K-means Clustering Algorithm

HIMANSHU KESARVANI

06/10/2024

1 Introduction

K-means is a popular unsupervised learning algorithm used for clustering, where the aim is to partition a set of n data points into k clusters. Each data point belongs to the cluster with the nearest mean, serving as the cluster's centroid. K-means minimizes the within-cluster sum of squares (WCSS), which is also known as the inertia or variance.

2 Mathematical Formulation

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of n data points, where each data point x_i is a d-dimensional vector: $x_i \in \mathbb{R}^d$. The goal of K-means is to divide X into k clusters, denoted as $C_1, C_2, \ldots, C_k$, such that the following objective function is minimized:

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

Here, μ_i represents the centroid of cluster C_i, and $\|x - \mu_i\|^2$ is the squared Euclidean distance between a data point x and the centroid μ_i.

The K-means algorithm iteratively updates the centroids and assigns data points to clusters based on the following steps.

3 Steps of the K-means Algorithm

3.1 Step 1: Initialization

Randomly initialize k centroids $\mu_1, \mu_2, \ldots, \mu_k$. These centroids can be initialized randomly from the data points or by using methods such as the K-means++ algorithm, which improves the initialization to avoid poor clustering.

3.2 Step 2: Assign Data Points to Clusters

For each data point x_j, assign it to the cluster with the nearest centroid. Mathematically, data point x_j is assigned to cluster C_i if:

$$C_i = \{x_j : \|x_j - \mu_i\|^2 \leq \|x_j - \mu_l\|^2 \text{ for all } l = 1, \ldots, k\}$$

This step minimizes the distance between each data point and its closest centroid.

3.3 Step 3: Update Centroids

After assigning all data points to clusters, recompute the centroid of each cluster as the mean of the points in that cluster:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

where $|C_i|$ is the number of points in cluster C_i.

3.4 Step 4: Convergence Check

Repeat Steps 2 and 3 until convergence. Convergence occurs when the centroids no longer change significantly or when the assignments of data points to clusters remain unchanged. Formally, you can stop the algorithm when:

$$\sum_{i=1}^{k} \|\mu_i^{(t)} - \mu_i^{(t-1)}\|^2 < \epsilon$$

where ϵ is a small threshold and $\mu_i^{(t)}$ and $\mu_i^{(t-1)}$ represent the centroids at iteration t and $t-1$, respectively.

4 Complexity and Limitations

The time complexity of the K-means algorithm is $O(n \cdot k \cdot d \cdot t)$, where:

- n is the number of data points,
- k is the number of clusters,
- d is the dimensionality of each data point,
- t is the number of iterations until convergence.

4.1 Limitations

- K-means is sensitive to the initial choice of centroids. Poor initialization can lead to suboptimal clustering.
- The algorithm assumes clusters to be spherical and equally sized, which might not be the case for real-world data.
- K-means requires the number of clusters, k, to be predefined.
- The algorithm converges to a local minimum, which might not be the global minimum.

5 Conclusion

The K-means algorithm is a simple yet powerful clustering technique that groups data into k clusters based on the nearest centroid. Although it has limitations such as sensitivity to initialization and the assumption of spherical clusters, it remains widely used due to its efficiency and ease of implementation. Proper initialization techniques, such as K-means++, can mitigate some of the algorithm's weaknesses.