Unsupervised learning

Unsupervised learning is a type of machine learning that learns from test data that has not been labeled, classified, or categorized. Instead of responding to feedback, unsupervised learning identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data.

One of the main advantages of unsupervised learning is its ability to discover previously unknown patterns in data. This is particularly useful when the experts do not know what to look for in the data.

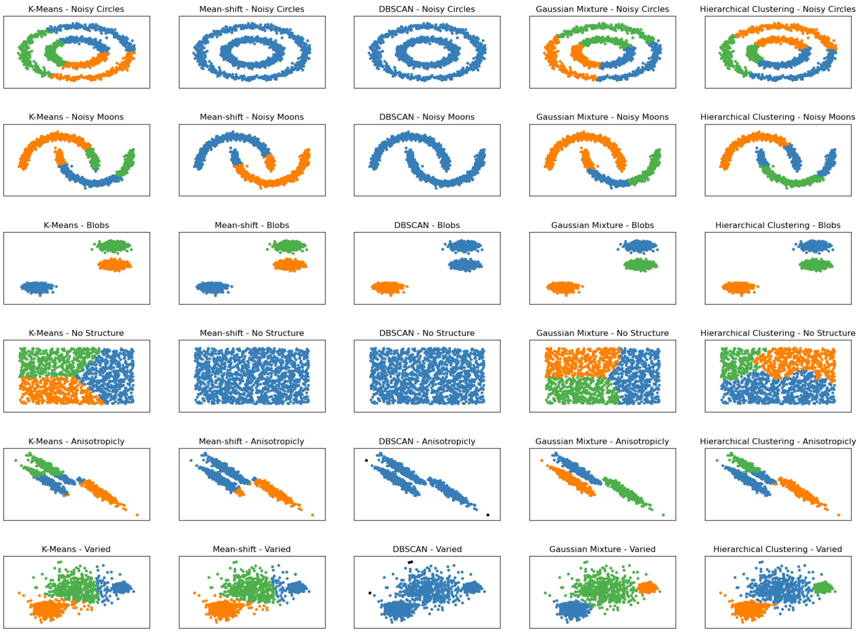
A common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data. The clusters are modeled using a measure of similarity which is defined upon metrics such as Euclidean or probabilistic distance.

Another unsupervised learning method is association rule learning, where the machine identifies common patterns of items in large datasets. For example, a supermarket might use association rule learning to understand the purchasing behavior of its customers.

Dimensionality reduction is another method under unsupervised learning. High-dimensional data can be very challenging to visualize, and more generally, can present a variety of challenges related to overfitting in many classes of statistical models. Dimensionality reduction techniques provide ways to convert high-dimensional data into a form that can be more readily analyzed.

What are the Clustering Algorithms?

According to scikit-learn official documentation, there are 11 different clustering algorithms: K-Means, Affinity propagation, Mean Shift, Special Clustering, Hierarchical Clustering, Agglomerative Clustering, DBScan, Optics, Gaussian Mixture, Birch, Bisecting K-Means. The following graph shows how these five algorithms work on six differently structured datasets.



There are six different datasets shown, all generated by using scikit-learn:

- **Noisy Circles:** This dataset consists of a large circle containing a smaller circle that is not perfectly centered. The data also has random Gaussian noise added to it.
- **Noisy Moons:** This dataset consists of two interleaving half-moon shapes that are not linearly separable. The data also has random Gaussian noise added to it.
- **Blobs:** This dataset consists of randomly generated blobs that are relatively uniform in size and shape. The dataset contains three blobs.
- **No Structure:** This dataset consists of randomly generated data points with no inherent structure or clustering pattern.
- **Anisotropicly Distributed:** This dataset consists of randomly generated data points that are anisotropically distributed. The data points are generated with a specific transformation matrix to make them elongated along certain axes.
- **Varied:** This dataset consists of randomly generated blobs with varied variances. The dataset contains three blobs, each with a different standard deviation.

Seeing the plots and how each algorithm works on them will help us compare how well our algorithms perform on each dataset. This may help you in your data project if your data have the same structure as in these graphs.

Clustering

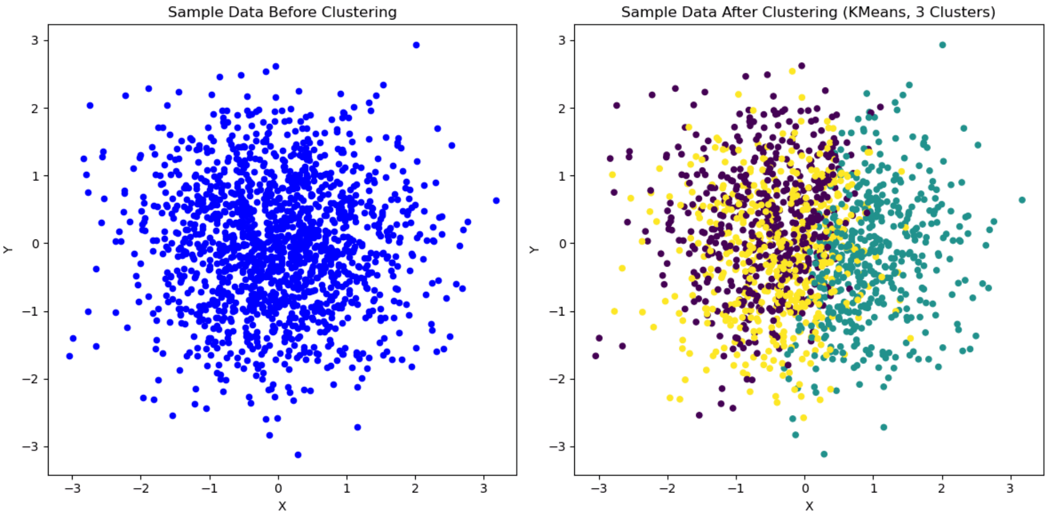
Clustering is the grouping of data points which are similar to each other. It can be a powerful technique for identifying patterns in data. Clustering analysis does not usually require any training and is therefore known as an unsupervised learning technique. Clustering can be applied quickly due to this lack of training.

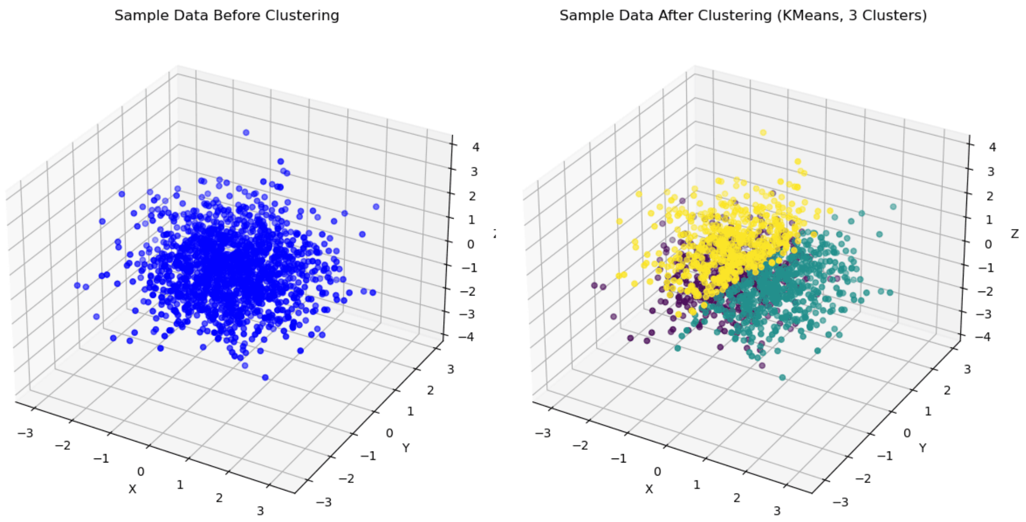
Applications of clustering

- Looking for trends in data
- Reducing the data around a point to just that point (e.g. reducing colour depth in an image)
- Pattern recognition

K-means clustering

The k-means clustering algorithm is a simple clustering algorithm that tries to identify the centre of each cluster. It does this by searching for a point which minimises the distance between the centre and all the points in the cluster. The algorithm needs to be told how many k clusters to look for, but a common technique is to try different numbers of clusters and combine it with other tests to decide on the best combination.





Sample Data Before Clustering

Sample Data After Clustering (KMeans, 3 Clusters)

Limitations of k-means

- Requires number of clusters to be known in advance
- Struggles when clusters have irregular shapes
- Will always produce an answer finding the required number of clusters even if the data isn't clustered (or clustered in that many clusters)
- Requires linear cluster boundaries

Advantages of k-means

- Simple algorithm and fast to compute
- A good choice as the first thing to try when attempting to cluster data
- Suitable for large datasets due to its low memory and computing requirements