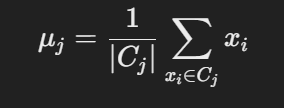
Mathematical Foundations of K-Means Clustering

The core steps involve assigning data points to the nearest cluster centroids and then updating the centroids based on these assignments.

1. Centroid Update Rule



Where:

Cj is the set of points that belong to the j-th cluster.

∣Cj∣ is the number of points in Cj.

x\_i stands for the data points in the j-th cluster.

μ\_j is the new centroid of the j-th cluster.

This formula calculates the arithmetic mean of all the data points in the cluster, ensuring that the new centroid is at the geometric center of the cluster. The centroid update rule ensures that the sum of squared distances from the data points to their respective centroids is minimized.

2. Algorithm Convergence Criteria

The K-means algorithm iterates through the following steps until convergence is achieved:

* Assign Data Points to Nearest Centroid:

For each data point x\_i , calculate the Euclidean distance to each centroid μ\_j .

Assign each data point x\_i to the cluster Cj whose centroid is the nearest.

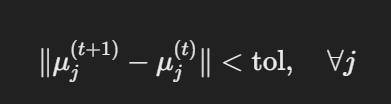
* Update Centroids:

Recalculate the centroids μ\_j using the centroid update rule described above.

* Check for Convergence:

The algorithm stops when one of the following convergence criteria is met:

Centroid Stability: The centroids no longer change significantly between iterations. This is measured by checking if the movement of the centroids is smaller than a predefined tolerance level tol:

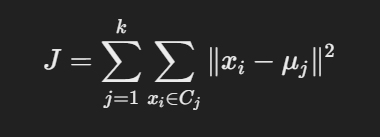


Where ∥⋅∥ denotes the Euclidean norm (distance), μ\_j^(t+1) is the centroid in the current iteration, and μ\_j^(t) is the centroid from the earlier iteration.

Maximum Iterations: The algorithm stops after a predefined maximum number of iterations max\_iter is reached. This prevents the algorithm from running indefinitely in cases where it might oscillate or converge very slowly.

3. Objective Function and its Minimization

The K-means algorithm looks to minimize the following objective function, also known as the within-cluster sum of squares (WCSS) or inertia:



Where:

J is the objective function that K-means minimizes.

∥x\_i−μ\_j∥^2 is the squared Euclidean distance between a data point x\_i and its assigned centroid μ\_j .

The sum is taken over all clusters k and all points within each cluster.

The centroid update step ensures that in each iteration, the objective function J decreases or stays the same, as the centroids are moved to the geometric center of their respective clusters. Since the function is non-increasing and bounded below by zero, the algorithm is guaranteed to converge to a local minimum.