COMP90051 Statistical Machine Learning
2015 Semester 2—Project 1 Spec

Due date: Report due Tues 5:00pm Sep 1 2015 (competition closes 9:59am same day).
Weight: 20% or 25% depending on midsemester performance (whichever is higher)
Competition link: `https://inclass.kaggle.com/c/comp90051-2015-link-prediction`
Team registration: `https://docs.google.com/forms/d/1XVhWpaz7aO1y2u3gta2CxQHrySWetFDawHp_ZUi1cOA/viewform?usp=send_form` (complete in first week)

1 Overview

Pairwise relationships are prevalent in real life. For example, friendships between people, communication links between computers and pairwise similarity of cars. A way to represent a group of relationships is using networks, which consists of a set of nodes and edges. The entities in question are represented as the nodes and the pairwise relations as the edges.

In real data, often there are missing edges between nodes. This can be due to a bug or deficiency in the data collection process, the 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 not be 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 a disease transmission network, 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.

Hence, 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 partial crawl of the *Twitter social network* from several years ago. The nodes in the network—Twitter users—have been given randomly assigned IDs, and a directed edge from node A to B represents that user A follows B. The training network is a subgraph of the entire network. Starting from several random seed nodes, we proceeded to obtain the friends of the seeds, then their friends' friends, and so on for several iterations.

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 Twitter 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 given in a (tab delimited) adjacency edge list format, where each row represents a node and its out neighbours (users being followed by that node). For example:

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

represents the network illustrated in Figure 1.

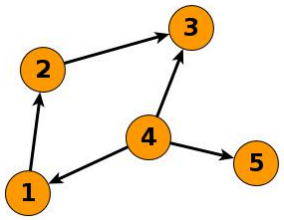

Figure 1: Network diagram for the adjacency list example.

The test edge set is in a (tab-delimited) edge list format also, where each represents an edge (source node, target node). Given this 2,000-row edge list, your implemented algorithm should take the test list in and return a 2,001 row CSV file that has a) in the first row, the header string "Id,Predictions"; b) in all subsequent rows, an ID integer representing row number (1 through 2000), 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 Twitter network or not. Higher predictions correspond to being more confident that the edge is real.

For example, given the test edge set of:

3 1
3 4

if your prediction probabilities are 0.1 for edge (3,1) and 0.99 for edge (3,4) is true, then your output file should be as follows:

Id,Predictions
1,0.1
2,0.99

The test set will be used to generate an AUC for your performance. During the competition AUC on a subset of the test set will be used to rank you in the leaderboard. We will use the complete test set to determine your final AUC and ranking (using the best of two of your chosen submissions). 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.txt", 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 (a validation set), and test your algorithms with that.

This process closely reflects how you would ideally practice machine learning in reality.

3 Kaggle Competition

The Kaggle in class competition allows you to compete and benchmark against your peers. Please do the following by the end of the first week:

- Setup an account on Kaggle with username and email both being your unimelb student email—only unimelb emails can access to the competition.
- Form your team of student peers
- Connect with your team mates on Kaggle as a Kaggle team. Only submit via the team
- Register your team using the Google forms link at the top of this spec

The performance measure we use to evaluate your prediction is the *Area Under Curve (AUC)*. The competition server computes an *receiver operating characteristics (ROC) curve* from your 2,000 probabilities, and the AUC score is the area under this curve.

During the course of the competition a public leaderboard will rank teams by the AUC of their latest entry. As stated in the previous section, the leaderboard AUC's are computed on a random subset of the data. The final leaderboard is computed on the entire test set. Teams should consist of three individuals. If you cannot find a team, please introduce yourself to fellow students in workshop, or post to LMS that you are in search of a team—we will try to assist. In only very rare occassions will we permit teams of less than three (and we will mark all teams based on our expectations of what a team of three could achieve). The motivation for working in teams is that in industry, practising machine learning experts work effectively in teams.

We encourage active discussion among teams, but please refrain from colluding. Given your marks are dependent on your final ranking in the competition, it is in your interest not to collude.

4 Report

A report describing and explaining your approach should be written and submitted. It should provide the following sections:

1. A very 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 we 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, you do not have to rewrite the complete description. Aim to provide a brief summary that shows your understanding and a few references to the relevant literature. In the report, we will be very 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.

The report should be submitted as a PDF, and be no more than three 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 three and ignore further pages.

5 Submission

The final submission will consist of three parts:

- A valid submission to the Kaggle in class competition. 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).
- Your source code of your link prediction algorithm. Your code can be in any of the following languages {C, C++, C#, Python, Java, R, Matlab}. If there is another language you like to use, please contact us first. If the language requires compiling, a makefile or script must be provide to build the executables. We may or may not run your code, but we will definitely read it.

- A written research report in PDF format (see Section 4).

Note: submission of the report and code will be done via LMS via the links under assessment.

6 Assessment and Marking Rubric

The project will be marked out of 20.[1] No late submission of Kaggle portion will be accepted; late submissions of reports will incur a deduction of 4 marks per day.

Based on our experimentation with the project task and the design of the marking scheme below, we expect that all reasonable efforts at the project will achieve a passing grade or higher. So relax and have fun!

Kaggle Competition (**10/20**): Your final mark for the Kaggle competition is based on your rank in that competition. Assuming N teams of 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

$$8 \times \frac{\max\{\min\{A, 0.90\} - 0.4, 0\}}{0.50} + 2 \times \frac{N - R}{N - 1} \ .$$

Ties are handled so that you are not penalised by the tie: by separating tied AUC's by subtracting very small random numbers from all but one, to break ties. All who are tied then gets the score of the highest (unperturbed) team. External teams of unenrolled students (auditing the subject) may participate, but their entries will be removed before computing the final rankings and the above expression, and will not affect registered students' grades.

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

This complicated-looking expression can result in marks from 0 all the way to 10. We are weighing slightly higher your absolute AUC than your ranking. **The component out of 8 for AUC gives a score of 0/8 for AUC of 0.4 or lower; 8/8 for AUC of 0.9 or higher; and linearly scales over the interval of AUCs [0.4, 0.9].** We believe that much higher than 0.5 (random classifier) AUC is achievable with minimal work, while 0.9 AUC is an excellent result deserving of full marks. *For example, an AUC of 0.8 for a team coming last would yield 6.4/10; or 7.4/10 if coming mid-way in the class.*

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. Also teams not registering their Kaggle user and team names with Ben within the first week will receive zero.

Report (**10/20**): The 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 cheating by students in any form is not permitted, and that work submitted for assessment purposes must be the independent work of the student(s) concerned.

For more details, please see the policy at `http://academichonesty.unimelb.edu.au/policy.html`.

[1] But its weight towards your final mark for the subject could be either 20% or 25% depending on your midsemester performance. Remember also that the project forms part of your hurdle requirements.

A Marking scheme for the Report

Critical Analysis (Maximum = 5 marks)	**Report Clarity and Structure** (Maximum = 5 marks)
5 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.
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.
3 marks Final approach is somewhat motivated and its advantages/disadvantages is 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.
2 marks Final approach is marginally motivated and its advantages/disadvantages is 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 mark Final approach is barely or not motivated and its advantages/disadvantages is 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.