KMeans++ From Scratch

Based on Professor Galletti's Medium Article

Objective

- **Goal**: Partition a dataset into k clusters such that similar data points are grouped together.
- **Cost Function**: Minimize the sum of variances within each cluster:

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

  where C_i is the set of points in cluster i and μ_i is the mean of C_i.

Standard KMeans Algorithm

1. Initialize k cluster centers randomly.
2. Assign each data point to the nearest cluster center.
3. Recompute the centers as the mean of the assigned points.
4. Repeat steps 2 and 3 until convergence (no change in assignments).

Convergence

- Each reassignment reduces the cost function.
- Since there are a finite number of possible partitions, the algorithm must converge.

Initialization Sensitivity

- Random initialization can lead to suboptimal clustering.
- Poor initial centers may cause the algorithm to converge to a local minimum.

KMeans++ Initialization

1. Choose the first center c_1 uniformly at random from the data points.
2. For each data point x, compute $D(x)$, the distance to the nearest already chosen center.
3. Choose the next center c_i from the data points with probability proportional to $D(x)^2$.
4. Repeat steps 2 and 3 until k centers have been chosen.

Advantages of KMeans++

- Provides a smarter initialization leading to better clustering results.
- Reduces the likelihood of poor clustering due to unfortunate initial center choices.