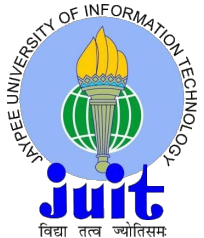


Machine Learning

(Course Code: 18B1WCI634)

K-Means Clustering

Mr. Sandeep Kumar Patel

Assistant Professor

Jaypee University of Information Technology

Waknaghat, Solan, HP-173234

Introduction to Clustering

- Clustering is an unsupervised learning technique.
- It groups similar data points together.
- Objects in the same cluster are similar.
- Objects in different clusters are dissimilar.

Goal of Clustering

- Maximize intra-cluster similarity
- Minimize inter-cluster similarity

What is K-Means?

Definition:
K-Means is a partition-based unsupervised clustering algorithm that divides data into K clusters.

Key Features

- Unsupervised algorithm
- Centroid-based clustering
- Iterative process
- Uses distance calculations

Why the Name K-Means?

K

Number of clusters

Means

Mean value (centroid) of data points in a cluster

Characteristics of K-Means

Feature	Description
Learning Type	Unsupervised
Method	Partitioning
Cluster Type	Hard Clustering
Distance Metric	Euclidean Distance
Output	K Clusters

Applications of K-Means

- Customer segmentation
- Image segmentation
- Recommendation systems
- Fraud detection
- Medical diagnosis
- Market analysis
- Document clustering

Working Principle of K-Means

1. Choose number of clusters K
2. Initialize centroids
3. Assign points to nearest centroid
4. Recalculate centroids
5. Repeat until convergence

Objective Function

K-Means minimizes the following objective function:

$$J = \sum_{i=1}^{K} \sum_{x_j \in C_i} ||x_j - \mu_i||^2$$

Where:

- K = Number of clusters
- C_i = Cluster i
- μ_i = Centroid
- x_j = Data point

Euclidean Distance

For two points:

$$P(x_1, y_1), Q(x_2, y_2)$$

Distance formula:

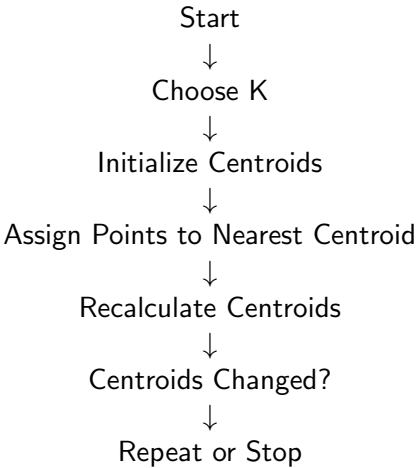
$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Used to calculate similarity between points.

Steps of K-Means Algorithm

1. Select K clusters
2. Randomly initialize centroids
3. Assign data points to nearest centroid
4. Update centroids
5. Repeat until convergence

Flowchart of K-Means



Numerical Example

Point	Coordinates
A	(1,1)
B	(2,1)
C	(4,3)
D	(5,4)

Assume:

$$K = 2$$

Initial Centroids

Choose initial centroids:

- $C_1 = (1, 1)$
- $C_2 = (5, 4)$

Next:

- Compute distances
- Assign nearest cluster

Distance Calculation Example

Distance of A(1,1) from $C_1(1,1)$:

$$0$$

Distance of A(1,1) from $C_2(5,4)$:

$$\sqrt{(5-1)^2+(4-1)^2}=5$$

Therefore: A belongs to Cluster 1.

Cluster Formation

Cluster 1

- A(1,1)
- B(2,1)

Cluster 2

- C(4,3)
- D(5,4)

Recalculation of Centroids

Cluster 1 Centroid

$$\left(\frac{1+2}{2}, \frac{1+1}{2}\right) = (1.5, 1)$$

Cluster 2 Centroid

$$\left(\frac{4+5}{2}, \frac{3+4}{2}\right) = (4.5, 3.5)$$

Convergence

The algorithm repeats:

- Assignment step
- Update step

Until:

- Centroids stop changing

This process is called convergence.

Pseudocode

Algorithm

1. Initialize K centroids
2. Repeat
   Assign points to nearest centroid
   Recalculate centroids
3. Until convergence
4. Return clusters

Choosing the Value of K

Importance of Choosing K

- Small K $\rightarrow$ Under-clustering
- Large K $\rightarrow$ Over-clustering
- Correct K gives better cluster quality

Common Methods

- Elbow Method
- Silhouette Score
- Gap Statistic
- Davies-Bouldin Index
- Calinski-Harabasz Index

Elbow Method

Idea:
Run K-Means for different values of K and compute WCSS.

WCSS Formula

$$WCSS = \sum_{i=1}^{K} \sum_{x_j \in C_i} ||x_j - \mu_i||^2$$

Where:

- x_j = data point
- μ_i = centroid
- C_i = cluster

Working of Elbow Method

1. Run K-Means for different K values
2. Compute WCSS for each K
3. Plot K vs WCSS
4. Identify the elbow point

The elbow point gives the optimal value of K.

Advantages and Limitations of Elbow Method

Advantages

- Simple and intuitive
- Easy visualization
- Widely used

Limitations

- Elbow may not be clearly visible
- Subjective interpretation

Silhouette Score Method

Idea:
Measures how well data points fit within their clusters.

Formula

$$S = \frac{b - a}{\max(a, b)}$$

Where:

- a = average intra-cluster distance
- b = average nearest-cluster distance

Interpretation of Silhouette Score

$$-1 \leq S \leq 1$$

Score	Meaning
Close to 1	Good clustering
Close to 0	Overlapping clusters
Negative	Incorrect clustering

Higher silhouette score indicates better clustering.

Gap Statistic Method

Idea:
Compares clustering performance with randomly generated data.

Procedure

1. Compute WCSS for actual data
2. Generate random datasets
3. Compare results
4. Select K with maximum gap

Davies-Bouldin Index

Idea:

Measures:

- Intra-cluster similarity
- Inter-cluster separation

Interpretation

- Lower DBI $\rightarrow$ Better clustering

Calinski-Harabasz Index

Idea:
Measures ratio of:

- Between-cluster dispersion
- Within-cluster dispersion

Interpretation

- Higher score $\rightarrow$ Better clustering

Comparison of Methods

Method	Best Criterion	Complexity
Elbow Method	Elbow Point	Low
Silhouette Score	Highest Score	Medium
Gap Statistic	Maximum Gap	High
Davies-Bouldin	Lowest DBI	Medium
Calinski-Harabasz	Highest Score	Medium

Summary

- Choosing K is important in K-Means.
- Elbow Method is the most common approach.
- Silhouette Score measures cluster quality.
- Gap Statistic is more statistically robust.
- Correct K improves clustering accuracy.

Time Complexity

Complexity of K-Means:

$$O(nkdi)$$

Where:

- n = Number of data points
- k = Number of clusters
- d = Dimensions
- i = Iterations

Advantages of K-Means

- Simple to implement
- Fast algorithm
- Scalable to large datasets
- Easy interpretation
- Computationally efficient

Disadvantages of K-Means

- Requires predefined K
- Sensitive to initialization
- Sensitive to outliers
- Poor for non-spherical clusters
- May converge to local optimum

K-Means++

Improved Initialization Method

Advantages:

- Better centroid initialization
- Faster convergence
- Improved clustering accuracy

Comparison with Other Algorithms

Algorithm	Type	Main Idea
K-Means	Partitioning	Mean-based
Hierarchical	Tree-based	Nested clusters
DBSCAN	Density-based	Dense regions
Mean Shift	Density-based	Density peaks

Industrial Applications

Industry	Application
Banking	Fraud detection
Healthcare	Disease grouping
Marketing	Customer segmentation
Telecom	User analysis
E-commerce	Recommendation systems

Summary

- K-Means is an unsupervised clustering algorithm.
- Divides data into K clusters.
- Uses centroids and Euclidean distance.
- Iterative refinement process.
- Widely used in machine learning.