Unsupervised ML

to discover groups in the data → clustering

Datasets ↳ target column → missing (unlabelled data)

Real-time applications

1. Customer segmentation (Marketing, EdTech, fintech)
2. Image Compression
3. Anomaly Detection
4. Document/topic grouping

Semi-supervised Technique

↳ 1000 Records

400 Records ↳ labeled data

600 Records ↳ unlabelled data → Clustering → labels

labeled data → labels → Supervised ML

Distance measures :-

1. Euclidean distance
2. Manhattan distance

Numeric points

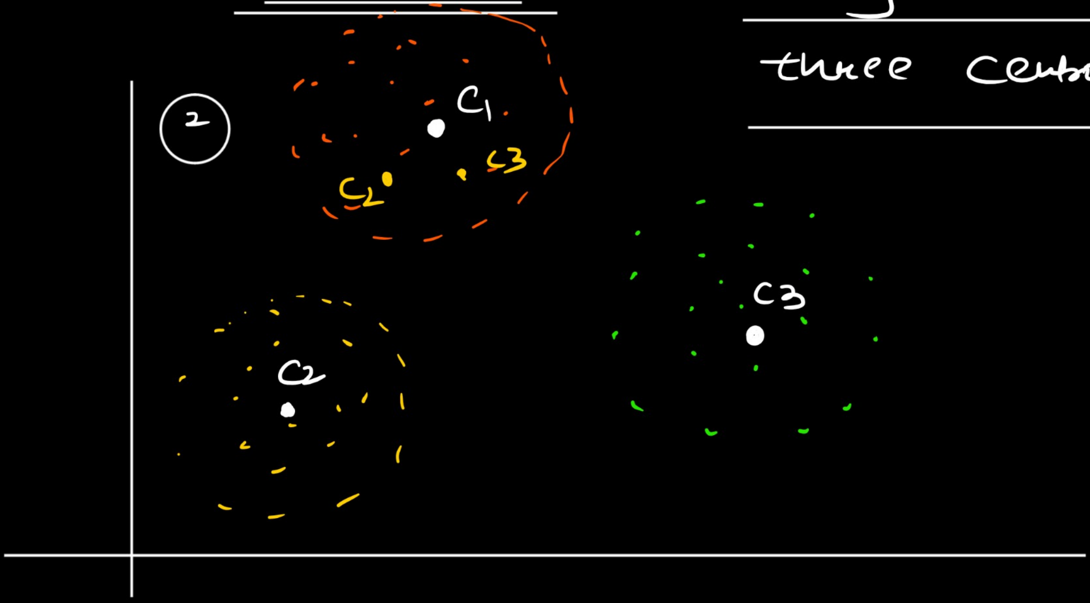
3. Hamming distance
4. Cosine similarity (for textual data)

K-Means clustering → customer segmentation

clusters

1. K = 3 → Randomly initialise three centroids

2.



3. Re-evaluation of centroids (c1, c2, c3)

$c_1 = \frac{P_1 + P_3 + P_5}{3}$ $c_2 = \frac{P_2 + P_4 + P_6}{3}$ $c_3 = \frac{P_7 + P_8 + P_9 + P_{10}}{4}$

C_1 ———— $P_1 \to 2$

C_1 ———— $P_2 \to 1$

C_1 ———— P_3

C_1 ————

|

C_1 ———— P_{10}

C_2 ———— $P_1 \to 1$

C_2 ———— $P_2 \to 2$

C_2 ———— P_3

————

|

C_2 ———— P_{10}

C_3 ———— $P_1 \to 5$

C_3 ———— $P_2 \to 3$

C_3 ———— P_3

|

C_3 ———— P_{10}

(x_1, y_1) (x_3, y_3) (x_4, y_4) (x_7, y_7)

$$\left(\frac{x_1+x_3+x_4+x_7}{4}, \frac{y_1+y_3+y_4+y_7}{4}\right) \longleftarrow C_1 \longrightarrow P_1, P_3, P_4, P_7$$

$C_2 \longrightarrow P_2, P_5, P_6, P_8$

$C_3 \longrightarrow P_9, P_{10}, P_{11}, P_{12}$

Steps of KMeans

1. Choose number of clusters
   k ✓
   k.
2. Initialize centroids randomly.
3. Assign points to nearest centroid.
4. Update centroid as mean of assigned points.
5. Repeat until convergence.

Advantages

↳ simple yet efficient for large datasets.

1 simple yet efficient for large datasets.

2 Works well with spherical clusters

Disadvantages

1 Requires # clusters (k) beforehand. → elbow Method

2 Sensitive to outliers

3 Randomly initialize centroids (within 1 cluster)

↓

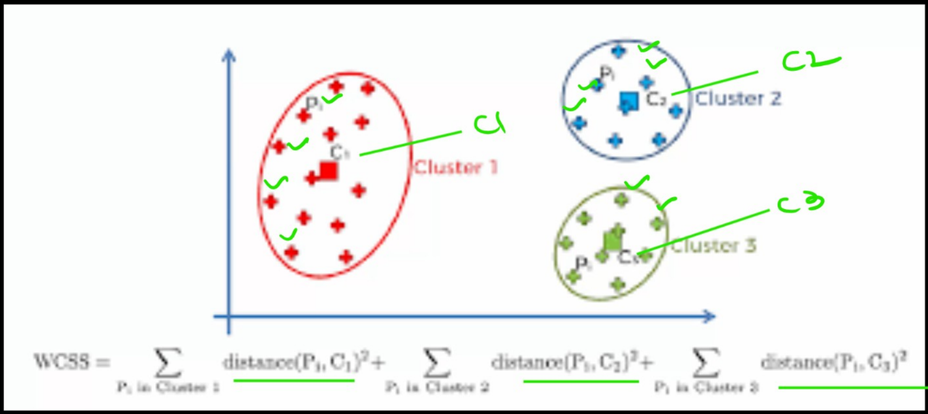
kmeans ++ → Task

Elbow Method

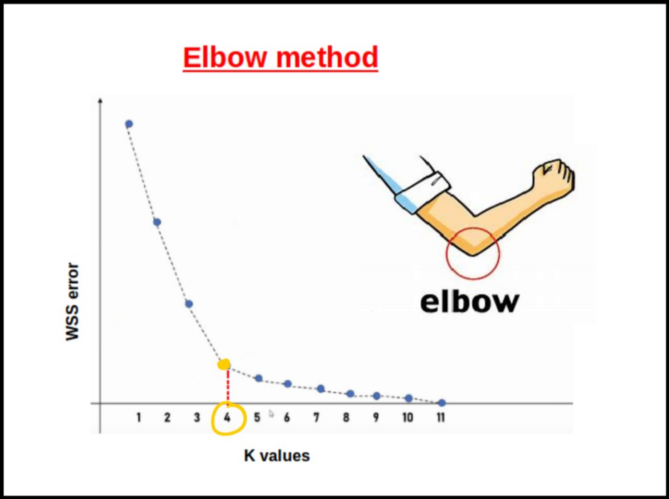
WCSS → Within cluster summation of squares

or

Inertia

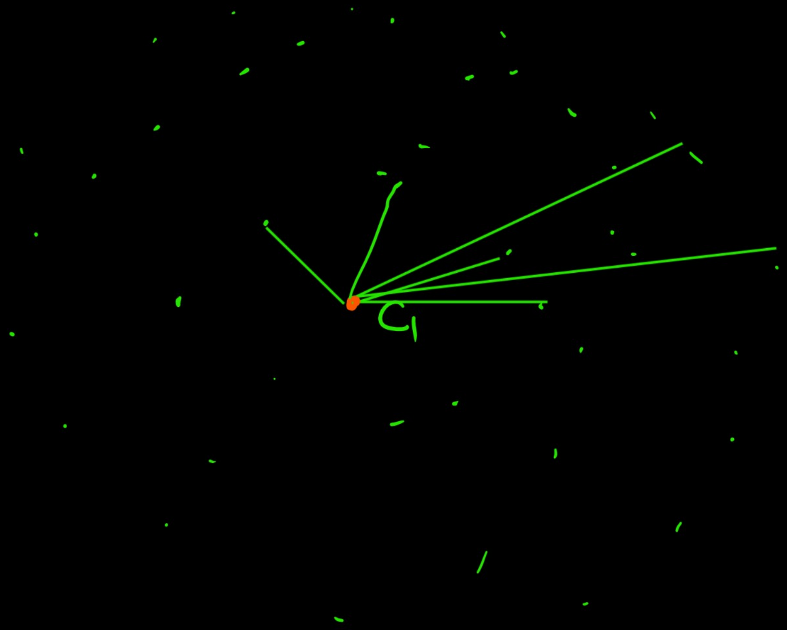


$$WCSS = \sum_{P_i \text{ in Cluster 1}} \underline{distance(P_i, C_1)^2} + \sum_{P_i \text{ in Cluster 2}} \underline{distance(P_i, C_2)^2} + \sum_{P_i \text{ in Cluster 3}} \underline{distance(P_i, C_3)^2}$$



K = 1 ———— WCSS = max (Peak)

K = n ———— WCSS = 0



$n = 4$