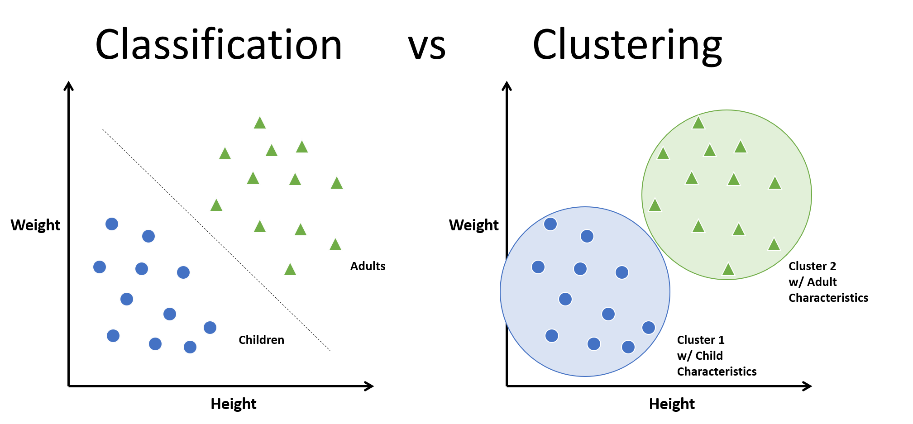
When given a bunch of movies to organize, you might sort them based on similarity measures such as genre, year of release, casting, revenue generated, and so on. While you categorize them based on a measure, it doesn’t mean that a measure used by others is of no good.

Our perception of data is greatly influenced by the clustering measure being used. When you believe that two sci-fi movies belong to the same cluster, someone else might consider two movies released in the same year to be in the same cluster. Because you might have used the genre as a measure to cluster and the other person might have used the year of release as a measure.

Classifications vs Clustering

As humans, in machine learning, a widely used unsupervised algorithm to group unlabeled data points by similarity and distance measures is clustering. If the data points are labeled, grouping is known as classification. Clustering algorithms have their application in many places including anomaly detection, image segmentation, search result grouping, market segmentation, and social network analysis. Clustering is one of the initial steps done in exploratory data analysis to visualize the similarity and to identify the pattern lying hidden in data points. The motive of clustering is to find the similarity within a cluster and the difference between two clusters.

Support Vector Machines, Decision trees, Random forests, Linear classifiers and Neural networks are few classification algorithms whereas K means clustering, Fuzzy analysis clustering, Mean shift, DBSCAN and Spectral are clustering algorithms.



There are two major approaches in clustering. They are:

1. Compactness
2. Connectivity

In compactness, the points are closer to each other and are compact towards the cluster center. Distance is used as a measure to compute closeness. There are different types of distance metrics that are in use. A few of them are Euclidean distance, Manhattan distance, Minkowski distance, and Hamming distance. K-means algorithm uses the compactness approach. In connectivity, the points in a cluster are either immediately next to each other (epsilon distance) or connected. Even if the distance is less, they are not put in the same cluster. Spectral clustering is one of the techniques to follow this approach.

