Centroids-based clustering

Centroids-based clustering, also known as K-means clustering, is a popular unsupervised machine learning algorithm used to partition a dataset into K clusters. The goal of the algorithm is to find the K cluster centers, also known as centroids, that minimize the distance between the data points and their assigned cluster centers.

To illustrate this concept, let's consider an example of a grocery store. Suppose the grocery store wants to segment their customers into different groups based on their purchasing behavior. They have data on each customer's purchases, such as the types of products, the frequency of purchases, and the amount spent on each transaction. Using this data, the grocery store can apply the K-means clustering algorithm to cluster their customers into different segments based on their purchasing behavior.

The working of the algorithm can be summarized as follows:

Choose the number of clusters K that you want to form

Randomly initialize K cluster centers, which represent the centroids of each cluster

Assign each data point to the nearest cluster center based on the distance metric (usually Euclidean distance)

Recalculate the cluster centers by taking the mean of all the data points assigned to each cluster

Repeat steps 3 and 4 until the cluster centers no longer change significantly, or until a predetermined number of iterations have been reached.

The algorithm aims to minimize the within-cluster sum of squares (WCSS) by finding the cluster centers that minimize the distance between each data point and its assigned cluster center.

The K-means clustering algorithm is widely used in a variety of real-world applications, including customer segmentation, image segmentation, anomaly detection, and market research.