Programming Project 3

November 25, 2012

Due date: Mon, Fri Dec. 14, 2012 at 4pm.

Description: *Summary:* You are requested to provide a classification of pairs of nodes as being friends or not in the social graph. More specifically you need to solve two classification tasks:

Task 1: Identification of social links:

> You will be working on a sample from Flickr's network. Each node in the network is a Flickr user and an edge between two nodes indicates that these users are friends. To help you with this task, you will have access to the groups in which each user is a member. Using this group information you will have to predict whether the

Task 2: Identification of co-authorship links:

> We provide you with a snapshot from DBLP. This network contains information about researchers and their work. The nodes of the network are people and the edges represent coauthorship. This means that if two researchers have collaborated on at least one paper, then there exists an edge that connects them. As supporting material, we give you, for each researcher, the terms of the titles of their work and their respective frequencies.

Evaluation: The evaluation method that will be used is the AUC. This is a measure that handles unbalanced data (i.e. if there exist more true negatives than true positives). You can see how well your solution scores if you go to the Leaderboard.

Data: The files train.csv and test.csv are Comma Separated Values files. Each line corresponds to a triplet (node1, node2, label), where node1 and node2 are nodes in the network and label indicates the presence or the absence of an edge between them.

What you are asked to do is, for each pair of nodes in test.csv, to predict if there is an edge between them. You need to upload your suggested solution as

a file in which, at line i there is a 0 or a 1, depending on your prediction for the i-th pair of the test file.

The baseline solutions provided are based on the random classifier. For each given pair of nodes, we flip a coin and with probability 0.5 we assign a label to that pair.

Writeup: In addition to submitting your solution online, you need to provide us with a 2-page writeup that describes the algorithm you have implemented and the special tricks you implemented in order to make it work. Also, describe your strategy for selecting a particular algorithm and how you did your offline evaluation of the method.

Instructions: If you don't have an account to the platform already, you will need to signup. Go to `http://inclass.kaggle.com` and use your BU email. After that, search for *BU CS 565 : Classification* and you will find 2 projects. **Take part and submit solutions fo rboth of these competitions.** Download the training file, the test file and the supporting file ("features.txt") from "Data Files". You may upload your predicted solution as a *.csv* file at "Make a submission". We suggest that you don't just upload your final solution, but submit all your intermediate ones to check your standing. You can see how well your solution is ranking if you go to the "Leaderboard". Notice that the deadline on the competition page may not be accurate.

The urls for the two competitions are:
`http://inclass.kaggle.com/c/bu-cs-565-classification-1-2`
`http://inclass.kaggle.com/c/bu-cs-565-classification-2-2`

If you encounter any problems, send an email at *cmav@bu.edu*

Privacy and confidentiality: Please use the user names you have used in your previous project and write them on your project report.