k-Means Clustering Algorithm

Jerome Chou

February 1, 2025

Introduction to k-Means

- k-Means is an unsupervised learning algorithm used for clustering data into k groups.
- It minimizes the variance within each cluster by iteratively updating cluster centroids.
- Common applications include image segmentation, document clustering, and anomaly detection.

Mathematical Formulation of k-Means

- Given a dataset $X = \{x_1, x_2, \ldots, x_n\}$ and a predefined number of clusters k, the objective is to minimize:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} \mathbb{1}(c_i = j) \|x_i - \mu_j\|^2 \tag{1}$$

- Where:
  - c_i is the cluster assignment of data point x_i.
  - μ_j is the centroid of cluster j.
  - $\mathbb{1}(c_i = j)$ is an indicator function that is 1 if x_i belongs to cluster j, otherwise 0.

k-Means Algorithm Steps

- Initialize k cluster centroids randomly.
- Assign each data point to the nearest centroid:

$$c_i = \arg\min_j \|x_i - \mu_j\| \tag{2}$$

- Update each centroid to be the mean of the assigned points:

$$\mu_j = \frac{\sum_{i=1}^{n} \mathbb{1}(c_i = j) x_i}{\sum_{i=1}^{n} \mathbb{1}(c_i = j)} \tag{3}$$

- Repeat until centroids converge.

Machine Learning Applications of k-Means

- Customer segmentation for marketing.
- Image segmentation and pattern recognition.
- Anomaly detection in network security.
- Document clustering in Natural Language Processing (NLP).