Graph Mining CSF426
Lab session 6 (evaluative)
Time: 5 pm – 7 pm
Date: 19-09-2024
Instructor IC - Vinti Agarwal

Instructions: All questions need to be answered. **You are required to write programs in jupyter notebook and submit .ipynb on canvas**. Please rename your solutions in format <ID-NAME-LABNO>. 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 a continuation of Lab session 5, and is based on **Correct and Smooth technique** explained in the research paper ***"Combining Label Propagation and Simple Models Out-performs Graph Neural Networks"*** (Link)

NOTE: Fill the table "Table 1" at the end of this lab sheet and upload it with your jupyter notebook file.

Objective: The objective of this lab session is to learn about process of label correction using error propagation.

Dataset: You are provided with cora dataset which contains information about 2708 research papers belonging to 7 categories of Machine learning. It is a citation network/graph where each publication/paper cites one or more other papers.

The dataset contains two files:

a. "Cora.content" – contains the description of research papers in format:

<PaperID> <1433 columns of word features> <class label>

PaperID – a unique identifier of each reseach paper.

Word features - each research paper is associated 1/0-valued word vector of length 1433 based on whether a word is present/absent in the corresponding paper.

Class label - contains the one category to which a paper belongs out of seven categories of : Case_Based, Genetic_Algorithms, Neural_Networks, Probabilistic_Methods, Reinforcement_Learning, Rule_Learning, Theory

b. "cora.cites" – contains the information about the citations. Each line describes a link in format:

<cited PaperID> <citing PaperID>

The first PaperID describes the ID of paper being cited and second PaperID is the ID of paper that contains the citation or that cited the first paper.

1. **Data loading:**
    a. Reuse code from your previous lab to load cora dataset.
2. **Feature augmentation using spectral embeddings:**
    a. Perform dimensionality reduction using Spectral embedding approach (Laplacian eigenmaps) to get a k-dimensional(**k=10**) feature vector for each datapoint. Now, augment (concatenate) these new features to the original features to get a feature vector of size 1443 for each datapoint.
3. **Base predictions:**
    a. Use MLP to make base predictions on all the datapoints. Hyperparameters are as follows:
        i. Layers = 3
        ii. Neurons per layer = 64
        iii. Learning rate = 0.01
        iv. Activation function= ReLU
        v. Dropout = 0.5
        vi. Optimizer = Adam Optimizer
    b. The result will be base predictions $Z \in R^{n \times c}$ where c is the number of classes in dataset.
4. **Error propagation:** Use label spreading iteratively to propagate the errors in graph,
    a. Compute error matrix $\hat{E} \in R^{n \times c}$, $where$

$$E_L = Z_L - Y_L, \ L - Training\ datapoint$$

$$E_v = 0\ , \ v - validation\ datapoint$$

$$E_U = 0\ , \ U - unlabeled/test\ datapoint$$

And $E^{t+1} = (1 - \alpha).\ E + \alpha.S.E^{t}\ , \ where\ E^{0} = E\ and\ S = D^{\frac{-1}{2}}\ .\ A.\ D^{\frac{-1}{2}}$

$A - Adjacency\ matrix\ of\ data,\ D - Degree\ matrix\ ,\ \alpha = [0.1, 0.2, ..., 0.9]$

Use validation data to find out the best alpha value.

    b. Corrected predictions,

$$Z\_c = Z + \ \hat{E}$$

5. **Scaling:** Scale the size of errors in $\hat{E}$ to be approximately the size of errors in E.
    a. Perform autoscaling using approach: $Z_{(i,:)\ c} = \ Z_{(i,:)} + \sigma\hat{E}_{:,i} / \left|\left|\hat{E}_{(:,i)}\right|\right|_1$ for i in *U*

$$\sigma = \frac{1}{|L|}\sum_{j\in L} \left\|e_j\right\|_1$$

6. **Final prediction:** Use label spreading iteratively to make the final predictions

   a. Initialize matrix $G \in R^{n\times c}$, $such\ that$

   $$G_L = Y_L\ ,\ G_v = Z_v,\ and\ G_U = Z\_c_U$$

   $$G^{t+1} = (1 - \alpha).\ G + \alpha.S.G^t, \quad \alpha = [0.1, 0.2, ..., 0.9],$$

   b. Final classification, Y` for a node $i\in U$ is $argmax_{j\in\{1,....,c\}} Y`_{ij}$

   c. Use validation data to find the best alpha value and report the accuracy results on best alpha.

7. Create a table of the accuracies achieved with base predictor, accuracy after error propagation, accuracy after final prediction with the accuracy achieved if we use Label Spreading algorithm (LSA) on the dataset directly.

   NOTE: (you can use the code of your previous lab for LSA)

Table 1: Accuracies achieved

Method	Best alpha	Accuracy
Label propagation (LSA)		
Plain linear		
Plain linear + C&S		
Plain linear + autoscaling		
MLP		
MLP + autoscaling		