K-Means Clustering

07 October 2025 17:18

1. What is K-Means Clustering?

- **K-Means** is one of the most popular **unsupervised learning algorithms** used to partition a dataset into **K clusters**.
- Each cluster is represented by its **centroid**, and every data point belongs to the cluster with the nearest centroid.

Goal: Minimize the total variance (distance) within each cluster.

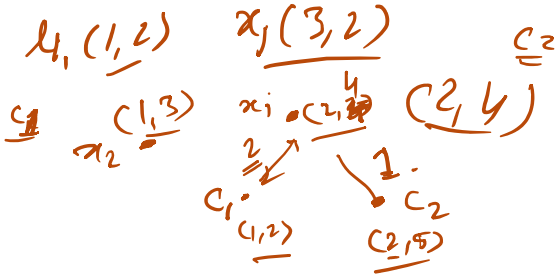
2. Objective Function (Mathematics Behind K-Means)

We aim to minimize the **Sum of Squared Errors (SSE)** — the distance between each point and its assigned centroid.

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

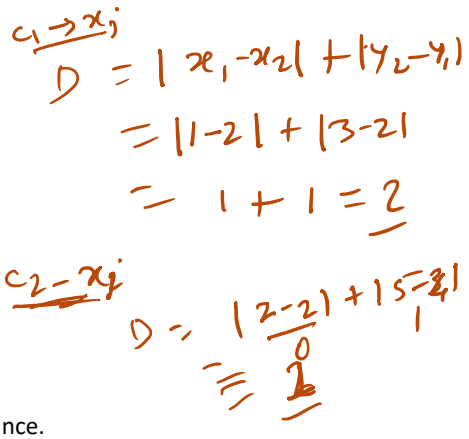
Where:

- K: number of clusters
- C_i: set of data points in cluster i
- μ_i: centroid of cluster i
- x_j: data point in cluster C_i



3. K-Means Algorithm Steps

Step	Description
1. Initialize	Choose K cluster centers (centroids) randomly or using K-Means++
2. Assignment	Assign each data point to the nearest centroid
3. Update	Recalculate centroids as the mean of all assigned points
4. Repeat	Repeat steps 2–3 until centroids do not change (convergence)



Numerical Example:

Short goal: cluster 7 two-dimensional points into **k = 2** clusters using Euclidean distance.

Data (7 points)

A = (1,2)

B = (1,4)

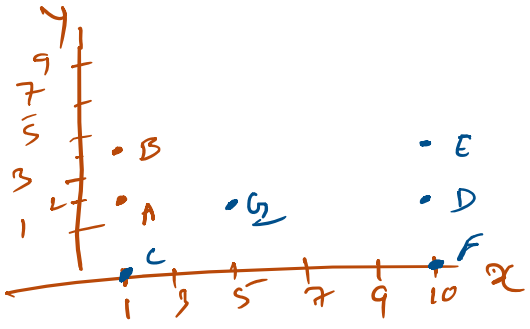
C = (1,0)

D = (10,2)

E = (10,4)

F = (10,0)

G = (5,2)



G = (3,2)

K = 2 (1) $C_2 (10,0)$

$C_1 = (1,0)$

(2) Assign each data point to the nearest centroid.

$D = |x_2 - x_1| + |y_2 - y_1|$

$A(1,2) \rightarrow C_1(1,0)$

$D_1 = |1-1| + |0-2|$

$0 + 2 = 2$

$A(1,2) \rightarrow C_2(10,0)$

$D_2 = |10-1| + |0-2|$

$= 9 + 2 = 11$

$A \rightarrow C_1$

$B(1,4) \rightarrow C_1(1,0) \rightarrow d \rightarrow 4$

$B(1,4) \rightarrow C_2(10,0) \rightarrow d \rightarrow 13$

$B \rightarrow C_1$

$C(1,0) \rightarrow C_1(1,0) \Rightarrow d = 0$

$C(1,0) \rightarrow C_2(10,0) \Rightarrow d = 9$

$C \rightarrow C_1$

$D(10,2) \rightarrow C_1(1,0) \Rightarrow d = 11$

$D(10,2) \rightarrow C_2(10,0) \Rightarrow d = 2$

$D \rightarrow C_2$

$E \rightarrow C_2$

E → c_2

F → c_2

$G_1(5,2) \rightarrow c_1(1,0) \rightarrow 6$

$G_2(5,2) \rightarrow c_2(10,2) \rightarrow 7$

$G_2 \rightarrow c_1$

(3) Update the centroid with mean of the assigned points.

$c_1 \rightarrow (A(1,2), B(1,4), C(1,0), G_1(5,2)$

$\bar{x} = \frac{1+1+1+5}{4} = 2$

$\bar{y} = \frac{2+4+0+2}{4} = 2$

$c_1(1,0) \rightarrow (2,2)$

c_1 (2,2)

$c_2 = (D \quad E, F$

D = (10,2)

E = (10,4)

F = (10,0)

$\bar{x} = \frac{10+10+10}{3} = 10$

$\bar{y} = \frac{2+4+0}{3} = 2$

c_2 (10,2)

$c_1(2,2), \quad c_2$ (10,2)

A = (1,2) $\rightarrow C_1$

B = (1,4) $\rightarrow C_1$

C = (1,0) $\rightarrow C_1$

D = (10,2) $\rightarrow C_2$

E = (10,4) $\rightarrow C_2$

F = (10,0) $\rightarrow C_2$

G = (5,2) $\rightarrow C_1$

$C_1 \Rightarrow \{A(1,2), B(1,4), C(1,0), G(5,2)\}$

$C_2 = \{D(10,2), E(10,4), F(10,0)\}$

$C_1(2,2)$

$C_2(10,2)$

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

$$[A(1,2) - C_1(2,2)]^2 + [B(1,4) - C_1(2,2)]^2 + [C(1,0) - C_1(2,2)] + [G(5,2) - C_1(2,2)]^2$$

$\text{Cluster}_1\text{-error} = 1 + 5 + 5 + 9 = 20$

$\text{Cluster}_2\text{-error} = [D(10,2) - C_2(10,2)]^2 + [E(10,4) - C_2(10,2)]^2 + [F(10,0) - C_2(10,2)]^2$

$= 0 + 4 + 4 = 8$

$\text{SSE} = \text{Cluster}_1\text{-err} + \text{Cluster}_2\text{-error}$

$= 20 + 8 = 28$

Find best K value

25 November 2025 09:50

How to Choose the Right Number of Clusters (K)?

1 Elbow Method

The **Elbow Method** is one of the most popular techniques to determine the optimal number of clusters in **K-Means**.

Concept

- For different values of K, compute the **inertia** (also known as **SSE – Sum of Squared Errors**):

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

where:

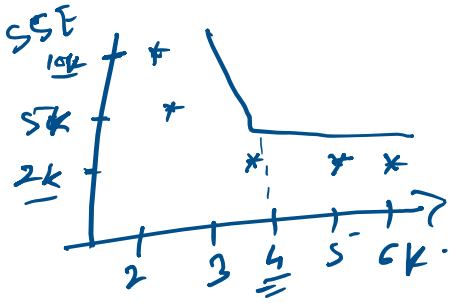
- C_i: cluster i
- μ_i: centroid of cluster i
- x_j: data point belonging to cluster i

- As K increases:
  - **SSE decreases** (clusters get smaller and fit data better).
  - But after a certain point, the **marginal gain drops sharply** — forming an "elbow" in the curve.

Visualization

Plot **SSE vs. K** and look for the **elbow point** — the value of KKK after which SSE reduction slows down significantly.

```
K →
|     *
|     *
|     *     ← Elbow Point (Optimal K)
|   *
|  *
|_*_____________________
   2  3  4  5  6
      Number of Clusters (K)
```



Interpretation

- The **"elbow" point** represents a good trade-off between:
  - Lower **SSE** (better compactness)
  - Simpler model (fewer clusters)
- Too few clusters → high SSE (underfitting)
- Too many clusters → very low SSE but may **overfit** and lose interpretability.

Silhouette Score

The **Silhouette Score** measures how similar each data point is to its **own cluster** compared to **other clusters**. It provides an indication of the **quality of clustering** — how well-separated and cohesive the clusters are.

Formula

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

$\frac{10-5}{10}$ $= \frac{5}{10} = \frac{1}{2}$

Where:

- a = Mean **intra-cluster distance**
  (the average distance between a sample and all other points in the **same cluster**)
- b = Mean **nearest-cluster distance**
  (the average distance between a sample and all points in the **closest neighboring cluster**)

Interpretation

Silhouette Score (S)	Interpretation
+1	Perfectly clustered — points are well-matched to their own cluster and far from others
0	Points are on or near the decision boundary between clusters
–1	Incorrect clustering — points are assigned to the wrong cluster

---

Formula Intuition

- If $a \ll b \rightarrow S$ approaches **1**, meaning well-separated clusters.
- If $a \approx b \rightarrow S$ is **0**, meaning overlapping clusters.
- If $a > b \rightarrow S$ becomes **negative**, meaning misclassified points.

A(1,2), B(1,4)
C1
C(1,0) G(5,2)

D(10,2)
C2
E(10,4), F(10,0)

A → B
→ C
→ D

$a = \frac{|(1-1)+(2-4)| + |(1-1)+(2-0)| + |(1-5)+(2-2)|}{3}$

$$a = \frac{2 + 2 + 4}{3} = \frac{8}{3} = 2.66$$

$$b = \frac{A \to D + A \to E + A \to F}{3} = \frac{9 + 11 + 11}{3}$$

$$= \frac{31}{3} = 10.33$$

$$S = \frac{b - a}{\max(a, b)} = \frac{10.33 - 2.66}{10.33} = \frac{7.67}{10.33}$$

$$S = 0.74$$