Assignment #1

IP Formulation for k-means problem

Consider n data points $\mathbf{x}_i \in \mathbb{R}^d$. We are asked to find k centroids such that the sum of distances between the points assigned to a certain centroid and the centroid is minimized.

Consider all possible clusters of the data $C_1, \dots, C_P \in \{0, 1\}^n$, where a 0 at position i indicates that $\mathbf{x}_i$ is included in the cluster. Then, calculate $c(C_i)$ as the sum of distances between the cluster's centroid (calculated as the average of the point assigned to the cluster) and the point assigned to the cluster. Introduce a variable for each cluster $y_j \in \{0, 1\}$ such that if the cluster is selected it is set as 1, otherwise as 0.

Then formulate the problem as follows:

$$
\begin{aligned}
\min \sum_{j=1}^{P} c(C_j) y_j & \\
\text{Such that } \sum_{j=1}^{P} y_j C_{ji} = 1 \ \forall i & \qquad \text{Each element belongs to a chosen cluster} \\
\sum_{j=1}^{P} y_j = k & \qquad \text{Choose } k \text{ clusters} \\
0 \leq y_j \leq 1, y_j \text{ integer } \forall j &
\end{aligned}
$$