SI 671/721 (Fall 2023) Data Mining: Methods and Applications

Instructor: Paramveer Dhillon

Homework 3 (Due: 2023-17-11):
Social Network Analysis (100 Points)

1 Summary

The first part of this homework contains hand-written problems for PageRank and Network Density. Please DO NOT use any Python libraries in this part. You can either write the answers on a piece of paper and take a picture, or type the answers in the Jupyter Notebook using Markdown and Latex.

For the last two parts of this homework, we will use the Amazon co-purchasing network dataset to perform social network analysis. This dataset contains various product networks including books, music CDs, DVDs, and VHS video tapes Leskovec et al. (2007). It was collected by crawling the Amazon website in March, 2003 according to `Customers Who Bought This Item Also Bought` on the Amazon website. So, if product A is always co-purchased with product B, the graph contains a directed edge from A to B.

We recommend that you use Jupyter Notebooks and Python libraries (Numpy, Sci-kit learn, Pandas, and NetworkX) for this homework.

2 Details

This homework is divided into three parts.

1. PageRank and Network Density computations (By hand).
2. Exploratory Social Network Analysis.
3. Predicting Review Rating using features derived from Network Properties.

2.1 Part 1: Network Calculations (By Hand) [20 Points]

This part of the homework is designed to help you familiarize yourself with the calculations for PageRank and Network Density. You should not use any Python libraries for this part of the homework. You can either submit your answers as a PDF document or write your answers in Markdown/Latex cells in a Jupyter Notebook.

1. **PageRank Computation** [10 Points]

   Please use the network graph shown in Figure 1 for these computations. All edges have a weight of 1.

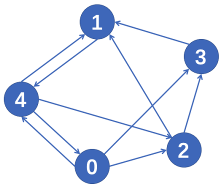

Figure 1: Network Graph (PageRank computation)

(a) Please write down the equations for each node's PageRank. [2 points]

(b) Please compute and write down the stochastic adjacency matrix of the network. [2 points]

(c) Calculate the PageRank of each node using the power iteration method for 2 iterations and show your work. [4 points]

(d) Which node has the highest PageRank after 2 iterations? What is the PageRank of that node? [2 points]

2. **Network Density Computation** [10 Points]

$$\text{Network Density} = \frac{\text{Actual Connections}}{\text{Potential Connections}},$$

where Potential Connections $= \frac{n\times(n-1)}{2}$.

Please compute the network density for network graphs shown in Figures 2, 3, 4. Please show all the calculations [3+4+3=10 points].

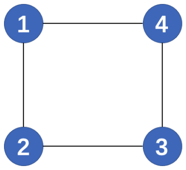

Figure 2: (a)

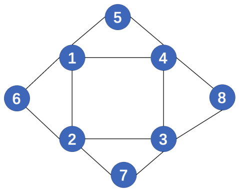

Figure 3: (b)

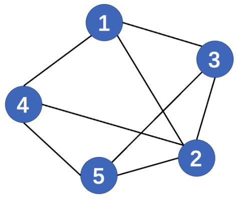

Figure 4: (b)

2.2 Part 2: Exploratory Social Network Analysis [20 points]

This part of the homework is designed to help you familiarize yourself with the dataset and basic concepts of network analysis. The insights from this part of the homework will help you in building the prediction models for Part 3 of the homework.

1. Read `NetworkX` library documentation closely to understand the context and review some code examples of network analyses. [0 point]
2. Read the document linked below to understand the basics of Social Network Analysis. https://www.datacamp.com/community/tutorials/social-network-analysis-python [0 point]
3. Perform some basic network analyses and briefly explain each of your findings:
   (a) Load the directed network graph (G) from the file `amazonNetwork.csv` and draw a graph of the network. [2 points]

   (b) How many items are present in the network and how many co-purchases happened? [2 points]

   (c) Compute the average shortest distance between the nodes in graph G. Explain your results briefly. [2 points]

   (d) Compute the transitivity, the average clustering coefficient, density, and degree assortativity of the network graph G. Explain your findings briefly based on the definitions. [4 points]

   (e) Apply the `PageRank` algorithm to network G with damping value 0.5 and find the 10 nodes with the highest `PageRank`. Explain your findings briefly.
   NetworkX document of the `PageRank` algorithm:
   https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html [5 points]

   (f) Read and explore the documentation of NetworkX:
   https://networkx.org/documentation/networkx-1.10/reference/introduction.html. Perform the average neighbor degree, closeness centrality, and edge betweenness centrality. Report the top 10 results with the highest value for each analysis and briefly explain your findings. [5 points]

The main deliverable for this part of the homework is 1) a step-by-step exploration of data in your Jupyter Notebook. 2) a PDF document containing the answers to each of the questions above. You should also describe your conclusions.

2.3 Part 3: Predicting Review-Rating using Features derived from network properties [50 points]

For this part of the homework, you will build a machine learning model to predict the review rating of the Amazon products on a scale of 0-5 using various network properties as features.

We provide you with the training dataset (`reviewTrain.csv`) which you should judiciously use to train your models. We also provide a test dataset `reviewTest.csv` where the "match" label is missing.

You need to extract at least 4 different features based on the network properties to train your model. The error-metric that we will use for evaluating your match labels on the test dataset is the mean absolute error (MAE). Some of the features that you can consider using include:

- Clustering Coefficient
- Page Rank
- Degree centrality
- Closeness centrality
- Betweenness centrality

Some of the models that you can consider using include:

- Logistic Regression
- Support Vector Machine (SVM)
- Multi-layer perceptron

The main deliverable for this part of the homework is a step-by-step analysis of your feature selection and extraction and model building exercise, describing clearly how you generated features from your dataset and why you chose a specific feature over the other. Your Jupyter notebook should contain the reproducible code for training various models as well as text descriptions of your conclusions after each step.

Your grade on this part of the homework will depend on the accuracy of your model on the test dataset as well as your step-by-step description of how you arrived at your final model. We will evaluate your model using mean absolute error (MAE).

Here are the sample steps you can follow (just for reference):

1. Load the data from the CSV file.

2. Construct some features that you think are relevant for the problem. You can use any of the features we have already computed or you can compute new ones. You can also use the features from the original dataset.
3. Perform feature selection since all features might not be equally informative. Chi-squared test and mutual information are some criteria that you can use to select features.
4. Perform Model Selection.
5. Make predictions on the test dataset.

3 Data Description

Here's the description of files included with this homework.

1. `amazonNetwork.csv`: This file contains the data for Part 2 of the homework. It contains 10841 observations and 2 columns with the numbers representing product IDs. Each node represents a product and each directed edge between two nodes represents a co-purchase. The column `fromNodeId` contains the ID of the main purchasing item and `ToNodeId` contains the ID of the co-purchased items.
2. `reviewTrain.csv`: This file contains the training data for Part 2 of the homework. It contains 1674 observations and 4 columns/features. The `review` column contains ratings on a scale of 1-5.
3. `reviewTest.csv`: This file contains the test data for Part 2 of the homework. Please insert your prediction results in the `review` column in the file.

4 Peer-assessed Exam question generation [10 points]

After receiving some great feedback from the students regarding the questions on the midterm exam, we thought of having a "tiny" competition among the students to generate potential midterm questions! Here are the details:

- You need to generate 1 question that can be a potential exam question for SI 671/721 based on the material that we have covered till 2023-11-08 (Time Series (II)).
- The question should be a multiple choice with 1 or more correct answers. In other words, questions with descriptive answers are not allowed.
- It can be a standalone question testing some course concepts, e.g., the midterm question "Which of the following are frequent itemsets…" OR it can be a composite question with few sub-questions similar to the scenario-based questions on the midterm, e.g., "Planning the course paths for students or Computing Edit Distance for Olivia Rodrigo…"
- You also need to provide the correct answer for the question.

- Your submitted questions will be evaluated anonymously by your fellow classmates! Each student will be assigned 5 questions (from other students), and they will 1) rank those questions from 1 to 5 in terms of quality, and 2) reply with yes/no regarding whether the submitted question was correct in the first place.

Here is how we will grade your submitted questions:

- Submitting the ranked list (and correctness) for the 5 questions assigned to you on time. [2 points]
- Correctness of your own submitted question. [3 points] (0 points if your submitted question/answer was incorrect as judged by your classmates)
- The remaining 5 points will be given based on the quality of your question. 1st rank= 5 points, 2nd rank = 4 points, 3rd rank = 3 points, 4th rank= 2 points, 5th rank = 1 points. For example, if my submitted question received 1 vote each for 1st, 2nd, 3rd, 4th, 5th ranks by the students, then I'd receive (5+4+3+2+1)/5= 3 points out of 5. If my question got all 5th rank votes, then I'd receive (1+1+1+1+1)/5= 1 points out of 5, and so on.

Note that questions that involve asking arcane facts embedded in a footnote on one of the slides might not be rated as high-quality by your peers!

So, get set to unleash your creativity!

5 Submission

All submissions should be made electronically by **11:59 PM EST on November 17, 2023.**

Here are the main deliverable files:

- HTML version of your Jupyter notebook.(Only one HTML file should be submitted)
- The actual Jupyter notebook with "step-by-step analysis," so that we could replicate your results.
- PDF document containing Part1's answer.
- File `reviewTest.csv` with your predicted ratings on a scale of 1-5 for Part 2 of the homework. Keep all the columns in the file `reviewTest.csv` which we shared with you, as they are. Just update the file with your predictions in the correct column.
- Submission details TBD for Question 4. (Most likely, the submission will be as an anonymous submission to Canvas).

6 Academic Honesty

Unless otherwise specified in the homework, all submitted work must be your own original work. Any excerpts, statements, or phrases from the work of others must be clearly identified as a quotation, and a proper citation provided. Any violation of the University's policies on Academic and Professional Integrity may result in serious penalties, which might range from failing a homework, to failing a course, to being expelled from the program. Violations of academic and professional integrity will be reported to the concerned authorities. Consequences of academic misconduct are determined by the faculty instructor; additional sanctions may be imposed.

References

Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007. ISSN 1559-1131. doi: 10.1145/1232722.1232727. URL http://doi.acm.org/10.1145/1232722.1232727.