K-Means Algorithm

Introduction

K-Means is one of the most popular unsupervised machine learning

algorithms used for clustering. It partitions a dataset into K distinct,

non-overlapping clusters based on feature similarity.

The algorithm was first proposed by Stuart Lloyd in 1957 and later

published in 1982. It is widely used in data mining, pattern

recognition, and image compression.

------------------------------------------------------------------------

Working of K-Means Algorithm

The K-Means algorithm follows an iterative refinement approach. The main

steps are:

1. Choose K: Select the number of clusters (K) you want to identify in

the dataset.

2. Initialize Centroids: Randomly select K points as the initial

cluster centroids.

3. Assign Clusters: Assign each data point to the nearest centroid

using a distance metric (commonly Euclidean distance).

4. Update Centroids: Recalculate the centroids as the mean of all data

points belonging to a cluster.

5. Repeat: Continue steps 3 and 4 until:

- Centroids no longer change significantly, or

- A maximum number of iterations is reached.

------------------------------------------------------------------------

Objective Function

K-Means aims to minimize the Within-Cluster Sum of Squares (WCSS), also

known as the inertia.

““” J = {i=1}^{K} {x C\_i} || x - \_i ||^2 ““”

Where: - ( C\_i ) = set of points in cluster i

- ( \_i ) = centroid of cluster i

- ( || x - \_i ||^2 ) = squared distance between point and centroid

------------------------------------------------------------------------

Key Properties

- Distance Metric: Usually Euclidean distance, but other metrics can

be used.

- Hard Clustering: Each data point belongs to exactly one cluster.

- Centroid-based: Clusters are represented by their mean value.

------------------------------------------------------------------------

Advantages of K-Means

1. Simple and easy to implement.

2. Works well with large datasets.

3. Computationally efficient: ( O(n K t) ), where n = number of data

points, K = number of clusters, t = iterations.

4. Produces tighter clusters than hierarchical clustering (in many

cases).

------------------------------------------------------------------------

Disadvantages of K-Means

1. Need to specify K in advance.

2. Sensitive to the initial placement of centroids.

3. Sensitive to outliers and noisy data.

4. Works only with numerical data (distance metric-based).

5. Assumes clusters are spherical and of equal size (not always

realistic).

------------------------------------------------------------------------

Variants of K-Means

- K-Means++: Improves initialization of centroids to avoid poor

clustering.

- Mini-Batch K-Means: Uses small random batches of data for faster

computation.

- Fuzzy K-Means (FCM): Allows soft clustering (a point can belong to

multiple clusters with probabilities).

------------------------------------------------------------------------

Applications of K-Means

1. Image Compression – Reduces the number of colors by clustering pixel

values.

2. Customer Segmentation – Groups customers by purchasing behavior.

3. Document Clustering – Organizes documents based on similarity.

4. Anomaly Detection – Finds unusual data points by distance from

centroids.

5. Recommendation Systems – Clusters users or products for better

recommendations.

------------------------------------------------------------------------

Limitations and Improvements

- Choosing K: The “Elbow Method” and “Silhouette Score” help determine

an optimal K.

- Scalability: For very large datasets, Mini-Batch K-Means is

preferred.

- Initialization Problem: K-Means++ helps in better centroid

initialization.

- Non-globular Clusters: K-Means may fail when clusters are not

spherical (e.g., concentric circles).

------------------------------------------------------------------------

Example (High-Level)

Suppose we want to cluster students based on their Math and English

scores into 2 groups (K=2).

1. Randomly select 2 students’ scores as initial centroids.

2. Assign all students to the nearest centroid.

3. Recalculate the centroids (average scores of students in each

group).

4. Repeat until centroids stabilize.

The result is two groups of students: one strong in Math, another in

English.

------------------------------------------------------------------------

Conclusion

K-Means is a fundamental clustering algorithm with wide applications.

Despite its limitations (sensitivity to K and outliers), it remains a

powerful tool, especially when combined with improvements like K-Means++

and Mini-Batch K-Means.