CSCI 347
Project 04: Final Project

This project may be completed individually or a group of up three people. You may ask others for help, but the submitted work must be from the group. 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 (for example, improvements/extensions of k-means), or explore a related topic that we did not have time to cover (for example, additional clustering or classification algorithms, advanced feature selection algorithms, non-linear dimensionality reduction algorithms, 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: Plan (20 points)

NOTE THAT PART 1 IS DUE EARLIER 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. For example, include how many instances and attributes, how many categorical and numerical features, how many nodes and edges if using graph data, etc.
- Lists the 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 the Part 1 due date. You are encouraged to visit the instructor or TA office hours to help develop your idea.

Part 2: Implement (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: Report (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? Make sure to 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? Make sure to 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.
  - visualization of a large graph if you are working with graph data.
  - one or more visualizations of your data in two dimensions (original dimensions or PCA dimensions).
  - for PCA, a plot of r vs. f(r).
  - for k-means, a plot of the objective function for various k's.
  - for DBSCAN, a plot or table of the precision at various parameters.
  - other visualizations or tables that you think will effectively communicate your ideas.
- 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: Present (10 points)

Make a 5-10 minute video presentation summarizing your findings. You may use whatever video editing technology you prefer. (The MSU supported tool is TechSmith Relay. See the UIT tutorial for more info.) 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.
- What else you would work on if you had more time.

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

Note: any presentation that exceeds 10 minutes or does not reach 5 minutes will be docked 1 point per minute.

Tips and Acknowledgements

Make sure to submit your code, report, and video on Brightspace. The report should also be turned in on Gradescope.

Acknowledgements: Project adapted from assignments of Veronika Strnadova-Neeley.