CSC 177 Data Warehousing and Data Mining (Spring 2019)

Final Project

The goal of this project is to provide you **hands-on experience** of applying data mining/machine learning techniques to one or more large data sets, going through the following steps:

- identifying data set(s)
- deciding on what you want to achieve with data mining
- choosing appropriate machine learning algorithms
- implementing your methods
- evaluating your methods on your data set(s)
- reporting conclusions by writing a paper

1. Deliverables

Here are the project deliverables and due dates:

Project Deliverable	Due Date
1. Project Proposal	On Canvas, 2 pm Monday April 8, 2019
2. PowerPoint File for Presentation	On Canvas, 2 pm Monday April 29, 2019
3. Project Report (IEEE format)	On Canvas, 2 pm Monday April 29, 2019
4. All Source Code (Jupyter notebook)	On Canvas, 2 pm Monday April 29, 2019

2. Demo Session:

Each team member must present the project work during the scheduled demo session. Failure to show up in demo session will result in zero point for the project. Each team has **15 minutes** to present your work in class. Use **visual aids** to make your presentations clearer and more interesting. The instructor will post presentation schedule on Canvas once receiving all the proposals.

3. Data Sets:

You are to find data set(s) for your project. The data set should be reasonably large and for a domain that you know something about or are interested in learning about.

See the separate document for a list of sample data sets.

4. Types of Projects:

Your grade is based on the novelty in your work.

Here are THREE possible types of projects:

- **Research-oriented** (Type A, difficulty degree = 1): Identify a paper recently published at an academic conference or journal. Read it thoroughly to understand the paper. Improve the proposed solution in the paper. Implement and evaluate your improved solution with the original solution proposed in the paper. You are required to do **feature engineering (i.e., feature normalization/encoding and feature selection) and parameter tuning.** Your goal here is to evaluate how your improved solution outperforms the original solution proposed in the paper.

- **Application-oriented** (Type B, difficulty degree = 1): Identify a well-known task or a dataset (e.g., a prediction/clustering problem from popular data mining websites such as Kaggle, UCI repository). Implement and evaluate **multiple models** in terms of their performance on the selected task and dataset. You are required to do **feature engineering (i.e., feature normalization/encoding and feature selection) and parameter tuning.** Your goal here is to evaluate how well each model performs for your selected task.

- **Survey-oriented** (Type C, difficulty degree = 0.9): You write a survey paper to provide reader with a state-of-the-art view of existing work on a particular data mining/machine learning topic. Do the following:
  - Survey the publications within a specific topic in last 5 to 10 years.
  - Summarize major accomplishments
  - Include your own comments on each surveyed paper
  - Summarize future research directions/challenges.

  Requirements and tips on writing a survey:

  - Everything you write in the survey has to be in your own words
  - All other people's idea/words must be correctly attributed to the actual researcher(s) using citations.
  - Pick a recent survey of the field so you can quickly gain an overview.
  - Go to the "related work" section of a paper, which serves as "a short survey" and allows you to find more related papers

5. Grading Breakdown

Your score of the project report is calculated as follows:

project score = your rubric score (see the next page for rubric) * difficulty degree

For example, if your rubric score is 90 and your difficulty degree is 0.9, then your project report score is 90*0.9 = 81

6. Where to find papers:

- Various digital libraries on computer science can be found at

  http://xerxes.calstate.edu/sacramento/new-databases/subject/computer-science

  including ACM Digital Library, IEEE Xplore and a few more…
- (Quickest way) The most comprehensive CS library:

  DBLP: http://dblp.uni-trier.de/
- Google Scholar

7. Teaming:

Students must work in teams of 2 or 3 people. Think clearly about who will do what on the project. Normally people in the same group will receive the same grade. However, the instructor reserve the right to assign different grades to team members depending on their contributions. So you should choose partners carefully!

8. Final Report Format:

See the separate final report formatting guideline for more details.

Project Report Rubric

Type Coefficient: **A(1)** **B(1)** **C(0.9)**

Team members:

Paper title:

Criterion	Percent	Score
Organization/Presentation: Good organization. Sections are logically ordered. Well-written. Reads with ease. No spelling/grammatical errors.	10 %	
Problem Statement / Motivation: The project's objectives are clearly stated. The motivation is clearly established by relating the project to related work.	10 %	
Design of Approach: Approach is clearly described. Design choices are mentioned and justified. Shows good understanding of material learned from class. **(Note: for survey paper, this counts for 0%)**	10 %	
Experimental Results: Experiments are well-designed and conducted to verify claims made in problem statement. Several baselines are performed to validate improvements. **(Note: for survey paper, this counts for 0%)**	10 %	
Analysis of Results: Thorough analysis of results, presenting not just raw experimental data but also conclusions. **(Note: for survey paper, this counts for 0%)**	10 %	
Related Work: Related work is acknowledged. **(Note: for survey paper, this counts for 40%)**	10 %	
Conclusion: Summarizes the problem statement, solution, and experimental results well to tell an overall story	10 %	
Work Division: Provides reasonable work division. Equal sharing of work.	5 %	
Learning Experience: Well-thought and reflective. Clearly states what you and your group learn from the project or from the class	5 %	
Overall Quality:	20 %	
Total	100%	