K-Means Clustering

1 Introduction

K-Means Clustering is a widely used unsupervised machine learning algorithm for partitioning data into clusters. Each cluster is represented by its centroid, which is the mean of all data points assigned to the cluster. The goal of K-Means is to find the clustering that minimizes the variance within each cluster.

2 Algorithm

The K-Means algorithm follows these steps:

1. **Initialization**: Choose k initial centroids randomly from the dataset.
2. **Assignment Step**: Assign each data point to the nearest centroid based on Euclidean distance.
3. **Update Step**: Compute the new centroids as the mean of all data points assigned to each centroid.
4. **Convergence Check**: Repeat the assignment and update steps until the centroids do not change significantly (i.e., the change is below a specified tolerance).

2.1 Mathematical Formulation

Objective Function: The goal of K-Means is to minimize the within-cluster sum of squares (WCSS), which is given by:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{1}$$

where:

- k is the number of clusters.
- C_i is the set of points in cluster i.
- μ_i is the centroid of cluster i.
- x is a data point.

2.2 Distance Metric

The Euclidean distance between a data point x and a centroid μ_i is computed as:

$$d(x, \mu_i) = \sqrt{\sum_{j=1}^{d} (x_j - \mu_{ij})^2} \tag{2}$$

where d is the number of dimensions, and x_j and μ_{ij} are the j-th features of the data point x and centroid μ_i, respectively.

2.3 Convergence Criteria

The algorithm converges when the centroids no longer change significantly between iterations. This can be checked using:

$$\|\mathbf{C}_{\text{new}} - \mathbf{C}_{\text{old}}\| < \text{tolerance} \tag{3}$$

where $\mathbf{C}_{\text{new}}$ and $\mathbf{C}_{\text{old}}$ are the centroids from the current and previous iterations, respectively.

3 Applications

K-Means Clustering is applied in various fields, including:

- Image compression
- Market segmentation
- Document clustering
- Anomaly detection