5CS037-

Concepts and Technologies of AI

Week 11: K-Means Clustering

Name: Nayanika Dubey
Group: L5CG14
Week: 11

What is K-means clustering?

K-means clustering is a popular unsupervised learning algorithm used in data mining and machine learning to group similar data points into clusters. The algorithm aims to partition the data into k clusters, where each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. K-means is required to uncover patterns, similarities, and differences in data that are not immediately apparent.

How does it work?

K-means clustering is a widely used unsupervised learning algorithm used to group together similar data points in a dataset. It is a simple but effective way to find patterns in data and is commonly used in various fields such as image segmentation, customer segmentation, and anomaly detection.

The K-means algorithm works in the following phases and sequence:

Phase 1: Initialization

The first step is to initialize the algorithm by selecting the number of clusters (K) to create and choosing random points as the initial centroids for each cluster. The value of K can be determined by prior knowledge or using techniques such as the elbow method or silhouette analysis.

Phase 2: Assignment

In this step, each data point in the dataset is assigned to its nearest centroid, based on the Euclidean distance between the data point and the centroid. This forms K clusters.

Phase 3: Update

The next step is to update the centroids of each cluster by calculating the mean of all the data points assigned to it. The updated centroid becomes the new center of the cluster.

Phase 4: Repeat

Steps 2 and 3 are repeated until the algorithm converges and the centroids no longer change significantly. This means that the clusters have been optimized and stabilized.

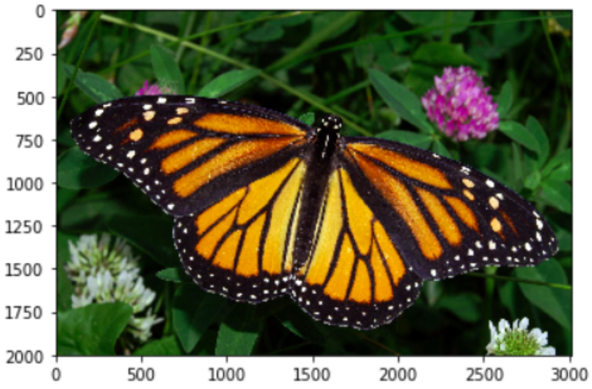
Phase 5: Output

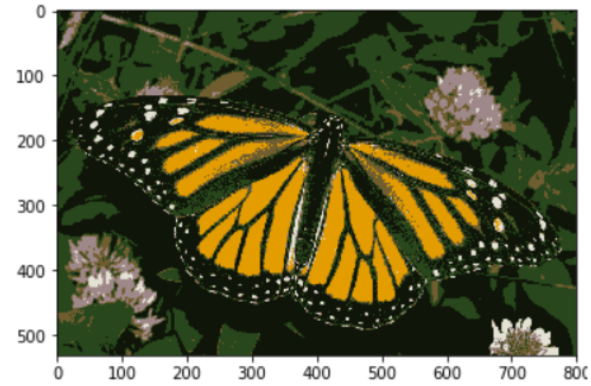
The final output of the K-means algorithm is the K clusters and their respective centroids. The data points in the dataset are assigned to the cluster with the closest centroid, based on the Euclidean distance.

Feature	**Lloyd's Algorithm (Standard K-Means)**	**K-Means++**	**Mini-Batch K-Means**	**Hierarchical K-Means**
Core Mechanism	Iterative assignment and centroid update	Improved centroid initialization	Uses small data batches per iteration	Builds a hierarchy of clusters
Centroid Update	Mean of assigned data points	Mean of assigned data points	Mean of assigned data points in the batch	Determined by the hierarchical process
Key Benefit	Basic, widely used	Improved convergence, avoids suboptimal solutions	Faster for large datasets	Identifies clusters at different granularities
Dataset Suitability	General purpose	General purpose	Large datasets	Exploring data structure, hierarchical data

Image Segmentation:

K-means clusters similar pixels to segment images, useful in healthcare and robotics. It iteratively assigns pixels to the nearest cluster centroid, recalculating centroids until convergence. This groups pixels by color, intensity, or texture, identifying objects or regions. Applications include medical image analysis, robotics, and computer vision. Though computationally intensive and potentially suboptimal, its simplicity makes it popular.





Customer Segmentation:

K-means segments customers based on purchase history, demographics, and behavior. It clusters data points by minimizing the distance to cluster centers. Retailers use it to group customers by purchase behavior (frequency, spending, products) for targeted marketing. Marketers segment based on online behavior (browsing, searches, social media) for personalized campaigns.