# K-Means Clustering

## Introduction

K-Means is a popular unsupervised learning algorithm used for clustering, where the goal is to group data points into k clusters based on similarity. Unlike supervised learning algorithms, K-Means does not use labeled data. Instead, it learns to partition the data by minimizing the distance between points and their assigned cluster centers (centroids). K-Means is widely used for its simplicity, efficiency, and scalability to large datasets.

## Algorithm Description

The K-Means algorithm follows an iterative refinement procedure:

1. Initialization: Choose the number of clusters $k$ and randomly initialize $k$ centroids.
2. Assignment Step: Assign each data point to the nearest centroid based on Euclidean distance.
3. Update Step: Update each centroid as the mean of the points assigned to its cluster.
4. Repeat: Continue the assignment and update steps until Convergence (i.e., the centroids no longer change or a maximum number of iterations is reached).

## Mathematical Formulation

Given a dataset $X = \{x_1, x_2, ..., x_n\} \subset \mathbb{R}^d$ and number of clusters $k$, K-Means solves:

$$\min_{C_1,...,C_k} \sum_{j=1}^{k} \sum_{x \in C_j} \|x - \mu_j\|^2 \tag{1}$$

Where:

- $C_j$ is the set of points assigned to cluster $j$.
- $\mu_j$ is the mean (centroid) of points in $C_j$.
- $\|x - \mu_j\|^2$ is the squared Euclidean distance between $x$ and its centroid.

## Choosing the Number of Clusters ($k$)

The optimal number of clusters is not always obvious. Common methods include:

- Elbow Method: Plot total within-cluster sum of squares (WCSS) versus $k$ and look for the "elbow" point.
- Silhouette Score: Measures how similar a point is to its cluster vs. other clusters.
- Gap Statistic: Compares total within intra-cluster variation to a null reference distribution.

## Key Characteristics

- Type: Unsupervised clustering algorithm.
- Output: Partition of data into $k$ clusters.
- Distance Metric: Typically uses Euclidean distance.
- Convergence: Guaranteed to converge to a local minimum.

## Strengths

- Easy to implement and computationally efficient.
- Scales well to large datasets.
- Works well when clusters are spherical and well-separated.

## Weaknesses

- Sensitive to initialization; different seeds can lead to different results.
- Assumes clusters of similar size and density.
- Not suitable for non-convex shapes or clusters with unequal variance.
- Requires specifying the number of clusters $k$ in advance.

## Applications of K-Means Clustering

- Image Compression: Reducing colors in an image by clustering pixel intensities.
- Market Segmentation: Grouping customers based on purchasing behavior.
- Document Clustering: Organizing documents into topics or themes.
- Anomaly Detection: Identifying outliers that do not belong to any cluster.

# Conclusion

K-Means is a foundational clustering algorithm in unsupervised learning, known for its simplicity and effectiveness. While it has limitations in handling complex cluster shapes and sensitive initialization, it remains a go-to method for exploratory data analysis and large-scale clustering tasks.