consulted solution

1. Let $X = x_1, x_2, \ldots, x_m$ be data pts, & let k be the # of clusters.

k-means alg:

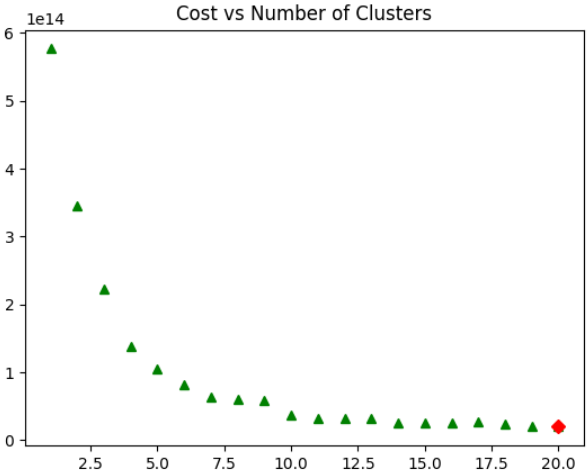
1. Randomly initialize k cluster centers, $\mu_1, \mu_2, \ldots \mu_k$, in the feature space.
2. Calculate the distance btwn each data pt & cluster centers
3. Assign each data pt. to the cluster center c whose distance between this data pt is the min of all cluster centers, namely,

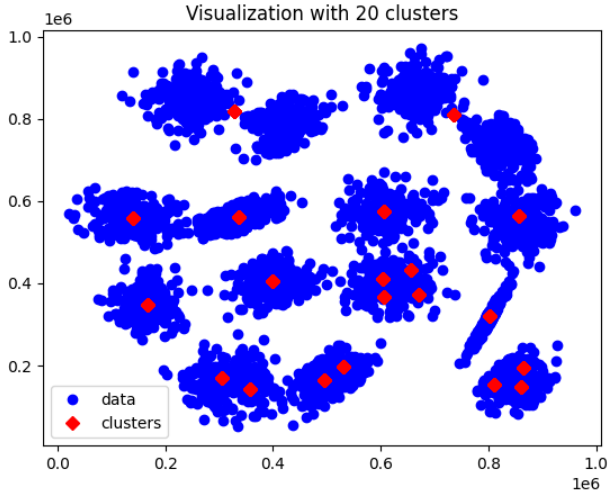
$$c_i = \arg\min_j \|x_i - \mu_j\|^2$$

4. Update each cluster center to be $\mu_j = \frac{\sum_{i=1}^{m} 1\{c_i = j\} x_i}{\sum_{i=1}^{m} 1\{c_i = j\}}$

5. Repeat step 2-4 until convergence or exhausted.

Objective cost fxn defined as: $J(c, \mu) = \sum_{i=1}^{m} \|x_i - \mu_{c_i}\|^2$


Cost vs Number of Clusters
1e14
6
5
4
3
2
1
0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Visualization with 20 clusters
1e6
1.0
0.8
0.6
0.4
0.2
data
clusters
0.0
0.2
0.4
0.6
0.8
1.0
1e6