COMP90051 Statistical Machine Learning
Project 1 Description

Due date: 5pm Monday 20th April 2020 (competition closes 12pm noon) **Weight:** 25%

1 Overview

Pairwise relationships are prevalent in real life. For example, friendships between people, communication links between computers and pairwise similarity of images. Networks provide a way to represent a group of relationships. The entities in question are represented as network nodes and the pairwise relations as edges.

In real network data, there are often missing edges between nodes. This can be due to a bug or deficiency in the data collection process, a lack of resources to collect all pairwise relations or simply there is uncertainty about those relationships. Analysis performed on incomplete networks with missing edges can bias the final output, e.g., if we want to find the shortest path between two cities in a road network, but we are missing information of major highways between these cities, then no algorithm will able to find this actual shortest path.

Furthermore, we might want to predict if an edge will form between two nodes in the future. For example, in disease transmission networks, if health authorities determine a high likelihood of a transmission edge forming between an infected and uninfected person, then the authorities might wish to vaccinate the uninfected person.

In this way, being able to predict (and correct for) missing edges is an important task.

Your task:

In this project, you will be learning from a training network and trying to predict whether edges exist among test node pairs.

The training network is a fragment of the *academic co-authorship graph*. The nodes in the network—authors—have been given randomly assigned IDs, and an undirected edge between node *A* and *B* represents that authors *A* and *B* have published a paper together as co-authors. The training network is a subgraph of the entire network, focussing on individuals in a specific academic subcommunity.

The test data is a list of 2,000 edges, and your task is to predict if each of those test edges are really edges in the authorship network or are fake ones. 1,000 of these test edges are real and withheld from the training network, while the other 1,000 do not actually exist.

To make the project fun, we will run it as a Kaggle in-class competition. Your assessment will be partially based on your final ranking in the privately-held competition, partially based on your absolute performance and partially based on your report.

2 Data Format

All data will be available in raw text. The training graph data will be given in a (tab delimited) edge list format, where each row represents a node and its neighbours. For example:

```
1   2
2   1   2   4
3   2   5
4   2   5
5   3   4
```

represents the network illustrated in Figure 1.

In addition to the edges, you are also provided with a file including several features of the nodes (authors). This file, “nodes.json” is in JSON format and includes for each author:

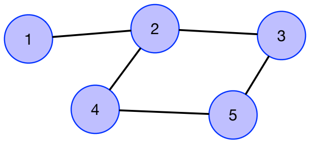

Figure 1: Network diagram for the adjacency list example.

- their `id` in the graph
- the number of years since their `first` and `last` publication
- their number of publications in total, `num_papers`
- presence of specific keywords in the titles and abstracts of their publications (denoted `keyword_X` where $X \in \{0,1,\ldots,52\}$, each being a binary value and only listed if its value is 1)
- publication at specific venues (denoted `venue_X` where $X \in \{0,1,\ldots,347\}$, each being a binary value and only listed if its value is 1)

This gives you some additional information beside the network structure for your prediction task.[1]

The test edge set is in a comma separated values (CSV) edge list format, which includes a one line header, followed by a line for each (source node, target node) edge. Your implemented algorithm should take the test CSV file as input and return a 2,001 row CSV file that has a) in the first row, the string "Id,Predicted"; b) in all subsequent rows, a consecutive integer ID, a comma, then a float in the range [0,1]. These floats are your "guesses" or predictions as to whether the corresponding test edge was from the co-authorship network or not. Higher predictions correspond to being more confident that the edge is real.

For example, given the test edge set of $\{(3,5),(4,12)\}$ as represented in CSV format by

```
Id,Source,Sink
1,3,5
2,4,12
```

if your prediction probabilities are 0.1 for edge (3,5), 0.99 for edge (4,12), then your output file should be:

```
Id,Predicted
1,0.1
2,0.99
```

The test set will be used to generate an AUC for your performance; you may submit test predictions multiple times per day (if you wish). During the competition AUC on a 30% subset of the test set will be used to rank you in the **public leaderboard**. We will use the complete test set to determine your **final AUC and ranking**. The split of test set during/after the competition, is used to discourage you from constructing algorithms that overfit on the leaderboard. The training graph "train.txt", the test edges "test-public.csv", and a sample submission file "sample.csv" will be available within the Kaggle competition website. In addition to using the competition testing and to prevent overfitting, we encourage you to generate your own test edge sets from the training graph, and test your algorithms with that.

3 Links and Check List

Competition link: `https://www.kaggle.com/t/7a46cb8512da4f58a99cd1c1be8ccc39`

The Kaggle in class competition allows you to compete and benchmark against your peers. Please do the following **by Monday 30th March 2020**:

[1] These features were calculated after excluding from the network the hidden test edges, to invalidate trivial approaches for prediction.

1. Setup one (and only one) account on Kaggle with your university email.

2. Register your entry using the 'registration' form link: `http://go.unimelb.edu.au/o99r`.

4 Report

A report describing your approach should be submitted through LMS **by 5pm Monday 20th April 2020** . It should provide the following sections:

1. A brief description of the problem and introduction of any notation that you adopt in the report.

2. Description of your final approach(s) to link prediction, the motivation and reasoning behind it, and why you think it performed well/not well in the competition.

3. Any other alternatives you considered and why you chose your final approach over these (this may be in the form of empirical evaluation, but it must be to support your reasoning - examples like "method A, got AUC 0.6 and method B, got AUC 0.7, hence I use method B", with no further explanation, will be marked down).

Your description of the algorithm should be clear and concise. You should write it at a level that a postgraduate student can read and understand without difficulty. If you use any existing algorithms, *please do not rewrite the complete description, but provide a summary* that shows your understanding and references to the relevant literature. In the report, we will be interested in seeing evidence of your thought processes and reasoning for choosing one algorithm over another.

Dedicate space to describing the features you used and tried, any interesting details about software setup or your experimental pipeline, and any problems you encountered and what you learned. In many cases these issues are at least as important as the learning algorithm, if not more important.

Report format rules. The report should be submitted as a PDF, and be no more than two A4 pages, single column. The font size should be 11 or above. If a report is longer than three pages in length, we will only read and assess the report up to page 2 and ignore further pages. (Don't waste space on cover pages. By 2 pages, I mean two sides, not two double-sided pages.)

5 Submission

The final submission will consist of three parts:

- A valid submission to the Kaggle in class competition **by 12pm noon Monday 20th April 2020** . This submission must be of the expected format as described above, and produce a place somewhere on the leaderboard. Invalid submissions do not attract marks for the competition portion of grading (see Section 6).

- To LMS **by 5pm Monday 20th April 2020** , a zip archive[2] of your source code[3] of your link prediction algorithm including any scripts for automation, and a README.txt describing in just a few lines what files are for (but no data please).

- To LMS **by 5pm Monday 20th April 2020** , a written research report in PDF format (see Section 4).

The submission link will be visible in LMS prior to deadline.

[2] Not rar, not 7z, not lzh, etc. Zip! Substantial penalties will be applied if you don't follow this simple instruction.

[3] We would encourage you to use Python, but we will also accept submissions in Matlab or R. Should you wish to use another language, please ask on the discussion board. You are welcome to use standard machine learning libraries, such as sklearn, pytorch, etc, but the code submitted should be your own.

Table 1: The function absolute(A), converting an AUC score A to a mark between 0 and 9

A	≥0.98	0.96	0.94	0.92	0.90	0.88	0.86	0.84	0.82	0.80	0.78	0.76	0.74
mark	9	8.1	6.8	5.9	5.2	4.6	4.1	3.7	3.3	2.9	2.6	2.4	2.1
A	0.72	0.70	0.68	0.66	0.64	0.62	0.60	0.58	0.56	0.54	0.52	≤0.50	
mark	1.9	1.6	1.4	1.2	1.1	0.9	0.7	0.6	0.4	0.3	0.1	0	

6 Assessment

The project will be marked out of 25. Note that there is a hurdle requirement on your combined continuous assessment mark for the subject, of 25/50, of which Project 1 will contribute 25 marks. **Late report submissions will incur a deduction of 2 marks per day—it is not possible to make late competition entries.**

The assessment in this project will be broken down into two components. The following criteria will be considered when allocating marks.

Based on our experimentation with the project task, we expect that all reasonable efforts at the project will achieve a passing grade or higher.

Kaggle Competition (**12/25**):

Your final mark for the Kaggle competition is based on your rank in that competition. Assuming N enrolled students compete, there are no ties and you come in at R place (e.g. first place is 1, last is N) with an AUC of $A \in [0,1]$ then your mark is calculated as

$$\texttt{absolute}(A) + 3 \times \frac{N-R}{N-1} \ .$$

where absolute(·) returns a value between 0 and 9, and is defined in Table 1. Ties are handled so that you are not penalised by the tie: tied students receive the rank of the highest student (as if no entries were tied). This expression can result in marks from 0 to 12. For example, if students A, B, C, D, E came 1st, 4th, 2nd, 2nd, 5th, then the rank-based mark terms (out of 3) for the five students would be 3, 0.75, 2.25, 2.25, 0.

This complicated-looking expression can result in marks from 0 all the way to 12. We are weighing more towards your absolute AUC than your ranking. The component out of 9 for AUC gives a score of 0/9 for AUC of 0.5 or lower (no better than random guessing); 9/9 for AUC of 0.98; and logarithmically scales over the interval of AUCs [0.5, 0.98]. We believe that an AUC higher than 0.88 is easily achievable, while results of 0.92 or above will require more effort. *For example, an AUC of 0.92 for a student coming last would yield 5.9/12; or 7.4/12 if coming mid-way in the class.*

External unregistered students may participate, but their entries will be removed before computing the final rankings and the above expression, and will not affect registered students' grades. We do not actively invite such participation.

The rank-based term encourages healthy competition and discourages collusion. The other AUC-based term rewards students who don't place in the top but none-the-less achieve good absolute results.

Note that invalid submissions will come last *and* will attract a mark of 0 for this part, so please ensure your output conforms to the specified requirements.

Report (**13/25**):

The marking rubric in Appendix A outlines the criteria that will be used to mark your report.

Plagiarism policy: You are reminded that all submitted project work in this subject is to be your own individual work. Automated similarity checking software will be used to compare submissions. It is University policy that academic integrity be enforced. For more details, please see the policy at `http://academichonesty.unimelb.edu.au/policy.html`.

A Marking scheme for the Report

Critical Analysis (Maximum = 8 marks)	**Report Clarity and Structure** (Maximum = 5 marks)
8 marks Final approach is well motivated and its advantages/disadvantages clearly discussed; thorough and insightful analysis of why the final approach works/not work for provided training data; insightful discussion and analysis of other approaches and why they were not used	5 marks Very clear and accessible description of all that has been done, a postgraduate student can pick up the report and read with no difficulty.
6.4 marks Final approach is reasonably motivated and its advantages/disadvantages somewhat discussed; good analysis of why the final approach works/not work for provided training data; some discussion and analysis of other approaches and why they were not used	4 marks Clear description for the most part, with some minor deficiencies/loose ends.
4.8 marks Final approach is somewhat motivated and its advantages/disadvantages are discussed; limited analysis of why the final approach works/not work for provided training data; limited discussion and analysis of other approaches and why they were not used	3 marks Generally clear description, but there are notable gaps and/or unclear sections.
3.2 marks Final approach is marginally motivated and its advantages/disadvantages are discussed; little analysis of why the final approach works/not work for provided training data; little or no discussion and analysis of other approaches and why they were not used	2 mark The report is unclear on the whole and the reader has to work hard to discern what has been done.
1.6 mark Final approach is barely or not motivated and its advantages/disadvantages are not discussed; no analysis of why the final approach works/not work for provided training data; little or no discussion and analysis of other approaches and why they were not used	1 mark The report completely lacks structure, omits all key references and is barely understandable.