Homework 1

Special topics on Machine Learning (SDML)

Fall 2018, NTU CSIE

Prof. Shou-De Lin

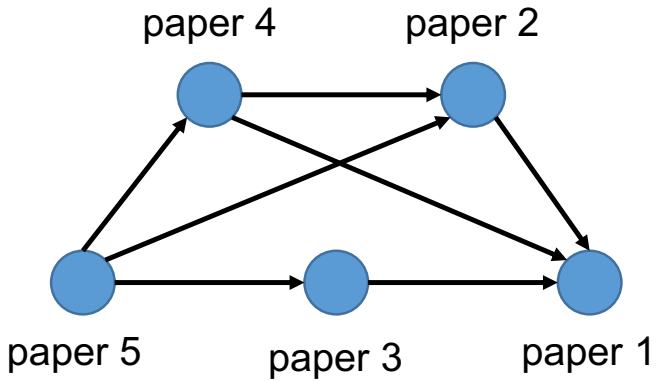
TA: Skyly Yang

Introduction

- **Topic**: Link prediction problem on networks.
- **Data:** Real-world citation networks.
- For HW1, three sub-problems are prepared with different problem settings:
  - **Task 1**: typical link prediction setting given pure graph information (the edge list).
  - **Task 2**: Similar to Task 1, but additional meta-data are provided for each node.
  - **Task 3**: Addition to Task 2, temporal information of edges (date of links) are provided.

Citation Networks

- **Node**: publications (e.g., papers, articles, books)
- **Edge**: citation references
  - Directed edges (arcs) <*a*, *b*> =/= <*b*, *a*>
- Directed Acyclic graphs (DAGs)
  - Temporal topological order on citations (i.e., links).



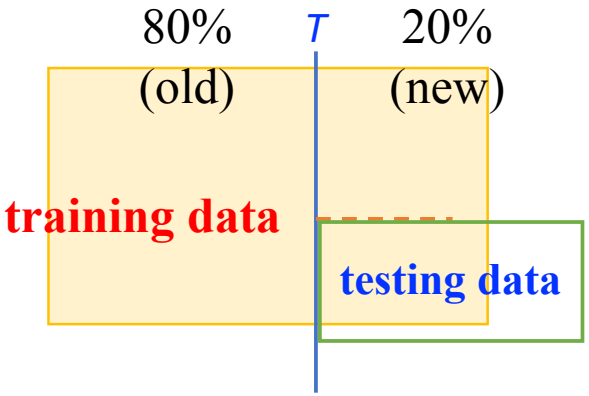
paper 5 cites paper 4;
paper 5 cites paper 3;
paper 5 cites paper 2;
paper 4 cites paper 2;
paper 4 cites paper 1;
paper 3 cites paper 1;
paper 2 cites paper 1

Link Prediction

- **Goal**: to determine whether a link <*u*, *v*> will occur, based on the partial graphs in training data.

- For all the three tasks of HW1:
  - Real-world datasets are used.
  - The given data are relabeled and randomly shuffled.
  - All the queries in the testing data are guaranteed to be no earlier than the links provided in training data.

Task 1



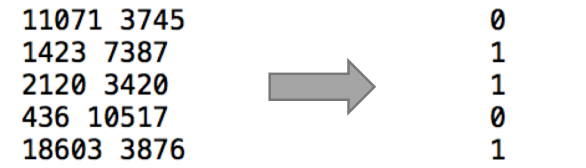
- Given data for training:
  - 80% nodes as papers published earlier (before *T*) and all their links (t1-train.txt)
  - About 50% of links of the 20% testing nodes published after *T* (t1-test-seen.txt)
- Queries for testing:
  - The remaining links (positives) and roughly the same number of *dummy* negative links (t1-test.txt)
- Meaning: given some citations of a paper, can you predict what other papers it cites?

- Data format:
  - Each line denotes a citation record:
    <from-id> <to-id> (separated by a space)
    means publication *from-id* cites *to-id*.
  - The *id*s are randomly shuffled, no temporal meaning on them.

```
7695 450
12392 51
733 11061
22742 5719
4829 7296
18947 1825
16899 16900
10582 16533
2871 9700
25812 6744
23919 7722
25273 5993
7617 7892
4519 5934
```

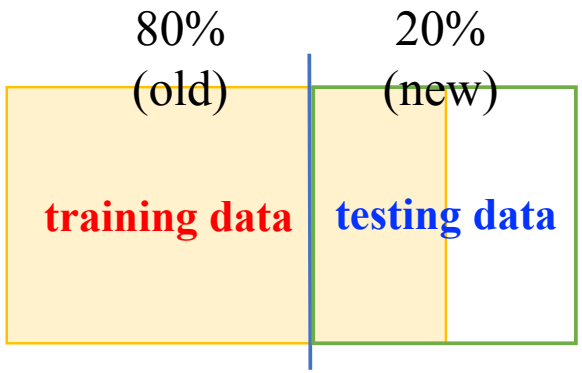
Task 1 (cont.)

- Prediction outputs:
  - To distinguish whether each queried link is positive (output "1") or not (output "0"). (pred.txt)
- Output format:
  - If there are *Q* queries (#lines in t1-test.txt), then the output (pred.txt) should contain exactly *Q* lines.
  - For each line, your model should output a single integer (0 or 1) to indicate the prediction of corresponding query.



(t1-test.txt) (pred.txt)

Task 2



- Similar to Task 1, but no link information of the 20% testing nodes are given in the training data. (t2-train.txt)
- However, this time the title and abstract of each publication are provided (in raw texts, .xml format)
  - That is, we have some additional meta-data of nodes.
  - You will need to consider *document embedding* methods

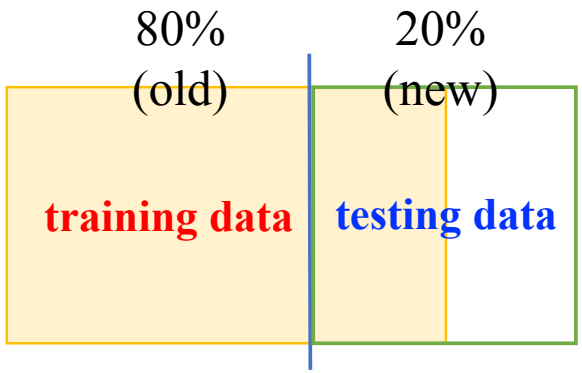
```
<title>
Confining Strings in the Abelian-Projected SU(3)-Gluodynamics
</title>

<abstract>
String representation of the Wilson loop in 3D Abelian-projected
SU(3)-gluodynamics is constructed in the approximation that Abelian-projected
monopoles form a gas. Such an assumption is much weaker than the standard one,
demanding the monopole condensation. It is demonstrated that the summation over
world sheets, bounded by the contour of the Wilson loop, is realized by the
summation over branches of a certain effective multivalued potential of the
monopole densities. Finally, by virtue of the so-constructed representation of
the Wilson loop in terms of the monopole densities, this quantity is evaluated
in the approximation of a dilute monopole gas, which makes confinement in the
model under study manifest.
</abstract>
```

- Meta-data filename: t2-doc/<*pub-id*>.xml

- For example, for publication-id (node id) 217, its corresponding meta-data file is named t2-doc/217.xml .

Task 3



- For our last task, the publication date is also provided for each node.
  - Leverage temporal information to improve embeddings.
  - Titles and abstracts are still available.

```
<date>
2001/01/05
</date>

<title>
Perturbative Quantum Field Theory in the String-Inspired Formalism
</title>

<abstract>
We review the status and present range of applications of the
``string-inspired'' approach to perturbative quantum field theory. This
formalism offers the possibility of computing effective actions and S-matrix
elements in a way which is similar in spirit to string perturbation theory, and
bypasses much of the apparatus of standard second-quantized field theory. Its
development was initiated by Bern and Kosower, originally with the aim of
simplifying the calculation of scattering amplitudes in quantum chromodynamics
and quantum gravity. We give a short account of the original derivation of the
Bern-Kosower rules from string theory. Strassler's alternative approach in
terms of first-quantized particle path integrals is then used to generalize the
formalism to more general field theories, and, in the abelian case, also to
higher loop orders. A considerable number of sample calculations are presented
in detail, with an emphasis on quantum electrodynamics.
</abstract>
```

- Publication date
  => citation date
  => date of link

- Meta-data filename:
t3-doc/<*pub-id*>.xml

Kaggle Competitions

- Testing data will be divided into private and public testing.
  - You will be evaluated based on the performance of private+public testing.
- Maximum 5 submissions a day per task are permitted.
  - You should choose one final choice among all your valid submissions before the deadline.
  - Remember to declare your team on the Kaggle platform for Task 2 and 3.
- Using *extra data* from the Internet is prohibited.
- Please use the <Student-ID>_<Chinese Name> as the Kaggle nickname to show on the leaderboards (for task 2 and 3, use the team leader's name).
  - For example, r05922000_王小明 as the nickname.
- Competition pages:
  - Task 1: https://www.kaggle.com/c/ntucsie-sdml2018-1-1
  - Task 2: https://www.kaggle.com/c/ntucsie-sdml2018-1-2
  - Task 3: https://www.kaggle.com/c/ntucsie-sdml2018-1-3

Grading Policy

- Task 1: 50%, Task 2: 30%, Task 3: 20%
- Performance (50%)
  - *Performance Ranking:* all the participant will be ranked according to the Kaggle testing scores.
  - *Baselines*: you need to beat our baseline for a basic score.
- Report (50%)
  - *Coverage* (25%): #methods you tried; please describe and analyze the approaches with experiment results.
  - *Novelty* (25%): how novel is your model designed.

    (Ensemble techniques are valid, however we encourage novel single models.)

Available Programming Languages

- For this homework, the following languages are allowed:
  - C (up to C11) / C++ (up to C++17)
  - Java 7, 8 or 9
  - Python 2.x or 3.x
  - Ruby 2.5.1
  - Perl 5 or 6
  - MATLAB (up to 2018a, only basic toolboxes)
  - R 3.5.1
- Basically, you can use any properly licensed third-party source codes.
  - However, the programming language restrictions above are applied on the third-party codes as well.
  - Third-party executables or any platform-sensitives are NOT allowed.
  - You should include their licenses in CEIBA submissions.

CEIBA Submissions

- You should submit your source codes along with reports to the corresponding CEIBA entries.
  - Including any third-party source codes you used.
  - A .zip file should be uploaded for each task. (i.e., you should upload three .zip files in total for HW1)
  - The format of CEIBA submissions is stated later.
- Your CEIBA submissions should match your final output on the Kaggle platform.
  - That is, your source codes should be able to reproduce your final performance scores on Kaggle.

CEIBA Submissions (cont.)

- For the source codes, you should also make sure that all of them can be correctly compiled and executed on the CSIE workstation.
  - Please contact R217 for further information.
  - You may have to register a temporary account as well.
  - You should provide precise README files.
  - Avoid heavy CPU/GPU computations if possible.
- **Plagiarism is strictly prohibited.**
  - You should clearly mention *all* the third-party codes (if any) used in your submissions.
  - We will check your source codes via professional softwares.

Reports

- *Three* report files should be submitted for the *three* tasks, respectively.
  - Your reports should be formatted in PDF files.
  - Only digital submissions on CEIBA are acceptable.
- The reports should include:
  - Official name and the student ID of each member.
  - Attempted approaches to solve specific problems.
  - Analyses and observations based on experiment results.
  - Difficulties encountered, unsolved issues, etc.
- No page limit. ☺
  - Feel free to include all the experiment results, reference theorems or other appendices.

Format of CEIBA Submissions

[student-id].zip (team leader's for task 2 and 3, e.g., r05922000.zip)

|-- src/ (the source codes written by you)

|-- lib/ (all the libraries, third-party source codes you used)

|-- report.pdf

|-- README (a 'plaintext' file to explain how to reproduce your results)

(You *must* submit this .zip to get the 50% performance points.)

<Important> You should upload in **.zip** format.

.rar, .tar, .gz, .7z, or any other formats will receive **0 points** without grading.

Submission Deadlines

- Due time (task 1): 2018/10/03 23:59:59 (Taiwan time)
  - According to the system times of Kaggle and CEIBA.
  - Since the network status is unpredictable, please make your submissions as earlier as possible.
  - Report due on: 2018/10/13 23:59:59
- Due time (task 2, 3): 2018/10/10 23:59:59 (Taiwan time)
  - For Task 2 and 3, only team leader has to submit the .zip & report.
  - Report due on: 2018/10/13 23:59:59
- HW1 presentation: 10/11
- For the delayed submissions:
  - Within 24 hours: original_task_score * 0.5
  - More than 24 hours: zero point for that task.

Contact TA

- If you have any problems, feel free to contact TAs.

- TA in charge: 楊鈞百
  - TA hour: Tue. 13:00~14:00