Final Project Instructions: CSCI 347, Introduction to Data Mining

Part 1 Due Date: April 17th, 2020 at 9:00 p.m.

Parts 2 - 4 Due Date: May 5th, 2020 at 9:00 p.m.

This project is to be completed individually. You may ask others for help, but the submitted work must be your own. You may use online resources, but they must be cited.

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 15th at 9:00 p.m.**) Find a problem and a data set of interest. Describe your proposed approach to apply data mining to solve the problem.

[NOTE THAT PART I HAS AN EARLIER DUE DATE THAN THE REST OF THE PROJECT]

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 paragraph that:

- Summarizes the problem
- Summarizes the data set (how many instances and attributes, how many categorical and numerical features, how many nodes and edges if using graph data, …)
- Lists data mining techniques you would like to use to help solve this problem
- Describes what part of your proposed solution may need to be left for future work if you run out of time

The paragraph summarizing your proposed work must be turned in by **April 15th at 9:00 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)?
- 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.

Make a short video, that must be between 5 and 10 minutes, summarizing your findings. The TechSmith Relay Tutorial page has a tutorial to help you create your video. TechSmith Relay is freely available to all MSU students. The video should:

- State your name
- Summarize your project, including:
  - the problem you are interested in
  - what data mining techniques you used to analyze data related to the problem
- Your key findings and any surprising results
- You can also mention (but are not required to) what else you would work on if you had more time.

The goal is to summarize the work you've done and what you've learned from the process.

Turn in your code, report, and video on Brightspace. The report should also be turned in on Gradescope.