CptS 315: Introduction to Data Mining
Course Project

Proposal (75 points)

You are required to write a 1 page proposal for your project as a pdf. Your proposal must include the following pieces of information:

1. Data Mining Task: What is your data mining task? This task could be a series of exploratory questions that you want to investigate or analyze. What is your motivation behind choosing this task for your project?
2. Dataset: What is the source of your data? Provide a link to your data source if you acquired it online.
3. Methodology: How will you solve the data mining task? You should have some idea of the algorithms or software tools you plan to investigate.

   Please feel free to use existing data mining and machine learning tool kits (e.g., Weka, Scikit-Learn) as needed for your project.
4. Final product: What will be the outcome of this project? How will you measure the success of your course project? Will this project help you explore or learn something new?

Report (325 points)

The project report (and code submission) should be written to address the grading criteria. (Ideally, you may want to write a project report that you can share with your recruiter when you apply for internships and full-time jobs at companies including Google, Microsoft, Amazon, Facebook etc. to improve your chances of getting hired)

Grading Criteria

- Clear Statement of the data mining task and/or questions.
- Methodology to solve the data mining task and/or questions.
- Evaluation of the methodology for the task and/or to answer the identified questions.
- Quality of the written report. This includes figures and illustration of the various concepts/algorithms and also the experimental results.
- Quality of your data mining code.

Written Report Format

The report will have 7 sections.

1. Introduction

   - Motivation from real-world applications for the data mining task you have chosen.
   - Give some examples of data mining questions you set out to investigate in this project.
   - State personal motivation to select this particular project and what were your goals.
   - Briefly describe the challenges and your approach to this task.
   - Briefly summarize your results.

2. Data Mining Task

   - Clearly describe all the details of the task. What is the input data? What is the output of data mining approach? Give examples to illustrate them.
   - List all the data mining questions that you set out to investigate in this project.
   - List the key challenges to solve this task

3. Technical Approach

   - Describe all the details of your algorithmic approach to solve this data mining task and/or answering the data mining questions.
   - How are you addressing the challenges mentioned above
   - An algorithmic pseudo-code and/or a figure (block diagram) to illustrate the approach will be good.

4. Evaluation Methodology

   - Explain the dataset and its source that you employed to study this task. Any specific challenges to use this data for your task.
   - List the metrics you employed to evaluate the output of data mining task and/or questions investigated. Justify their choice from real-world applications perspective.

5. Results and Discussion

   - Present and explain results in a step-by-step manner to tell us a story about what you have discovered by doing this project (all graphs and tables should be properly labeled with legends and captions. they should be self-sufficient to understand the results)
   - What worked and why?

- What didn't work and why not?

6. Lessons Learned
    - What did you learn by doing this project? In the hindsight, would you have made some different decisions to improve the project further?
7. Acknowledgments
    - Acknowledge all the sources of help you got to do this project.

When you turn in your report, you will include the following items in your zip file.

- A pdf file for the report itself.
- Source code for compiling and running your program.
- The data that your code is processing.
- A script file to execute your code on the given data (similar to regular homeworks).

Presentation (100 points)

The course project presentation will be evaluated in terms of "education" metric with the main goal of educating your fellow classmates about your project. You need to tell an interesting and engaging story about your project in less than 5 minutes. You can save all the low-level details for your project report.

Presentation Slides

Use a presentation application such as PowerPoint or Google slides to prepare your presentation slides. Be sure to write what you will say in the video presentation into the notes of your slides. This will give you a script to work from when you record and will provide a transcript for anyone with hearing impairment. Include as many pictures and figures as you need; the less text on your slides the better (save complete sentences for your notes section/transcript.)
Please plan on having around 5 slides ($\pm$ 2 slides) covering the following topics:

1. Motivation for your project
2. Precise problem you are addressing as part of your project
3. High-level solution methodology
4. Results

Your presentation should be self-sufficient (do not assume any additional domain knowledge) to meet the "education" goal as stated above. This kind of short and compact presentation skill is very important for internships and full-time job interviews in data mining and machine learning. Hence, this exercise.

Video Presentation

You will be recording and saving your presentation using Panopto in Blackboard. With the Panopto software, you will be able to record your voice while displaying your PowerPoint presentation. Please see the attached document for detailed instructions.
After recording and saving your video presentation, you will create a thread in the Discussion Board forum entitled, "Project Presentations," and place the link to your video presentation for others to view.

Critiques (50 points - 2 at 25 points each)

You are required to write a critique of 2 fellow students video presentations. Go to the discussion thread of the student you are reviewing and then describe at least 1 thing that student did well and suggest at least 1 thing they could improve. You will do this for 2 different students. Each critique should be about half a page if it were to be typed in document.

Possible Software Tools

- WEKA data mining tool kit (`https://www.cs.waikato.ac.nz/ml/weka/`)
- Sckit-learn (`http://scikit-learn.org/stable/`)
- XGBoost (`http://xgboost.readthedocs.io/en/latest/model.html`)