K-Means Clustering - Mathematical Notes

Mathematical Formulation:
The K-Means algorithm aims to minimize the objective function (inertia or WCSS):

$J = \Sigma \text{ (from k=1 to K) } \Sigma (x \in C_k) \; || \; x - \mu_k \; ||^2$

Where:
- K = number of clusters
- μ_k = centroid of cluster k
- x = data point in cluster C_k
- $|| \; x - \mu_k \; ||^2$ = squared Euclidean distance

Example:
Suppose we have the following 4 data points in 2D:
P1 = (2, 3), P2 = (3, 3), P3 = (8, 8), P4 = (9, 8)
We want to cluster them into K = 2 clusters.

Step 1: Initialize Centroids
Let initial centroids be:
C1 = P1 = (2, 3), C2 = P3 = (8, 8)

Step 2: Assign Points to Nearest Centroid
Distance formula: $d(x, y) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2}$

- $d(P1, C1) = 0$, $d(P1, C2) = \sqrt{(2-8)^2+(3-8)^2}=\sqrt{61} \approx 7.81 \rightarrow P1 \rightarrow C1$
- $d(P2, C1) = \sqrt{(3-2)^2+(3-3)^2}=1$, $d(P2, C2)=\sqrt{(3-8)^2+(3-8)^2}=\sqrt{50}\approx 7.07 \rightarrow P2 \rightarrow C1$
- $d(P3, C1) = \sqrt{(8-2)^2+(8-3)^2}=\sqrt{61}\approx 7.81$, $d(P3, C2)=0 \rightarrow P3 \rightarrow C2$
- $d(P4, C1) = \sqrt{(9-2)^2+(8-3)^2}=\sqrt{74}\approx 8.60$, $d(P4, C2)=\sqrt{(9-8)^2+(8-8)^2}=1 \rightarrow P4 \rightarrow C2$

So clusters are:
C1: {P1, P2}, C2: {P3, P4}

Step 3: Update Centroids
- New C1 = mean of P1 and P2 = ((2+3)/2, (3+3)/2) = (2.5, 3)
- New C2 = mean of P3 and P4 = ((8+9)/2, (8+8)/2) = (8.5, 8)

Step 4: Repeat Assignment
Recalculate distances with new centroids:
- P1 closer to C1
- P2 closer to C1
- P3 closer to C2
- P4 closer to C2

Assignments unchanged → Algorithm has converged.

Final Clusters:
C1 = {P1, P2} with centroid (2.5, 3)
C2 = {P3, P4} with centroid (8.5, 8)