K-Means Clustering

Notation:

K	number of clusters
$\mathbb{C}_i$	cluster i, $i = 1, 2, \ldots, K$
$\mathbf{c}_i$	centroid of ith cluster, $i = 1, 2, \ldots, K$
d	number of attributes
N_i	number of points in ith cluster, $i = 1, 2, \ldots, K$
$\mathbf{x}_{ij} \in \mathbb{R}^d$	jth point in ith cluster, $j = 1, 2, \ldots, N_i$, $i = 1, 2, \ldots, K$

The Sum Square Error (SSE) of a clustering of points is the sum of the squared distances from each point, $\mathbf{x}_{ij}$ to its corresponding cluster centroid, $\mathbf{c}_i$. SSE is also called the *inertia* of the clustering.

46 Definition (Sum Square Error (SSE))

$$\text{SSE } = \sum_{i=1}^{K} \underbrace{\sum_{j=1}^{N_i} \text{dist}(\mathbf{x}_{ij}, \mathbf{c}_i)^2}_{\text{SSE}_i}$$

where $\text{dist}(\mathbf{x}_{ij}, \mathbf{c}_i)$ is the distance between $\mathbf{x}_{ij}$ and $\mathbf{c}_i$.

Cluster SSE_i is minimized if the centroid is chosen to be the mean of the cluster points.

47 Theorem (Cluster Means Minimize Cluster SSE)
The centroid $\mathbf{c}_i$ that minimizes SSE_i is given by

$$\mathbf{c}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_{ij}, \; i = 1, 2, \ldots, K.$$

Solution: In order to minimize SSE_i, we will set $\dfrac{d}{d\mathbf{c}_i}\text{SSE}_i = 0$ and solve for $\mathbf{c}_i$. We will also use the fact that $\text{dist}(\mathbf{x}_{ij}, \mathbf{c}_i)^2 = (\mathbf{x}_{ij} - \mathbf{c}_i)^\top (\mathbf{x}_{ij} - \mathbf{c}_i)$. We have that

$$\begin{aligned}
\frac{d}{d\mathbf{c}_i}\text{SSE}_i &= \frac{d}{d\mathbf{c}_i}\sum_{j=1}^{N_i} \text{dist}(\mathbf{x}_{ij}, \mathbf{c}_i)^2 \\
&= \sum_{j=1}^{N_i} \frac{d}{d\mathbf{c}_i}\text{dist}(\mathbf{x}_{ij}, \mathbf{c}_i)^2 \\
&= \sum_{j=1}^{N_i} \frac{d}{d\mathbf{c}_j}(\mathbf{x}_{ij} - \mathbf{c}_i)^\top (\mathbf{x}_{ij} - \mathbf{c}_i) \\
&= \sum_{j=1}^{N_i} 2(\mathbf{x}_{ij} - \mathbf{c}_i)(-1)
\end{aligned}$$

Setting $\frac{d}{d\mathbf{c}_i}\text{SSE}_i = 0$ implies

$$\begin{aligned}
\sum_{j=1}^{N_i} 2(\mathbf{x}_{ij} - \mathbf{c}_i)(-1) &= 0 \\
\sum_{j=1}^{N_i} (\mathbf{x}_{ij} - \mathbf{c}_i) &= 0 \\
\sum_{j=1}^{N_i} \mathbf{x}_{ij} - \sum_{j=1}^{N_i} \mathbf{c}_i &= 0 \\
\sum_{j=1}^{N_i} \mathbf{x}_{ij} - N_i\mathbf{c}_i &= 0 \\
\mathbf{c}_i &= \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{x}_{ij}
\end{aligned}$$

48 Theorem (K-Means Clustering Always Converges)
K-Means Clustering always converges to a (local) minimum SSE.

Solution: We will use the following fact from Calculus II:

> *A decreasing sequence bounded below must converge.*

In particular, we will show that:

(I) SSE ≥ 0, so it is bounded below by 0.

(II) K-Means always decreases SSE.

To see why (I) must be true, observe that SSE is a sum of non-negative terms dist$(\mathbf{x}_{ij}, \mathbf{c}_i)^2$, so we must have SSE ≥ 0.

To see why (II) is true, observe that the K-Means algorithm has two steps:

(i) Update Centroids

(ii) Update Clusters

Next we show that each step decreases SSE. The Update Centroids step computes new centroids to equal the means of the points in each cluster. We have already shown that these means minimize cluster SSEs. Since SSE is the sum of cluster SSEs, the overall SSE must be minimized also. The Update Clusters reassigns points to their nearest centroid, i.e. a point is reassigned only if

$$\text{dist}(\mathbf{x}_{ij}, \mathbf{c}_{\text{new}})^2 < \text{dist}(\mathbf{x}_{ij}, \mathbf{c}_{\text{old}})^2.$$

Since SSE is the sum of dist$(\mathbf{x}_{ij}, \mathbf{c}_i)^2$ terms, SSE can only decrease after this step is completed.

Therefore, since SSE is decreasing and bounded below by 0, it must converge.

Note, the K-Means algorithm is guaranteed to find a local minimum of SSE. Running K-Means multiple times from random starting points can produce a smaller SSE.