Final Project Instructions: CSCI 347, Introduction to Data Mining

Part 1 Due Date: April 26th, 2024 at 11:59 p.m.
Parts 2 - 4 Due Date: May 8th, 2023 at 11:59 p.m.
Presenting the selected videos and peer grading: May 9, 2:00-3:55 pm

This project may be completed in teams of 2-4 partners. Solo participation is not permitted and will result in an automatic deduction of 10% from your final project grade. So make sure you start early with finding teammates It's crucial to initiate the process of forming your team early on. While collaboration and seeking assistance from peers is encouraged, the work you submit must be distinctly your own. Should you utilize any online resources, proper citation is mandatory.
This project is much more open-ended than previous projects. You are encouraged to explore a data mining topic of interest. You may choose to dive deeper into a topic covered in class (ex: improvements/extensions of k-means applied to a data set of interest), or explore a related topic that we didn't get have time to cover (for example, a different clustering or classification algorithm, advanced feature selection or feature extraction algorithms, other items mining approaches, other graph models, etc…). The learning objectives of this project are to:

- Identify problems that can be solved or partially solved using data mining techniques.
- Apply appropriate data mining algorithms to a real-world data set using the Python programming language
- Construct an end-to-end computational pipeline to solve a data mining problem
- Explore a data mining application of interest

Keep in mind that we have limited time for this project. Some exploration may therefore need to be left for future work.

Part 1 [20 points]: (Due Date: April 26th, 2023 at 11:59 p.m.) Find a problem and a data set of interest. Describe your proposed approach to apply data mining to solve the problem.

Find a problem that you are interested in that has an associated data set. You can browse the UCI Machine Learning Repository, the SNAP collection, Kaggle, or any other source of publicly available data. Think about how you might apply data mining to this problem. Write one page proposal that includes:

- A concise overview of the problem you aim to solve.
- Identify the stakeholders involved. Describe the potential value your solution aims to deliver. Specifically, what types of benefits will it bring, and to which stakeholders will these benefits accrue?
- A summary of the dataset, detailing the number of instances and attributes, the breakdown of categorical versus numerical features, and, for graph data, the number of nodes and edges.
- A list of the data mining techniques you intend to employ in addressing the problem.

- A strategy for evaluating the outcomes of your data analysis.
- An acknowledgment of any aspects of your proposed solution that might need to be deferred to future work due to time constraints.

The proposal summarizing your proposed work must be turned in by **April 26th at 11:59 p.m.**

You are encouraged to visit office hours or send an email to the instructor to help develop your idea.

Part 2 [30 points]: Write code to analyze your data. This should include pre-processing such as missing value imputation and one-hot encoding, dimensionality reduction, and any data mining algorithms that you want to apply to your data.

Part 3 [40 points]: Write up a report summarizing your findings. Summarize the methods you applied, from beginning to end, including pre-processing techniques, dimensionality reduction, clustering or classification, etc… Include answers to the following questions in your report:

- What problem were you trying to solve or help solve?
- Describe the data
    - How many instances?
    - How many attributes?
    - Any missing values?
    - Number of categorical and numeric attributes?
- What pre-processing techniques did you apply and why (justify the use of each technique you used, for example label encoding vs. one-hot encoding)?
- What data mining techniques did you apply and why (justify the use of each technique you used, for example why did you use k-means instead of DBSCAN)?
- What techniques or metrics you used to evaluate the results
- Include relevant visualizations and tables summarizing your data and your findings
    - This may include a table listing the number attributes, missing values, number of classes, parameter settings, etc…, a visualization of a large graph if you are working with graph data, one or more visualization of your data in two dimensions (original dimensions or PCA dimensions), a plot of r vs. f(r) for PCA, a plot of the objective function for various values of k for k-means, a plot or table of the precision of a clustering for different parameter settings, etc…
- What did you learn through your analysis?
- Was anything about your results surprising or unexpected?
- How will your work help with understanding the problem you set out to solve?
- What else would you do if you had more time?

Part 4 [10 points]: Make a video presentation of your project.

Create a concise video presentation, lasting strictly **between 5 and 6 minutes**, to showcase the essence of your project. You may use Panopto or a platform of your choice for recording. Begin the video by introducing yourself and then proceed to:

- Provide a brief overview of the problem your project addresses.
- Detail the data mining techniques employed in your analysis, explaining why they were chosen and how they were applied to the problem at hand.
- Highlight the principal findings of your research, noting any unexpected outcomes or insights.
- Optionally, discuss potential future directions for your project if time permitted further exploration.

This presentation aims to encapsulate the entirety of your project journey, from problem identification to the analytical methods used, leading to your significant conclusions and learnings.
Please ensure your video does not surpass the 6-minute mark or fall short of 5 minutes, as deviations from the set duration will result in a penalty of 1 point.

Submit your video, along with the supporting code and written report, via Brightspace. Ensure the report is also uploaded to Gradescope.