Project Assignments

CZ4032 Data Analytics and Mining (Data Mining)

Important Contact:

DAI Wenting DAIW0004@e.ntu.edu.sg
ZHANG Yu yu007@e.ntu.edu.sg
ZHANG Su su012@e.ntu.edu.sg
SONG Qun SONG0167@e.ntu.edu.sg

Due Date:

- Group Formation   :   (Week 4)
- Project Presentation :   (Week 12 and 13)
- Report Submission   :   (week 13)

I.   Introduction

This project aims at familiarizing the concepts of data mining learnt through the course in order to provide some insights into the topic of interest. With the available software (in Java) such as Weka (http://www.cs.waikato.ac.nz/ml/weka/) and others; and analyzing the datasets obtained from public databases.

Another aim is to promote teamwork and self-learning: Working in a team pays big dividends, it is less stressful and peer support does make learning much easier. Furthermore, in line with NTU vision to 'Teach less, learn more', you are encouraged to take the basic skills and principles, coupling with self-reading to handling real work problems.

II.   GENERAL GUIDELINES:

> This project is a required part of the course. It shall account for the main coursework component in your final grade.
> This project is to be accomplished in a *group of around FIVE (5) and Maximum Six (6) students*. A supportive and conducive environment within each group is beneficial; hence you are free to determine your own group members. While the accomplishments must genuinely belong to your group, you are free to have discussions with me and most importantly, your classmates. Utilization of the *discussion board* in the NTULEARN is strongly encouraged. **Marks** will be awarded to those actively participated and contributed in the discussion board.

III.   Project Topics for Data Analytics and Mining

Project Objective

The students are expected to practice hand-on skills for how to perform a real-world KDD (knowledge discovery from data) task from the beginning (data pre-processing - data collection, cleaning, etc) to the final stage (data post-processing - evaluation and presentation, etc).

You are encouraged to test your approach using existing tools such as Weka, R, Matlab toolbox etc. After familiarizing the basic steps and see the results, you are encouraged writing your own scripts and you may submit it together with your report.

Before You Start

Plenty of tools are available for data mining tasks using artificial intelligence, machine learning and other techniques to extract data. It is recommended you choose many of the freely available tools for data analysis. Simply search from the internet will return many lists of popular open source data mining tools available, for example:

- https://thenewstack.io/six-of-the-best-open-source-data-mining-tools/

- https://www.softwaretestinghelp.com/data-mining-tools/

As a start, you may try the "Data Mining and Predictive Analytics training course" using the open-source Weka tool. The tool is sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modeling. Its free under the GNU General Public License. Videos are produced by the University of Waikato, New Zealand. Who also authored the book: Ian H. Witten, Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Elsevier, 2005

Weka Predictive Analytics Tutorial
https://www.youtube.com/watch?v=Fg2x_zM3YTo&list=PLzVF1nAqI9VmC96TbvOPMkXToSmBMHJn7

Pay attention to the limitation of free and online tools in handling datasets. These will affect your selection and increase the challenge and complexity: Algorithms on a large dataset will run very slow. And the tasks on a very small dataset may be too easy.

IV.   Suggested Sources

The following are some example datasets for your consideration. However, you are strongly encouraged to propose your own new data analytics and mining problem.

1. Singapore's open data portal https://data.gov.sg/
2. Weka databases http://www.cs.waikato.ac.nz/ml/weka/datasets.html
3. UCI machine learning database http://archive.ics.uci.edu/ml/
4. Kaggle datasets for competitions https://www.kaggle.com/datasets

5. A topic-centric list of high-quality open datasets in public domains
   https://github.com/awesomedata/awesome-public-datasets
6. The PHM Data Challenge
   https://www.phmsociety.org/events/conference/phm/18/data-challenge
7. Aircraft Engine Simulator Data from NASA https://data.nasa.gov/dataset/C-MAPSS-Aircraft-Engine-Simulator-Data/xaut-bemq/data
8. Experimental acoustic and vibration measurements http://data-acoustics.com/measurements/bearing-faults/bearing-4/

As everyone might have chosen different data mining tools and dataset, you may consult my TAs for your choices. TAs will be able to help you with generic problems, data and problem-specific issue may need time to resolve, be kind to my TAs.

Reminders
- You are NOT allowed to COPY code/report directly from others / Internet (unless specified for special cases). Any plagiarism case will be seriously punished!
- For late submissions, a penalty of **1 mark** per day will be applied after the deadline. The assignment will not be accepted if more than a **7-day delay**. Please remember to submit your assignment before the deadline.

V. Marking Scheme

This is a rough guide on how a marking will be done.

Assignment Grading Criteria

- **Technical Depth**
  How challenging is your selected problem? How difficult is your selected methodology/solution? Is it trivial to implement the selected idea? What kinds of tools/knowledge/code required in order to implement the chosen approach?
- **Algorithm/Script Quality**
  How good and robust is your choose algorithm given the characteristic of your data; and how your scripts (or software) was designed?
- **The significance of Experimental Results**
  Are your experimental results significant? Can your results answer the question or achieve the objectives of your application?
- **Project Report**
  This is to evaluate the quality of your project report, including the organization, presentation, and comprehensiveness of the write-up.

VI.   What should be included in your project report?

Cover page: your group ID, your team members and their student ID (and their respective contribution in %)

- **Abstract**
  (use no more 300 words to summarize your whole project)
- **Problem Description**
  - Motivation
  - Problem Definition
  - Related Work
- **Approach**
  - Methodology
  - Algorithms
- **Implementations**
- **Experimental Results and Analysis**
  - Experimental Setup
  - Comparison Schemes
  - Results and Analysis
- **Discussion of Props and Cons**
- **Conclusions**
  - Summary of project achievements
  - Directions for improvements
- **References**
- **Appendix: (optional)**
  - **Datasets** *(if you collect your own data sets)*
  - **Scripts/Source Codes** *(if you implement your own codes)*
  - **Implementation Guidelines** (instructions for using any tools)

VII.   Submission Guidelines

What To Submit:

1. A file called **Project_Report_Group_XX.docx.** Please show your group members' names and IDs on the cover page of your project report. If you are using Latex, submit the latex source and the compiled PDF.
2. Source code (if any)
   - Well-commented code
   - Include Makefiles if necessary
   - Instructions on how to run your programs.
   - Remove the binary executable program if any
3. A README file. Please name it **README.txt** This file should include three sections:

- o   Your group ID and group member names
- o   Scripts or detailed instructions on how to reproduce your results using any toolbox.
4. The page limit of the report is 20 pages. The over-length case may be penalized. Please do not simply attach your source code in the report. However, if necessary, you can show some code segment or pseudo code to describe your key algorithm.

Submission Instructions

Submit your report to "Data Mining Assignment" under Assignments of NTULearn and **submit a hard copy of your project report** to my pigeonhole in general office.

If you have additional data which you want to submit, please package all of your files (including your report "Project_Report_Group_XX.docx" the README.txt file, and source code if any) into a ZIP file, named as "Project_Group_XXXXXXX.zip", where XXXXXXX is your group Leader name in upper case.

Submit the package file with the Subject "**PROJECT SUBMISSION GROUP** XXXXXXX", where XXXXXXX in the email subject is your group leader name in uppercase, to the following course E-mail: ntu.kdd@gmail.com and send an email to the respective TAs to keep them in the know.

VIII. Additional Reference
Data Mining, 4th Edition
By: Ian H. Witten; Eibe Frank; Mark A. Hall; Christopher J. Pal