Indian Institute of Technology Jodhpur
CSL2010: Introduction to Machine Learning
Lab 6&7, Due Date: Oct 5 , 2025, Max Marks: 70+30 for Viva

1. Consider the dataset $\mathcal{D} = \{x_1, \ldots, x_n\}$ given in the file `dataset.txt`. Here, each point $x_i \in \mathbb{R}^2$ is described by two features. The last column of the sheet contains the ground-truth cluster labels. However, these labels will only be used to measure the performance and will not be used anywhere in the training. Now, do the following:

   k-Means Clustering Algorithm

   (a) Implement the k-Means algorithm to cluster the points into two clusters. You can use any two data points from the dataset $\mathcal{D}$ uniformly at random to initialize the cluster centres.
   [Compulsory to implement in the lab 6].

   (b) Plot the obtained clusters using the k-Means algorithm with different colors for each cluster.

   (c) In order to evaluate the performance of the k-Means algorithm, find the percentage of the points for which the estimated cluster labels are correct.

   Spectral Clustering Algorithm:

   (a) Use the similarity function $W_{ij} = \begin{cases} e^{-\frac{\|x_i - x_j\|_2^2}{\sigma}}, & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases}$ to define the adjacency matrix $W \in \mathbb{R}^{n \times n}$. Choose the appropriate value of σ. Define the degree matrix $D \in \mathbb{R}^{n \times n}$ where $D_{rr} = \sum_{i=1}^{n} W_{ri}$ and then the Laplacian matrix as $L = D - W$. Now, find the eigenvalue decomposition of the Laplacian matrix as $Lu_i = \lambda_i u_i$. Ensure that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. You can use any inbuilt function to find the eigenvalue-eigenvector decomposition of the Laplacian matrix L.
   [Compulsory to implement in the lab 7].

   (b) Let $\mathbb{H} = \begin{bmatrix} r_1^\top \\ \vdots \\ r_n^\top \end{bmatrix} \in \mathbb{R}^{n \times 2}$ be the optimal cluster assignment matrix where $r_i \in \mathbb{R}^2$ represents the spectral embedding of the data point x_i. Plot the spectral embeddings and verify that the two clusters are now linearly separable.

   (c) Perform the k-means clustering on the spectral embeddings $\{r_1, \ldots, r_n\}$. Plot the obtained clusters with different colors.

   (d) In order to evaluate the performance of the spectral clustering algorithm, find the percentage of the points for which the estimated cluster labels are correct.