Graph Mining CSF426
Lab session 7 (evaluative)
Time: 2 pm - 4pm
Date: 07-10-2023

Instructions: All questions need to be answered. **You are required to write programs in jupyter notebook and submit .ipynb**. For theoretical questions, you can type answers in the jupyter notebook itself. There is no need to create a separate text file. **[Total Marks =10]**

This lab exercise is based on Label Propagation algorithm (LPA), explained in the research paper *"Learning from labeled and unlabeled data with label propagation"* (Link)

You are provided with Iris Dataset which contains 150 data points and 3 class labels.

Objective: Predict the labels of unlabeled data points (test data) using labels of train data points.

Steps to perform the experiments are as follows:

1. Extract the features from the dataset corresponding to each data point. You will have a feature matrix (X) of size (150 * 4).
2. Using Gaussian similarity measure, construct a graph from X with σ = 0.1

$$W_{ij} = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right) if\ i! = j, \quad W_{ii} = 0\ otherwise$$

3. Compute transition matrix T where,

$$T_{ij} = P(j \rightarrow i) = \frac{w_{ij}}{\sum_{k=1}^{n} w_{kj}}$$

4. Extract the labels corresponding to each data point and convert into one-hot vector form. You will get a matrix Y of size (150 * 3)
5. Split your data into train/test into 70/30 ratio and convert the labels for test nodes to 0 vector. That is, a test node will have the corresponding label of (0 0 0) in one-hot encoded form.

   The label matrix Y will be of form $Y = \begin{bmatrix} Y_L \\ Y_U \end{bmatrix}$

6. Using the formula below, predict the labels for unlabeled test nodes.

$$Y_U = \left(I - \overline{T_{uu}}\right)^{-1} \overline{T_{ul}}\, Y_L$$

Where $\overline{T}$ *is row normalized form of transition matrix* T *i.e.* $\overline{T_{ij}} = \frac{T_{ij}}{\sum_k T_{ik}}$ *and is in the form* $\overline{T} = \begin{bmatrix} \overline{T_{ll}} & \overline{T_{lu}} \\ \overline{T_{ul}} & \overline{T_{uu}} \end{bmatrix}$ *and* Y_L *is label matrix for labeled data points (training data points)*

7. Plot and compute accuracy of original and LPA predicted labels for test data points.

8. Compare the prediction from LPA and LSA for test data points by computing accuracy, precision and recall. Please note that precision and recall are class specific and will be computed for all classes separately. Suppose three class problem with three class labels 1,2 and 3, when computing precision and recall for class 1, assume class 1 as positive and class 2 & 3 as negative.