MSBD 5001 Project Specifications

MSBD 5001 team
{kaichen,kccecia,jxiaab}@cse.ust.hk

1 Introduction

The MSBD 5001 group project (abbr. project) is designed for students to form groups and solve real-world data analytic problems. This document would be updated regularly to describe the latest specifications and of group project. Both written report and oral presentation are required.

The document is organized as follows: We first describe the problem and group size requirements (§2) Then we describe the specific requirements on proposals (§3) and final presentations & reports (§4).

2 Requirements

The goal of the project is to find and solve problems based on the real world datasets. We list some datasets for your reference, but you are encouraged to explore datasets with reasonable techniques.

Problems: The project problem / goal is open to any topics relevant to data analytic. Some suggested topics are:

- Determine a problem that you can analyze from real world datasets. For example, you can do a classification/regression task on a specific dataset, after necessary preprocessing and manual split of train-test. You need to explain us the method you are using, and why do you think it works well on your specific problem, and properly present your results.
- Propose a novel method of data analytic. On the specific problem that you describe with your dataset, current methods may have certain disadvantages, and you can purpose a new method (a new model / learning algorithm) to outperform previous methods. In this case you need to explain the correctness of your method and use experiments to demonstrate the advantages.
- Any other topics you can think of that is related to data analytic.

Dataset: Pick ONE (or more if you like) dataset to work with. Be sure to cite your dataset or declare its source clearly in your proposal and report. We will provide you some datasets in the reference, you are free to use (or to not use) any of them.

Team work: Please form a team of exactly FIVE (5) people, and email your team members to TA on or before Oct. 2nd. Please note that:

- People without group will be randomly assigned to form groups.
- Groups less than 5 people may be assigned members to keep the size of 5.
- Changing / Swapping groups are not allowed after the group list is finalized.

3 Proposal

Each group need to do one proposal presentation and submit one copy of written proposal. For the proposal presentation, you can get 1) a pass grade 2) some advice to modify you proposal. You will have one week in case of 2) to modify your written proposal and submit again.

3.1 Proposal Presentation

In the class of Oct. 24th, each group should do a 3-minute in-class presentation + 2-minute Q&A about your proposal, that describes:

- The dataset you are using.
- The problem you are exploring with your dataset.
- Expected outcome of your project.

This proposal is compulsory. You can decide the who & number of people in your group to do the presentation.

3.2 Written Proposal

Each group needs to submit a written proposal due at **5pm, Oct. 24th**. The proposal should be typeset in LaTeX, using the usenix2019 template: https://www.usenix.org/conferences/author-resources/paper-templates

Please use the template for 2019 conferences. The authors should include the name of every member in the group, rather than putting a group number only. The proposal should not exceed 2 pages including references.

4 Final Report

4.1 Presentation

A 10-minute final presentation is required to describe the outcomes and findings of your project. Everyone in your group should be doing the presentation and the timespan for each group member should be roughly the same. The time of final presentation is the last two lectures. The final presentation will be peer-reviewed.

4.2 Written Report

Each group needs to submit a final report by **5pm on Saturday of the week of your group presentation**. The report should not exceed 6 pages, including at most 1 page of References and Acknowledgements (see below). It should also be typeset in LaTeXfollowing the template of §3.2. Please note that there is no requirements on the minimal length of your report, so long as the contents of your project are clearly written.

References and Acknowledgements Your paper should contain References and Acknowledgements. The acknowledgements part should describe the specific contribution of each group member in this project.

5 Data Sources

Below are datasets and data sources for your reference.

5.1 Datasets

We provide some datasets that you can work with:

1. Bikesharing (NYC) https://www.citibikenyc.com/system-data
2. Taxi (NYC) http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
3. Public Service Data (NYC 311) https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9
4. P2P Lending (Kiva) http://build.kiva.org/docs/data/snapshots
5. P2P Lending (Lending Club) https://www.lendingclub.com/info/download-data.action
6. Yelp dataset https://www.yelp.com/dataset/challenge
7. Movielens https://grouplens.org/datasets/movielens/

5.2 Data Sources

Alternatively, you can find open data sources to download interesting datasets:

- Kaggle [1]. https://www.kaggle.com
- UCI Machine Learning Repository [2] http://archive.ics.uci.edu/ml

Or any other sources of datasets.

6 Grading Scheme

Grading Scheme will be released by Oct. 5th.

References

[1] Kaggle. https://www.kaggle.com, 2018.

[2] Dheeru, D., and Karra Taniskidou, E. UCI machine learning repository, 2017.