm. It determines clusters based on how dense regions are. It is able to find irregular-shaped clusters and outliers really well.

OPTICS (Ordering Points to Identify the Clustering Structure): This is also a density-based clustering algorithm. It is very similar to the DBSCAN described above, but it overcomes one of DBSCAN's limitations, which is the problem of detecting meaningful clusters in data of varying density.

K-Means: This algorithm is one of the most popular and commonly used clustering technique. It works by assigning data points to clusters based on the shortest distance to the centroids or centre of the cluster. This algorithm's main goal is to minimise the sum of distances between data points and their respective clusters.

Mini-Batch K-Means: This is a k-means version in which cluster centroids are updated in small batches rather than the entire dataset. When working with a large dataset, the mini-batch k-means technique can be used to minimise computing time.

Mean Shift Clustering: Mean shift clustering algorithm is a centroid-based algorithm that works by shifting data points towards centroids to be the mean of other points in the feature space.

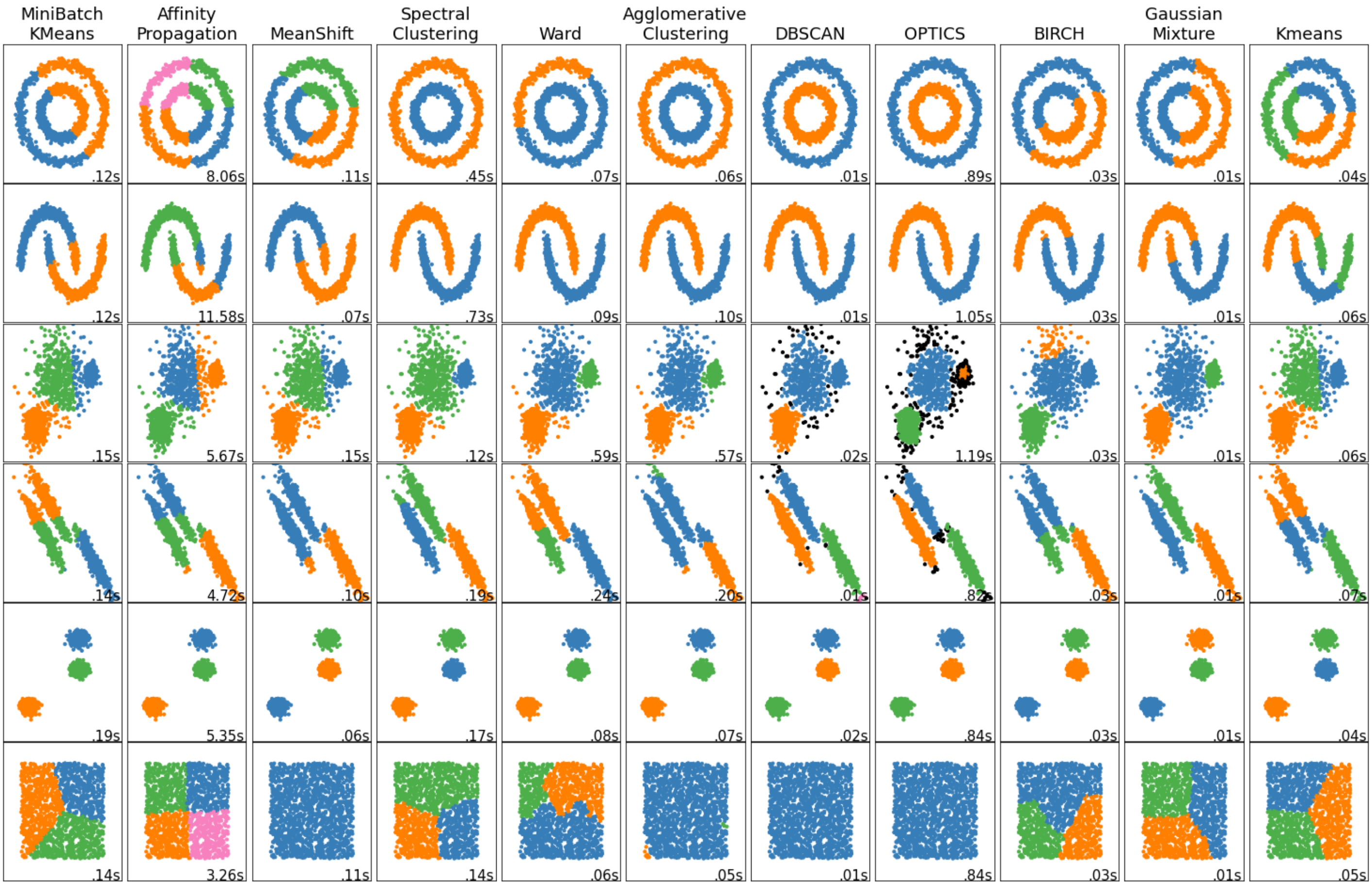
Spectral Clustering: Spectral clustering is a graph-based algorithm where the approach is used to identify communities of nodes based on the edges. Because of its ease of implementation and promising performance, spectral clustering has grown in popularity.

Gaussian Mixture Models (GMM): The Gaussian mixture models is an extension of the k-means clustering algorithm. It is based on the idea that each cluster may be assigned to a different Gaussian distribution. GMM uses soft-assignment of data points to clusters (i.e. probabilistic and therefore better) when contrasting with the K-means approach of hard-assignments of data points to clusters.

Each method described above approaches the problem of detecting patterns in datasets very differently. It would be interesting to demonstrate how these clustering algorithms work on different datasets. To do this, we can leverage the scikit-learn (https://scikit-learn.org/stable/modules/clustering.html) machine learning library. Below is a visual comparison of the top 10 clustering algorithms discussed in this article and how they work on different types of datasets, such as:

1. **Circles** - This dataset consists of two circles, one circumscribed by the other.
2. **Moons** - This dataset has two interleaving half circles.
3. **Varied variance blobs** – This dataset consist of blobs that have different variances.
4. **Anisotropically distributed blobs** - These blobs have unequal widths and lengths.
5. **Regular blobs** - Just three regular blobs
6. **Homogenous data** – This is an example of a 'null' situation for clustering.

From the figure below, you can quickly see the effectiveness of the clustering algorithm depends on the underlying dataset type. For example, if you are working in a dataset that looks similar to the circles dataset, then DBSCAN, OPTICS or Agglomerative clustering will be more effective than the rest of the algorithms.



WHAT IS THE BEST CLUSTERING ALGORITHM?

There is no such thing as the "best" clustering method; nonetheless, there are several clustering algorithms that are better suited to a particular dataset having a specific distribution.

WRAPPING UP

In this article, we highlighted how clustering algorithms can be grouped into four categories. In addition to this, we also included a list of the top ten clustering algorithms that are often used by data scientists and machine learning practitioners.