2. **(10 points)** In class, we analyzed the per-iteration complexity of k-means. Here, we will prove that the k-means algorithm will terminate in a finite number of iterations. Consider a data set $X = \{x_1, \ldots, x_n\} \in \mathbb{R}^d$.

   a. Show that the k-means loss function can be re-written in the form:

   $$F(\eta, \mu) = \sum_{i=1}^{n} \sum_{j=1}^{k} \eta_{ij} \|x_i - \mu_j\|^2$$

   where $\eta = (\eta_{ij})$ is a suitable binary matrix (with 0/1 entries). Provide a precise interpretation of η.

   b. Show that each iteration of Lloyd's algorithm can only decrease the value of F.

   c. Conclude that the algorithm will terminate in no more than T iterations, where T is some finite number. Give an upper bound on T in terms of the number of points n.

2a) S.T K-means loss function can be re-written in the form:

$$F(\eta, \mu) = \sum_{i=1}^{n} \sum_{j=1}^{K} \eta_{ij} \|x_i - \mu_j\|^2$$

where η is a binary matrix

Ans: $x_i \in \mathbb{R}^d$ where $i \in \{1, \ldots n\}$ and we are supposed to find K clusters with centers (means) $\mu = \{\mu_1, \mu_2 \ldots \mu_K\}$.

Consider M as a matrix of cluster centroids $\mu_i \in \mathbb{R}^d$

$\eta \in \mathbb{R}^{n \times K}$ → matrix of binary indicator variables

$$\eta_{ij} \begin{cases} 1, \text{ if } x_j \in C_i \\ 0, \text{ otherwise} \end{cases}$$

where $C_i \in \{C_1, C_2 \ldots C_K\}$ → This is a set of clusters.

$i \rightarrow 1$ to n (data)

$j \rightarrow 1$ to K (cluster)

$d \rightarrow$ dimensions

Some properties and interpretations of the binary indicator matrix

- Clusters $C_1, C_2 \ldots C_K$ have centroids $\mu_1, \mu_2 \ldots \mu_K$, so every row of η will have a single 1 and $K-1$ zeros. So every single row of η will sum to 1.

$$\eta = \begin{matrix} & K_1 & K_2 & \ldots & K_n \\ 1 \\ 2 \\ i \\ n \end{matrix} \begin{bmatrix} 0 & 1 & \ldots & 0 \end{bmatrix} \Big\} n \qquad \underbrace{\qquad\qquad}_{K}$$

Every row here represents to which cluster 'K' does the data point n belong to.
In this example data point 1 belongs to the cluster K_2.

$$\sum_j \eta_{ij} = 1$$

- The column sum indicates the number of elements per cluster.

$$\sum_i \eta_{ij} = n_j = |C_j|$$

- $\eta_{ij} \in \{0, 1\}$ and every row of η contains a single entry for 1, $\therefore$ the columns of η are pairwise perpendicular.

$$\eta_{ij}\,\eta_{ij'} = \begin{cases} 1, & \text{if } j = j' \\ 0, & \text{otherwise} \end{cases}$$

- Because of this matrix $\eta\eta^T$ is a diagonal matrix

$$(\eta\eta^T)_{jj'} = \sum_j (\eta)_{ij}(\eta)^T_{ji} = \sum_i \eta_{ij}\,\eta_{ij'} = \begin{cases} n_j, & \text{if } j = j' \\ 0, & \text{otherwise} \end{cases}$$

- Consider $\sum_{i,j} \eta_{ij} \|x_i - \mu_j\|^2 = \sum_{i,j} \eta_{ij} \left[x_i^T x_i - 2 x_i^T \mu_j + \mu_j^T \mu_j \right]$

$$= \underbrace{\sum_{i,j} \eta_{ij} x_i^T x_i}_{(1)} - 2 \underbrace{\sum_{i,j} \eta_{ij} x_i^T \mu_j}_{(2)} + \underbrace{\sum_{i,j} \eta_{ij} \mu_j^T \mu_j}_{(3)}$$

(1)

$$\sum_{i,j} \eta_{ij} \|x_i\|^2$$

$$\sum_i \|x_i\|^2$$

$$= X^T X$$

(2)

$$\sum_{i,j} \eta_{ij} \sum_p x_{pi} \mu_{pj}$$

$$\sum_{i,p} x_{pi} \sum_j \mu_{pj} \eta_{ij}$$

$$\sum_{i,p} x_{pi} (M\eta)_{pj}$$

$$\sum_i \sum_p (X^T)_{ip} (M\eta)_{pj}$$

$$\sum_i (X^T M \eta)_{ii}$$

$$= X^T M \eta$$

(3)

$$\sum_{i,j} \eta_{ij} \|\mu_j\|^2$$

$$= \sum_j \|\mu_j\|^2 \cdot n_j$$

This can be written as, $(\eta^T M^T M \eta)$, because

$$\eta^T M^T M \eta = M^T M \eta \eta^T$$

$$= \sum_j (M^T M \eta \eta^T)_{jj}$$

$$= \sum_j \sum_p (M^T M)_{jp} (\eta \eta^T)_{pj}$$

$$= \sum_j (M^T M)_{jj} (\eta \eta^T)_{jj}$$

$$= \sum_j \|\mu_j\|^2 n_j$$

So, from the above simplification, we get

$$\sum_{i,j} \eta_{ij} \|x_i - \mu_j\|^2 = XX^T - 2(X^T M \eta) + (\eta^T M^T M \eta)$$

We need to minimize this loss function, so differentiate w.r.t M.

$$\frac{\partial}{\partial M}\left[XX^T - 2(X^T M \eta) + (\eta^T M^T M \eta)\right] = 0$$

$$2\left[M\eta\eta^T - X\eta^T\right] = 0$$

$$\boxed{M^* = X\eta^T(\eta\eta^T)^{-1}}$$

- So, from this derivation we can see that the matrix M^*, which is a matrix of centroids of the K clusters can be obtained by computing the above equation. M^* is the optimal matrix of mean values of the cluster centroids.
- So, when the above stated precise interpretation of η is used, we can obtain a closed form expression for the optimal cluster center values.

26) b. Show that each iteration of Lloyd's algorithm can only decrease the value of F.

Ans: We wish to minimize the loss function,

$$F(\eta, \mu) = \sum_{i=1}^{n} \sum_{j=1}^{K} \|x_i - \mu_j\|^2$$

Let $z_i \in \arg\min_{j \in \{1,\dots,K\}} \|x_i - \mu_j\|^2$ for every data point x_i. The K-means algorithm iterates between updating z_i, which is basically the step of assigning a cluster to a data point (Assignment step) and updating the cluster centers $\mu_j = \frac{1}{|S_{ij}|} \sum_{x_i \in S_j} x_i$

where S_1 S_2 .... S_K are the clusters (Refitting step). The algorithm stops when there is no change in the assignment step.

Let $z = \{z_1, z_2 \dots z_n\}$ denote the cluster assignment for n points.

— In the assignment step, $L(\mu)$ can be written as

$$L(\mu, z) = \sum_{i=1}^{n} \|x_i - \mu_{z_i}\|^2$$

For a data point x_i, let z_i be the cluster assignment from the previous iteration and z_i^* be the new assignment obtained as:

$$z_i^* = \arg\min_{j \in \{1 \dots K\}} \|x_i - \mu_j\|^2$$

let z^* denote the new cluster assignment for all data points. The change in loss function after the assignment step is given by:

$$L(\mu, z^*) - L(\mu, z) = \sum_{i=1}^{n} \left[\|x_i - \mu_{z_i^*}\|^2 - \|x_i - \mu_{z_i}\|^2 \right] \leq 0$$

The inequality holds true by the rule using which z_i^* is calculated, i.e to assign x_i to the nearest cluster.

— In the refitting step, $L(\mu)$ is written as

$$L(\mu, z) = \sum_{j=1}^{K} \left(\sum_{i : z_i = j} \|x_i - \mu_j\|^2 \right)$$

For the j^{th} cluster, $\mu_j \rightarrow$ cluster center from previous iteration

$\mu_j^* \rightarrow$ new cluster center

$$\mu_j^* = \frac{1}{|\{i : z_i = j\}|} \sum_{i : z_i = j} x_i$$

Let μ^* denote the new cluster center for all the K clusters. The change in loss function after this refitting step will be,

$$L(\mu^*, z) - L(\mu, z) = \sum_{j=1}^{K} \left[\sum_{i : z_i = j} \|x_i - \mu_j\|^2 - \sum_{i : z_i = j} \|x_i - \mu_j\|^2 \right] \leq 0$$

This inequality holds because the update rule of μ_j^* minimizes this quantity.

2c) c. Conclude that the algorithm will terminate in no more than T iterations, where T is some finite number. Give an upper bound on T in terms of the number of points n.

Ans: Running time for 1 iteration, we get $O(nK)$ distances in each iteration in time $O(ndK)$. Where n- no of input points, K is the no of clusters and d- dimensions. The worst case running time to compute a K-clustering of n points in $\mathbb{R}^d$ is upper bounded by $O(ndK \cdot T)$, where T is the no of iterations. The number of iterations of the algorithm is bounded by the number of partitionings of the data points induced by K centers. This number can be bounded by $O(n^{dK^2})$ because for K cluster

centers, we can move each of the $O(K^2)$ bisectors such that they coincide with d linearly independent points without changing the partition.

For $d = 1$, $K < 5$ → upper bound is $O(n)$ iterations

For $d = 1$, any K → upper bound is $O(n\Delta^2)$ iterations.

Δ → ratio b/w diameter and smallest pairwise distance of input points.