Machine Learning 101

Rajdeep Chatterjee, Ph.D.
Amygdala AI, Bhubaneswar, India *

January 2025

K-Means Clustering Algorithm

1 K-means Algorithm

The **K-means algorithm** is a clustering technique that partitions a dataset into K distinct non-overlapping clusters. It minimizes the sum of squared distances between each data point and the centroid of its assigned cluster. The algorithm iteratively alternates between assigning points to clusters and updating cluster centroids.

1.1 Pseudo-code for K-means Algorithm

Algorithm 1 K-means Algorithm

Input: Dataset $X = \{x_1, x_2, \ldots, x_n\}$, number of clusters K
Output: Clusters $C_1, C_2, \ldots, C_K$ and centroids $\mu_1, \mu_2, \ldots, \mu_K$
Initialize K centroids $\mu_1, \mu_2, \ldots, \mu_K$ randomly.
while not converged **do**
    Assign each data point x_i to the nearest centroid:

$$C_k = \{x_i : \|x_i - \mu_k\|^2 \leq \|x_i - \mu_j\|^2, \forall j, 1 \leq j \leq K\}$$

    Update each centroid μ_k based on the mean of points in C_k:

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$

end while
Return C_k and μ_k

*Amygdala AI, is an international volunteer-run research group that advocates for *AI for a better tomorrow* `http://amygdalaai.org/`.

1.2 Key Equations

1. Distance between a point and a centroid:

$$\|x_i - \mu_k\|^2 = \sum_{j=1}^{d} (x_{ij} - \mu_{kj})^2$$

2. Objective function to minimize:

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

2 Within-Cluster Sum of Squares (WCSS)

In the K-means algorithm, the **Within-Cluster Sum of Squares (WCSS)** measures the compactness of clusters by calculating the sum of squared distances between each data point and the centroid of its cluster.

2.1 Mathematical Definition

Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$ partitioned into K clusters $C_1, C_2, \ldots, C_K$ with centroids $\mu_1, \mu_2, \ldots, \mu_K$, the WCSS is defined as:

$$\text{WCSS} = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2$$

where: - $\|x_i - \mu_k\|^2 = \sum_{j=1}^{d} (x_{ij} - \mu_{kj})^2$ is the **squared Euclidean distance** between data point x_i and centroid μ_k, - d is the dimensionality of the data.

Minimizing the WCSS is the objective of the K-means algorithm, as it ensures the clusters are compact.

2.2 Limitations of the K-means Algorithm

- The K-means algorithm **does not always produce globally optimal results**, as it may converge to a local minimum of the objective function.
- The result of the K-means algorithm **depends on the initial cluster centroids**, which can lead to different cluster configurations for different initializations.
- The K-means algorithm can only be applied to datasets where data points lie in a **Euclidean space**. It cannot handle datasets with non-Euclidean distances or categorical data effectively.

3 Toy Problem: Step-by-Step Example

Dataset: $X = \{(1,1), (2,1), (4,3), (5,4), (3,2), (6,5)\}$, $K = 2$

Step 1: Initialization

Randomly initialize centroids:

$$\mu_1 = (1,1), \quad \mu_2 = (5,4)$$

Step 2: Assign points to clusters

Compute distances:

$$\|x_1-\mu_1\|^2=0, \quad \|x_1-\mu_2\|^2=25$$
$$\|x_2-\mu_1\|^2=1, \quad \|x_2-\mu_2\|^2=20$$
$$\|x_3-\mu_1\|^2=13, \quad \|x_3-\mu_2\|^2=1$$
$$\|x_4-\mu_1\|^2=25, \quad \|x_4-\mu_2\|^2=0$$
$$\|x_5-\mu_1\|^2=8, \quad \|x_5-\mu_2\|^2=5$$
$$\|x_6-\mu_1\|^2=50, \quad \|x_6-\mu_2\|^2=2$$

Clusters:

$$C_1=\{(1,1),(2,1),(3,2)\}, \quad C_2=\{(4,3),(5,4),(6,5)\}$$

Step 3: Update centroids

$$\mu_1=\frac{(1,1)+(2,1)+(3,2)}{3}=\left(\frac{6}{3},\frac{4}{3}\right)=(2,1.33)$$

$$\mu_2=\frac{(4,3)+(5,4)+(6,5)}{3}=\left(\frac{15}{3},\frac{12}{3}\right)=(5,4)$$

Step 4: Repeat until convergence Reassign clusters based on updated centroids and recompute centroids until centroids do not change.

Step 5: Final Result After convergence, the final clusters and centroids are:

$$C_1=\{(1,1),(2,1),(3,2)\}, \quad C_2=\{(4,3),(5,4),(6,5)\}$$

$$\mu_1=(2,1.33), \quad \mu_2=(5,4)$$

Step 4: Repeat until convergence Reassign clusters and recompute centroids until centroids do not change.