K-Means Clustering

Shah Md. Arshad Rahman Ziban

What is K-Means Clustering?

Imagine you have a big box of different-colored marbles, and you want to group them based on their colors. Instead of sorting them by hand, you use an algorithm to do it for you. K-Means is an algorithm that finds groups (**clusters**) in a dataset based on similarity.

How It Works (Step by Step)

1. **Choose the Number of Clusters** K
   First, decide how many clusters you want. Let's say you choose $K = 3$ because you think your marbles can be divided into 3 groups.

2. **Select** K **Initial Centroids Randomly**
   Randomly select K points from the dataset as the initial centroids (leaders of the clusters).

3. **Assign Each Data Point to the Nearest Centroid**
   For each point in the dataset, compute the distance to all centroids and assign it to the nearest one. This forms K groups.

4. **Update the Centroids**
   Compute the mean (average) position of all points in each cluster. This new mean becomes the new centroid.

5. **Repeat Until Convergence**
   Steps 3 and 4 are repeated until centroids no longer change significantly, meaning the clusters are stable.

Mathematical Representation

1. Compute the Euclidean distance between each point x_i and each centroid c_k:

$$d(x_i, c_k) = \sqrt{\sum_{j=1}^{n} (x_{ij} - c_{kj})^2}$$

2. Assign each point x_i to the cluster with the nearest centroid:

$$C_k = \{x_i \mid d(x_i, c_k) \leq d(x_i, c_j) \text{ for all } j \neq k\}$$

3. Update each centroid to the mean of its assigned points:

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

4. Repeat the process until centroids no longer change significantly.

Applications of K-Means

- **Customer Segmentation** in marketing (grouping customers with similar purchasing behavior)
- **Document Clustering** (grouping articles by topics)
- **Image Compression** (reducing colors in an image by clustering similar colors)