COMP90051 Statistical and Evolutionary Learning
Project 1 Description

Due date: 5:00pm (competition closes 9:59am) Thursday, 4th September 2014
Weight: 20% or 25% depending on midsemester performance (whichever is higher; see slides01.pdf)
Competition link: https://kaggle.com/join/comp90051where

1 Overview

Social networks are fast becoming the dominant platform for communication online, with many of you likely holding accounts on half-a-dozen networks or more. Facebook, Linkedin, Twitter, Google+, Foursquare, Renren, Tencent Weibo, Vkontakte are a few of the very many social networks in use today.

Several aspects of online social networks offer unique insights into users: friendships or connections between users as represented by the social graph; location data of users which often varies as many prefer mobile access to social networks; profile and post content which can be quite voluminous and reveal much about users.

The interest of companies and governments in online social networks is largely due to the potential insights that can be gained. On one hand, advertising on social networks can be highly effective and is worth billions; on the other, user's privacy can be breached by unexpected access to data, or by insights that machine learning can derive from what is published intentionally. It is clear that machine learning is an important tool for social network analysis, and for any activity that derives value from social networks. In this project you will explore how machine learning can be used to make predictions on social network data.

Your task: In this project, your task is to predict user locations for a test set of 1,000 users. For a large number of users from a real social network (57,562 total), you have access to the social graph (undirected friendships between users, including the test users) and limited profile information on all users: the first, second and third most common hours of the day they post and the number of posts made. For all but the test users you will be given the latitude-longitude coordinates most common among each user's posts. Your task is to predict both coordinates for each of the test users.

2 Data Format

All data is available as raw text from the Kaggle site. Several files comprise the data.

2.1 graph.txt

This file represents the social graph. It is tab-delimited with each line having two user IDs from 1 through 57,562 representing that the two users are friends in the social network. Note edges are undirected (like Facebook, not directed like Twitter or Google+). There are 420k lines in the file representing 210k edges (each is representing twice in the file in both orders). There is no header line in the file. The first few lines of the file are:

```
1 52781
2 8100
2 27339
```

Note that some users in the range 1 through 57,562 may not have an edges incident to them. They are still valid users. Note that the test users' connections have been left visible in this file: there is no distinction between train and test users in the graph.

2.2 posts-train.txt

This file represents all available meta-data for users not in the test set. The first few lines of the file are:

```
Id,Hour1,Hour2,Hour3,Lat,Lon,Posts
1,18,19,20,28.6,77.2,38
2,13,07,08,18.490,73.912,13
```

The file has a header line, and represents 7 comma-delimited columns:

1. **Id**. The user ID, same as in graph.txt
2. **Hour1.** The first most frequent hour of the day the user posts. In UTC 24hr time. Valid values are 00 through 23 inclusive. Missing values are denoted as 25, meaning the user had no posts.
3. **Hour2.** The second most frequent hour of the day the user posts. In UTC 24hr time. Valid values are 00 through 23 inclusive. Missing values are denoted as 25, meaning the user only posted within the hour recorded in Hour1.
4. **Hour3.** The third most frequent hour of the day the user posts. In UTC 24hr time. Valid values are 00 through 23 inclusive. Missing values are denoted as 25, meaning the user only posted within the hours recorded in Hour1 and Hour2.
5. **Lat.** The latitude most common to the user's posts.[1] Users without posts, or who post without location, have 0.0 Lat.
6. **Lon.** The latitude most common to the user's posts. Calculated along with Lat above. Users without posts, or who post without location, have 0.0 Lon.
7. **Posts.** The total number of posts made by the user (integer 0 or more)

There is at most one line per user: only 49,813 here out of 56,562 training users: a few thousand users had no posts so are not represented in this file, but may be represented in graph.txt.

NOTE users may choose not to enable location on their posts. In those cases the posts had lat-lon of 0.0,0.0. So lat-lon of 0.0,0.0 can occur with non-zero Posts.

2.3 posts-test-x.txt

This file represents meta-data for the 1000 test users. All of the columns from posts-train.txt are represented here for the test users, *except Lat and Lon are ommitted* since it is your task to predict these values. As stated above, these 1000 test users are represented in graph.txt

2.4 submission-example.txt

This is an example submission file. Try uploading it to get started! its first few lines are:

```
Id,Lat,Lon
18,35.699,139.578
65,40.804,-73.951
```

For each of the test users this file represents the user's ID followed by your predicted Lat and Lon for the user. Note the file is comma-delimited, and has a specific header. In total it has 1001 lines.

To make a submission in the competition (you can make many – up to 5 per day if you wish) you simply upload a file like this to the Kaggle server through the competition webpage.

[1] To calculate this, we took all post latitudes and longitudes, truncated the two coordinates to at most 3 decimal places, then concatenated the two coordinates as a string. The most frequent such string for the user's posts is taken to be Lat and Lon.

3 Kaggle Competition

To make the project fun, we will run it as a Kaggle in-class competition, link at top of handout.

As discussed below, the performance measure is the root-mean squared error (RMSE) calculated from the 1,000 lat-lon pairs in each of your submissions. During the course of the competition a public leaderboard will rank teams by the RMSE of their latest entry. The leaderboard RMSE's are computed on a random subset of half of the data. The final leaderboard is computed on the entire test set. Your assessment will be partially based on your final ranking, your RMSE score and on a short report.

You must work in a team of 2. We will mark all teams based on our expectations of what a team of two could achieve. Both team members will need to create a Kaggle account, and should do so using their unimelb email address so that they have access to the competition. Those two users can then form a Kaggle team for the competition. Please register one Kaggle team only and only submit as a team. *Ben will ask in the first week for Kaggle usernames and team names so that your performance can be included in your final mark.* Those looking for a team-mate after a few days should post to the LMS discussion board. See Ben if you are unable to find a team-mate in a timely fashion as he can help!!

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 PDF report describing and explaining your approach should be written and submitted. It should be: **no more than three A4 pages single column, margins 1cm or more, font size 11 or above.** Any pages beyond three will be ignored. The report 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 location 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 RMSE 0.6 and method B, got RMSE 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, but provide a brief summary that shows your understanding and 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.

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 in a zip archive. Your code can be in any of the following languages {C, C#, Python, Java, R, Matlab}. If there is another language you like to use, please ask Ben. 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 is via LMS. We will provide details closer to the submission deadline.

6 Assessment

The project will be marked out of 20.[2] No late submission of the 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**): Part of your mark for the project comes from the Kaggle competition. The evaluation metric in use is the root mean-squared error (RMSE) which is computed on both your Lat and Lon predictions then summed. See `https://www.kaggle.com/wiki/RootMeanSquaredError` for details. Zero RMSE means you have perfect predictions, bigger means more error. You will get a final rank in the 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) then your mark for the competition part is calculated as

$$7 \times \frac{\min\{(140 - RMSE)_+, 120\}}{120} + 3 \times \frac{N - R}{N - 1} \ .$$

Note $(x)_+$ returns x for $x > 0$, zero otherwise. The first term from RMSE can be up to 7.0 while the second term from rank can be up to 3.0. Ties are handled so that you are not penalised by the tie: by separating tied RMSE's by subtracting very small random numbers from all but one, to break ties. All who are tied then gets the score out of 3.0 of the highest (unperturbed) team. We believe that with a small amount of effort you could get RMSE well below 80 (performance of random!); 35 is a fine performance and much lower is possible.

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

For example: a team getting a (quite achievable) 50 RMSE and ranking 30 out of 40 would get a respectable 6/10; a good 40 RMSE and ranking 10 out of 40 would yield over 8/10. Note: RMSE of 140 or higher will get a 0/7 for the first term, while 20 or lower gets 7/7.

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. Have at least some kind of valid submission early on!

Report (**10/20**): The marking sheet in Appendix A outlines the criteria that will be used to mark your report. With an average report and effort in the competition, even if far from the best team, any team should pass the project overall. We hope that you'll have lots of fun!

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, see the policy at `http://academichonesty.unimelb.edu.au/policy.html`.

[2] 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. See slide01.pdf for details.

A Marking scheme for the Report

Critical Analysis (Maximum = 6 marks)	**Report Clarity and Structure** (Maximum = 4 marks)
6 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	4 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.
4.8 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	3.2 marks Clear description for the most part, with some minor deficiencies/loose ends.
3.6 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	2.4 marks Generally clear description, but there are notable gaps and/or unclear sections.
2.4 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	1.6 mark The report is unclear on the whole and the reader has to work hard to discern what has been done.
1.2 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	0.8 marks The report completely lacks structure, omits all key references and is barely understandable.