K-means

K-means is one of the most popular and widely used clustering algorithms in machine learning. It's a simple and efficient method for partitioning a dataset into K distinct, non-overlapping subgroups (clusters) where each data point belongs to only one group.

Algorithm Steps:

a. Initialization:

- Choose the number of clusters, K.
- Randomly initialize K centroids in the feature space.

b. Assignment:

- Assign each data point to the nearest centroid based on Euclidean distance.

c. Update:

- Recalculate the centroids of each cluster by taking the mean of all points assigned to that cluster.

d. Repeat:

- Iterate steps b and c until convergence (centroids no longer move significantly) or a maximum number of iterations is reached.

Advantages:

- Simple and easy to implement
- Efficient and scalable to large datasets
- Guarantees convergence (to a local optimum)

Limitations:

- Requires specifying the number of clusters (K) in advance
- Sensitive to initial centroid positions
- Assumes clusters are spherical and equally sized
- Can be affected by outliers
- May converge to local optima

Applications:

- Customer segmentation
- Image compression
- Document clustering
- Anomaly detection
- Feature learning