SI 671/721 (Fall 2019) Data Mining: Methods and Applications

Instructor: Paramveer Dhillon

Homework 3 (Due: 11/25/2019):
Social Network Analysis

1 Summary

For this homework we will use the Amazon co-purchasing network dataset Leskovec et al. (2007) to perform social network analysis. This dataset contains various products' networks including books, music CDs, DVDs, and VHS video tapes. It was collected by crawling Amazon website in March, 2003 according to `Customers Who Bought This Item Also Bought` on the Amazon website. So, if a 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. Exploratory Social Network Analysis.
2. Predicting Review Rating using features derived from Network Properties.
3. Link Prediction.

2.1 Part 1: Exploratory Social Network Analysis [30 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 Parts 2 and 3 of the homework.

a). Read `NetworkX` library documentation closely to understand the context and review some code examples of network analyses.

b). Read the document linked below to understand the basics of Social Network Analysis.
`https://www.datacamp.com/community/tutorials/social-network-analysis-python`

c). Perform some basic network analyses and briefly explain each of your findings:

- Load the directed network graph (G) from the file `amazonNetwork.csv`.

- How many items are present in the network and how many co-purchases happened?
- Compute the average shortest distance between the nodes in graph G. Explain your results briefly.
- Compute the transitivity and the average clustering coefficient of the network graph G. Explain your findings briefly based on the definitions of clustering coefficient and transitivity.
- 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. https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html

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.

2.2 Part 2: Predicting Review-Rating using Features derived from network properties [30 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 use judiciously 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
- Degree of the node

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).

2.3 Part 3: Link Prediction [30 Points]

Next, we will use the Graph G from Part 1 of the homework to identify the edges with missing values and predict whether or not these edges will have a future connection. You need to create a matrix of features (social network properties such as common neighbors or shortest path length) and train a classifier to predict the existence of the edges.

You need to extract at least 3 features in the network properties to train the model. The error-metric that we will use for evaluating your match labels on the test dataset is the Area Under the ROC Curve (AUC). Some of the features that you can consider using include:

- Jaccard coefficient of the nodes.
- The common neighbors of the pair of nodes.
- Resource allocation index of the pair of nodes.
  https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_prediction.resource_allocation_index.html
- Preferential Attachment score of the pair of nodes.
  https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_prediction.preferential_attachment.html
- Shortest path between the pair of nodes.
  https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.shortest_paths.generic.shortest_path_length.html

Some of the models that you can consider using include:

- Logistic Regression.
- Gradient Boosting.
  https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoosting
  html
- Multi-layer Perceptron.

We provide you with the training dataset which you should use judiciously to train our models. We also provide a test dataset `reviewTest.csv` where the review ratings are absent.

The predictions generated by your model should be the probability of the corresponding edges being formed in future. The evaluation metric that we will use is the Area Under the ROC Curve (AUC).

As earlier, your Jupyter Notebook should contain a step-by-step analysis of your feature extraction and model building exercise describing clearly why you chose one feature or model.

3 Data Description

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

1. `amazonNetwork.csv`: This file contains the data for Part 1 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. `linkTrain.csv`: This file contains the training data for Part 3 of the homework. The complete training data set has 8000 observations. The object in `nodes` column are a tuple indicating a pair of nodes that do not have a connection currently. The `connection` column indicates if an edge between those two nodes will be formed in the future, where a value of 0 indicates no future connection and a value of 1.0 indicates a future connection.
5. `linkTest.csv`: This file contains the test data for Part 3 of the homework. Please insert your predicted ratings in the `connection` column in the file.

4 Submission

All submissions should be made electronically by **11:59 PM EST on November 25, 2019.**

Here are the main deliverable files:

- HTML version of your Jupyter notebook.(Only one HTML files should be submitted)
- The actual Jupyter notebook with "step-by-step analysis" for all the three parts of the homework. It's fine to submit everything in a single notebook. (So that we could replicate your results.).(Only one Jupyter notebook files should be submitted)

- 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.
- File linkTest.csv containing the predicted probability of the corresponding edges being future connections for Part 3 of the homework. Again, keep all the columns in the file.

5 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.