Gradient Descent for K-Means Clustering

The K-means clustering algorithm aims to partition n data points into K clusters in such a way that each data point belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The goal is to minimize the within-cluster sum of the squares (WCSS), also known as the inertia.

Here's a brief outline of how K-Means clustering works:

1) **Initialization**: Randomly select K initial centroids (mean points of clusters).
2) **Assignment step**: Assign each data point to the nearest centroid. This step can be represented as

$$\boxed{C_i = \{x_j : \|x_j - \mu_i\|^2 \le \|x_j - \mu_l\|^2 \ \forall l, 1 \le l \le K\}}$$

where C_i is the set of points assigned to cluster i, x_j is a data point, and μ_j is the centroid of cluster i.

3) **Update step**: Update the centroids to be the mean of the points in the cluster. This step can be represented as:

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

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

4) Repeat : Repeat steps 2 and 3 until the centroids no longer change (or the changes are below a small threshold), indicating the convergence.

• Objective Function : The objective is to minimize the within-cluster sum of the square (WCSS) defined as

$$\boxed{J = \sum_{i=1}^{K} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2}$$

• Gradient Descent Analogy : While K-Means is not gradient descent in the traditional sense (it's more of a coordinate descent algorithm), the update of the centroids can be seen as a step towards minimizing the objective function.

To derive the gradient descent update rule for the centroids, consider the partial derivative of J with respect to μ_i :

$$\frac{\partial J}{\partial \mu_i} = \frac{\partial}{\partial \mu_i} \sum_{i=1}^{K} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$

Focusing on one cluster i and one centroid μ_i :

$$\frac{\partial J}{\partial \mu_i} = \sum_{x_j \in C_i} \frac{\partial}{\partial \mu_i} \|x_j - \mu_i\|^2$$

expanding the squared team:

$$\|x_j - \mu_i\|^2 = (x_j - \mu_i)^T (x_j - \mu_i)$$

Taking the derivative with respect to μ_i:

$$\frac{\partial}{\partial \mu_i} \|x_j - \mu_i\|^2 = \frac{\partial}{\partial \mu_j} (x_j - \mu_i)^T (x_j - \mu_i)$$

$$= -2(x_j - \mu_i)$$

Summing over all points x_j in cluster i:

$$\frac{\partial J}{\partial \mu_i} = \sum_{x_j \in C_i} -2(x_j - \mu_i) = -2 \sum_{x_j \in C_i} (x_j - \mu_i)$$

Setting this gradient to zero to find the min:

$$0 = -2 \sum_{x_j \in C_i} (x_j - \mu_i)$$

$$\Rightarrow \sum_{x_j \in C_i} x_j = \sum_{x_j \in C_i} \mu_i$$

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