Clustering

Clustering is the process of grouping data into classes or clusters.
The grouping is done in such a manner that the objects within the same cluster are very similar to each other but they are very dissimilar to the objects in some other cluster.

Clustering is a form of "learning by observation". It is an unsupervised learning method and does not require a training data set to generate a model. Clustering can lead to the discovery of previously unknown groups within the data.

Clustering Example:

You have a pack of cards. But these cards have Chinese letters/words printed on them. You cannot simply divide them into four groups of Spades, Hearts, Diamonds and Clubs, since you do not recognise the Chinese letters on them. You do not know Chinese. So, how do you put these cards into different groups?

You observe the letters on each card. You try to find the letters that look similar, and put these cards in a group. Once you come across a card that does not look similar to any card you have seen so far, you make a new group for this card. Eventually, you will have made a few groups, on the basis of how similar the letters looked to each other.

This is called Clustering. We do not have any previous knowledge about the letters on the cards. We make groups only on the basis of similarity between the cards. The cards in one group are as similar to each other as possible, while cards in different groups are as dissimilar to each other as possible.

k-Means Clustering

One of the popular examples of clustering is k-Means clustering. The k-Means algorithm is a way to find clusters or groups in your data. Often, you may hypothesize that groups exist in your data, but those groups haven't been identified or labeled ahead of time. The k-Means clustering algorithm is a set of steps that work iteratively to find the groups and label the data. "k" is a variable that represents the number of groups.

How does it work?

The following are the steps in the k-Means alogrithm:

- Select an initial partition of k clusters.
- Assign each object to the cluster with the closest center (mean of all data points).
- Compute the new centers of the clusters (mean of all data points).
- Repeat step 2 and 3 until no object changes cluster.

You'll need to run the k-Means clustering algorithm for a range of k values and compare the results to find the value of k that best represents the number of clusters in your data.

Applications of Clustering

- Segmenting customers into groups with similar buying patterns for targeted marketing campaigns.
- Detecting anomalous behavior, such as unauthorized network intrusions, by identifying patterns of use falling outside the known clusters.
- Simplifying extremely large datasets by grouping features with similar values into a smaller number of homogeneous categories.

Overall, clustering is useful for transforming diverse and varied data to much smaller number of groups. It results in meaningful and actionable data structures that reduce complexity and provide insight into patterns of relationships.

References:

For more details on clustering can refer below:
`https://en.wikipedia.org/wiki/Cluster_analysis`
`https://www.geeksforgeeks.org/different-types-clustering-algorithm/`
`https://en.wikipedia.org/wiki/K-means_clustering`