ECE657A: Data and Knowledge Modeling and Analysis
Project Description W2019

There are two broad types of projects: (a**) Application-oriented projects**: You have a problem, perhaps in your field of research, that you would like to analyze using the concepts and algorithms of this course, and (b) **Algorithm-oriented projects**: You select an interesting data analysis/machine learning technique that you want to learn more about. Then you find multiple datasets to test out the algorithm on and compare its performance against other algorithms. A good source of ideas for this type would be academic papers proposing a new or seminal algorithm. You could recreate the experiments in a paper and try to validate their results.

You may of course choose to develop or extend your own data analysis techniques, or simply apply existing techniques (they do not necessarily need to be covered in the class) to data. Similarly, the dataset may be a pre-existing one, or part of your work could be to collect and pre-process the raw data needed for a new dataset.

Note that pure literature surveys are *not* acceptable. There must be a hands-on, experimental and comparative element in your work.

Something new: One general requirement of the project is that somewhere in it you should have present something new which goes beyond what we've learned in class. This could mean you used an algorithm we did not discuss in detail, or at all. It could mean you go in depth into the original paper for some famous algorithm and re-implement their experiments to show the class. You may not be sure what the new part is until later in the project but you should highlight it in your presentations and reports.

Projects should be worked on in groups of 2-3 people. Doing the project on your own will be discouraged without a good reason. The project consists of proposal, presentation, and a written report. This document explains what is expected in each of these milestones. Marking schemes for the proposal, presentations and final report will be available on the course page in the LEARN system.

1 - Recommended Topics

- Possible Application-oriented projects

Sentiment analysis from social media	Topic discovery and analysis from text
Mining of scientific publications	Political event data analysis
Financial data mining	Purchasing behaviour analysis and recommendation systems
Crime Statistics Analysis and Mapping	Flu/Disease/Trend prediction from social media data

- Possible Algorithm-oriented projects
    - Comparison of performance of different algorithms in a sub area:

Cost-sensitive classification	Kernel-based clustering
Semi-supervised classification	Clustering ensembles
Frequent pattern mining	Supervised clustering
Neural network based classification	Decision tree based classification

-
    - Or find a recent paper introducing a relevant Data Analysis method and implement their experiments on new data.

2 - Proposal

Your proposal should contain:

- Description of the project. You should clearly mention your main goal: is it classification, clustering, or mining association rules or something else?
- A comprehensive review of 3 to 4 well-recognized research papers. At least a few sentences for each paper.
- A paragraph to discuss the expected challenges / difficulties.
- A sketch of your planned approach, algorithms, preprocessing methods, evaluation metrics, etc.
- Description of datasets that you plan to use. It should include a link and a brief description about the properties of the data, such as its features, instances, preprocessing techniques, etc. If you are going to use your own dataset, then a description of its source and preprocessing steps is needed.
- List of key references.

Your proposal should be no longer than 2 pages, and submitted as a PDF file via the LEARN dropbox. It will be graded. If not being approved, you will need to revise it based on our feedback

3 – Presentation

Your X minute group presentation would be via projected slides. Please allocate 3-5 minutes for questions. (note: the presentation length will depend on the number of groups that can be fit in the allotted days, it should be 15-20 minutes). Each presentation should include the following:

- Introduction: Basic definitions, background and terminology used.
- Literature review: Based on papers from your literature search, summarizing common variants of the method and data mining applications being used and the achieved results as claimed in the literature.
- Description of the goal and the use of method in your project, such as types of data mining, representation of the input, training requirements, output representation.

- Report and analyze your comparative experimental results.
- Something new: highlight one thing you discovered or explored that goes beyond what we talked about in class, or which goes into something mentioned briefly in class to much greater depth.
- A summary of your work: new findings and potential future directions.
- List of key references.

A copy of presentation slides is to be deposited to the LEARN dropbox before your presentation.

4 – Report

Your report should be in Springer LNCS format, as if you were planning to submit it as a conference paper. The suggested length of the report is ten (10) pages, but a maximum of up to thirteen (13) pages is allowed if you are having trouble fitting figures. Bibliography can be extra beyond this. See Springer's `Information for LNCS Authors' page with LATEX templates and a sample PDF (there is also a Word template but Latex is suggested).

The report should include the following:

1. Introduction to methods selected and task applied to.
2. Brief review of literature on the selected methods and their application to similar problems.
3. Description of the method selected with details on the options and parameters.
4. Implementation: Software used, data structures, program structures, data representation and any special set up needed. Don't put code in the report, only abstract descriptions and diagrams. Your code can be submitted to dropbox to be looked at separately. Data sets themselves do not need to be uploaded.
5. Testing: Test cases on the selected datasets and evaluation of the performance in comparison with base line methods (for example, K-means for clustering, KNN for classification, PCA for data reduction).
6. Discussion of results and conclusions: Provide a discussion on the use of the method and its suitability and/or limitations, and discuss the effect of each parameter on the trade-off in performance.
7. References to relevant literature you consulted about the data domain or the methods used.
8. Make sure the writing format is correct, spelling and grammar mistakes are removed and the description is clear and easy to follow.

The PDF format of report should be submitted via the LEARN dropbox, attached with related code, datasets, and other supporting materials.

5 - Timelines

Deliverable	Due	Grade
Proposal	Feb 15	5%
Presentations	Week of March 25, April 01?	10%
Report	April 5	15%

Late submissions up to 3 days are accepted with the penalty of 10% per day. You should not submit a work that you have performed for other classes, or have already been developed for your thesis. Although if you have two course projects this term that are related it is possible, but talk to the prof.

6- Plagiarism check:

Trunitin is going to be used in this term for academic integrity. You have the right to decline to do so and you can submit in the alternate Dropbox. However, a check for plagiarism will be implemented, so please refer to the university policies on Academic integrity.

https://uwaterloo.ca/academic-integrity/what-academic-integrity-0/integrity-policies