Hard Clustering Algorithms

⊛ Consider a set of unlabeled data points:

$$\boxed{\mathcal{X} = \{ \underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N \} \text{ where } \underline{x}_i \in \mathbb{R}^{\ell}, \forall j \in [N]}$$

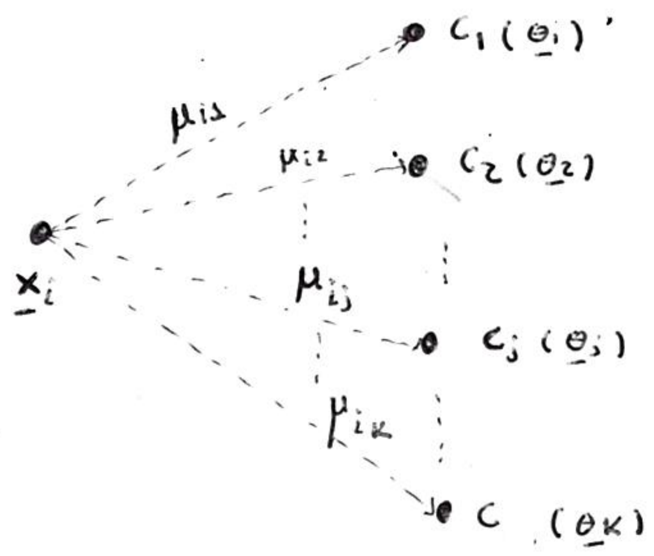
⊛ The clustering provides a partitioning of the original dataset into M clusters such that:

$$\boxed{\mathcal{X} = \bigcup_{j=1}^{K} C_j \quad \text{with} \quad C_r \cap C_m = \phi, \ \forall r \neq m}$$

⊛ **Problem Definition:**

- Assign each datapoint $\underline{x}_i$ to a unique cluster C_j such that the data points which are assigned to the j-th cluster exhibit minimum distance towards the cluster representative given by $\underline{\theta}_j$.
- The total number of clusters is K and each cluster is represented by a vector $\underline{\theta}_j \in \mathbb{R}^{\ell}$.

⊛ Let $\mu_{ij} \in \{0, 1\}$ denote the membership status of the i-th datapoint relative to the j-th cluster. $\{1 \le i \le N$ and $1 \le j \le K\}$.



Each datapoint is assigned to a single cluster.

$$\boxed{\sum_{j=1}^{K} \mu_{ij} = 1, \ \forall i \in [N]}$$

⊛ Consider that all membership values are organized in a membership matrix $\underline{\underline{M}} = [\mu_{ij}]$ such that $1 \le i \le N$ and $1 \le j \le K$.

⊛ In this setting, we may define the clustering problem as an optimization problem where the objective/cost function is given as:

$$\boxed{\begin{aligned} &\min_{(\underline{\underline{\Theta}}, \underline{\underline{M}})} J(\underline{\underline{\Theta}}; \underline{\underline{M}}) = \sum_{i=1}^{N} \sum_{j=1}^{K} \mu_{ij} \cdot \| x_i - \Theta_j \|^2 \\ &\text{s.t.} \quad \sum_{j=1}^{K} \mu_{ij} = 1, \quad \forall i \in [N] \qquad \boxed{\text{OP. I}} \\ &\text{where} \quad \underline{\underline{\Theta}} = [\underline{\Theta}_1, \underline{\Theta}_2, \ldots, \underline{\Theta}_K] \end{aligned}}$$

S.O.S: The objective function $J(\underline{\underline{\Theta}}; \underline{\underline{M}})$ is not differentiable since it's not continuous given that $\mu_{ij} \in \{0, 1\}$.

⊛ This is a hard mixed-integer (combinatorial) optimization problem since the number of possible clusterings of N datapoints into K clusters will be given by the so-called Stirling numbers of the second type:

$$\boxed{S(N, K) = \frac{1}{K!} \sum_{j=0}^{K} (-1)^{K-j} \binom{K}{j} j^N}$$

⊛ Therefore, we cannot use straightforward optimization techniques.

Heuristic Optimization Scheme:

(α): Considering that $\underline{\theta}_j$ with $1 \leq j \leq K$ are fixed.

Since for each vector $\underline{x}_i$ only one μ_{ij} is 1 and all the others are $\emptyset$, it is straightforward to see that $J(\underline{\theta};\underline{\underline{\mu}})$ is minimized by if we assign each $\underline{x}_i$ to each closest cluster-representative:

$$\boxed{\mu_{ij} = \begin{cases} 1, & j = \arg\min_{r \in [K]} \|\underline{x}_i - \underline{\theta}_r\|^2; \\ \emptyset, & \text{otherwise} \end{cases} \quad \forall i \in [N] \qquad [A]}$$

(β): Considering that μ_{ij}'s are fixed with $1 \leq i \leq N$ and $1 \leq j \leq K$.

In this setting, we need to evaluate the F.O.Cs with the respect to the parameters $\underline{\theta}_r$, $\boxed{\forall r \in [K]}$ as:

$$\frac{\partial J}{\partial \underline{\theta}_r} = \underline{\mathbb{0}} \in \mathbb{R}^l \ \text{or} \ \frac{\partial J}{\partial \underline{\theta}_r} = \frac{\partial}{\partial \underline{\theta}_r}\left\{\sum_{i=1}^{N}\sum_{j=1}^{K}\mu_{ij}\|\underline{x}_i - \underline{\theta}_j\|^2\right\} = \underline{\mathbb{0}}$$

⊛ Since we are differentiating w.r.t $\underline{\theta}_r$, all the other $\underline{\theta}_m$'s with $m \neq r$ as we do not contribute to the differentiation process. They are constant as far as the partial differentiation process is concerned.

⊛ Thus, we may write that:

$$\frac{\partial J}{\partial \underline{\theta}_r} = \sum_{i=1}^{N}\sum_{j=1}^{K}\frac{\partial}{\partial \underline{\theta}_r}\left[\mu_{ij}\|\underline{x}_i - \underline{\theta}_j\|^2\right] = \sum_{i=1}^{N}\frac{\partial}{\partial \underline{\theta}_r}\left[\mu_{ir}\|\underline{x}_i - \underline{\theta}_r\|^2\right] = \underline{\mathbb{0}}$$

$\forall r \in [K]$ →

$$\circledast \quad \sum_{i=1}^{N} \mu_{ir} \frac{\partial}{\partial \underline{\theta}_r} \| \underline{x}_i - \underline{\theta}_r \|^2 = \underline{0} \Rightarrow \sum_{i=1}^{N} \mu_{ir} (-2 \underline{x}_i + \overset{2}{\cdot} \underline{\theta}_r) = \underline{0} \Rightarrow$$

$$\| \underline{x}_i - \underline{\theta}_r \|^2 = (\underline{x}_i - \underline{\theta}_r)^T (\underline{x}_i - \underline{\theta}_r) = (\underline{x}_i^T - \underline{\theta}_r^T) (\underline{x}_i - \underline{\theta}_r) =$$

$$= \underline{x}_i^T \underline{x}_i - \underline{x}_i^T \underline{\theta}_r - \underline{\theta}_r^T \underline{x}_i + \underline{\theta}_r^T \underline{\theta}_r =$$

$$= \underline{x}_i^T \underline{x}_i - 2 \underline{x}_i^T \underline{\theta}_r + \underline{\theta}_r^T \underline{\theta}_r .$$

$$\frac{\partial}{\partial \underline{\theta}_r} \left[\underline{x}_i^T \underline{x}_i - 2 \underline{x}_i^T \underline{\theta}_r + \underline{\theta}_r^T \underline{\theta}_r \right] = -2 \underline{x}_i + 2 \underline{\theta}_r$$

$$\frac{\partial \underline{x}^T \underline{x}}{\partial \underline{x}} = \frac{\partial \underline{x}^T \underline{I} \underline{x}}{\partial \underline{x}} = (I + I^T) \underline{x} = 2 I \underline{x} = 2x$$

$$\circledast \quad -\cancel{2} \sum_{i=1}^{N} \mu_{ir} \underline{x}_i + \cancel{2} \sum_{i=1}^{N} \mu_{ir} \underline{\theta}_r = \underline{0} \Rightarrow$$

$$\underline{\theta}_r \cdot \sum_{i=1}^{N} \mu_{ir} = \sum_{i=1}^{N} \mu_{ir} \underline{x}_i \Rightarrow \boxed{\underline{\theta}_r = \frac{\sum_{i=1}^{N} \mu_{ir} \underline{x}_i}{\sum_{i=1}^{N} \mu_{ir}} \quad [B]}$$

Eqs. (A) and (B) can pave the way in order to formulate the Generalized Hard Clustering Scheme.

Generalized Hard Clustering Algorithmic Scheme

Step 1: Choose $\underline{\theta}_j(0)$ as initial estimates for $\underline{\theta}_j$, $\forall j \in [K]$.

Step 2: $t = 0$

Step 3: Repeat:

Step 3.1: $\forall i \in [N]$: Update $\underline{\mu}_i$ according to:

$$\boxed{\mu_{ij}(t) = \begin{cases} 1, & j = \arg\min_{r \in [K]} \|\underline{x}_i - \underline{\theta}_r(t)\|^2; \\ 0, & \text{otherwise.} \end{cases} \quad (1)}$$

Step 3.2: $t = t+1$.

Step 3.3: $\forall j \in [K]$: Update $\underline{\theta}_j$ according to:

$$\boxed{\underline{\theta}_j(t) = \frac{\sum_{i=1}^{N} \mu_{ij}(t-1)\underline{x}_i}{\sum_{i=1}^{N} \mu_{ij}(t-1)} \quad (2)}$$

Until $\|\underline{\underline{\theta}}(t) - \underline{\underline{\theta}}(t-1)\| < \varepsilon$

(A) Consider $\underline{\underline{\theta}} = [\underline{\theta}_1, \underline{\theta}_2, \dots, \underline{\theta}_K]$ and $\underline{\underline{\mu}} = \begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \\ \vdots \\ \underline{\mu}_N \end{bmatrix}$

where $\underline{\theta}_j$ the j-th cluster representative and $\underline{\mu}_i$ the i-th datapoint membership vector.