Introduction to Clustering

Clustering:

It is essentially a form of unsupervised learning. An unsupervised learning method is a process by which references are drawn from input data sets without marked answers. Generally speaking, it is utilised as a procedure to uncover significant structures, explaining processes, generative characteristics, and groups present in a number of cases.

Clustering is the task of dividing population or data points into a number of groups in order to make data points more similar to other databases in the same group in the same group and different from data points in others. It's basically a collection of objects based on their similarities and differences.

Hierarchical clustering:

While there is no need to specify the number of clusters to be produced, hierarchical clustering can be utilised as an alternative to partitioned clustering. The dataset is separated into clusters in this technique to form a tree-like structure known as a dendrogram. By pruning the tree at the appropriate level, the observations or any number of clusters can be selected. The Agglomerative Hierarchical algorithm is the most common example of this strategy.

Non-hierarchical clustering:

The K Means algorithm is an iterative procedure that attempts to partition the dataset into K unique non-overlapping subgroups (clusters), with each data point belonging to just one group. It attempts to keep intra-cluster data points as comparable as possible while keeping clusters as diverse (far) as possible. It distributes data points to clusters in such a way that the total of the squared distances between the data points and the cluster's centroid (the arithmetic mean of all the data points in that cluster) is as small as possible. The lower the variation within clusters, the more homogenous (similar) the data points inside the same cluster are.

Use cases of clustering:

- Identifying Fake News
- Spam filter
- Marketing and Sales

- Classifying network traffic
- Identifying fraudulent or criminal activity
- Document analysis
- Fantasy Football and Sports

How does the k-means clustering work?

K-means clustering is a method for grouping data points into numerous comparable groups, or "clusters," which are distinguished by their midpoints, which we refer to as centroids.

This is how it works:

1. Enter K, the number of clusters you want to find. Let's go with K=3.
2. Generate K (three) new points at random on your chart. These are the centroids of the first clusters.
3. Calculate the distance between each data point and each centroid, then assign each data point to the centroid closest to it and the cluster to which it belongs.
4. Recalculate the cluster's midpoint (centroid).
5. Repeat steps 3–4 to allocate data points to clusters based on the new centroid positions. Stop when one of the following occurs:
    a. The centroids have been stabilised; no data points are redistributed after determining the centroid of a cluster.
    b. The maximum number of iterations has been reached.

How does hierarchical clustering work?

Hierarchical clustering is another method of clustering. Here, clusters are assigned based on hierarchical relationships between data points. There are two key types of hierarchical clustering: agglomerative (bottom-up) and divisive (top-down). Agglomerative is more commonly used as it is mathematically easier to compute, and is the method used by python’s scikit-learn library, so this is the method we’ll explore in detail.

Here’s how it works:

1. Assign each data point to its own cluster, so the number of initial clusters (K) is equal to the number of initial data points (N).
2. Compute distances between all clusters.
3. Merge the two closest clusters.
4. Repeat steps two and three iteratively until all data points are finally merged into one large cluster.

Both K-means and hierarchical clustering are widely used algorithms, yet they serve different purposes. Because it is relatively computationally efficient, K-means works significantly better with larger datasets. Hierarchical clustering, on the other hand, does not perform well with huge datasets due to the quantity of computations required at each step, but it produces better results for smaller datasets and allows for hierarchy interpretation, which is useful if your dataset is hierarchical in nature.

There is no one-size-fits-all answer to practically every machine learning problem, but these two methods each have various use cases, and it is vital to evaluate the nature of your dataset when picking which approach to utilise.