CSE 158/258, DSC 256, MGTA 461, Fall 2023: Assignment 2

Instructions

This is an **open-ended** assignment in which you are expected to write a detailed report documenting your results. Please submit your solution electronically via gradescope, on or before Dec 5 (Tuesday week 10). This assignment is worth **25%** of the final grade.

This assignment may be conducted **in groups of 1-4 people**. Groups of four are allowed, but possibly consider splitting your project into two groups of two unless you have something in mind in which really benefits from four members. The marking scheme is the same regardless of your group's size. Make sure to specify the names of all of your group members when submitting. Submissions should be in the form of a written report, which is expected to be at least four pages (double column, 11pt), or roughly 2.5-3 thousand words, plus figures, tables, and equations. See an example template in the lecture slides to get an idea of the length expected.

Examples of datasets and projects that may be of interest in this assignment will be discussed in the lectures, though you may use any dataset you wish (including the ones we used for Assignment 1). For a selection of datasets that I frequently use, see `https://cseweb.ucsd.edu/~jmcauley/datasets.html`

Tasks

Assignments will be graded based on their coverage of the following five components. Examples of what might be included in these sections and previous assignment examples shall be described in more detail in class. Each of the five sections below will contribute approximately 5 percent of your grade, for a total of 25 percent for the whole assignment.

1. Identify a **dataset** to study, and perform an exploratory analysis of the data. Describe the dataset, including its basic statistics and properties, and report any interesting findings. This exploratory analysis should motivate the design of your model in the following sections. Datasets should be reasonably large (e.g. large enough to run the kinds of methods we've discussed in class).

2. Identify a **predictive task** that can be studied on this dataset. Describe how you will evaluate your model at this predictive task, what relevant baselines can be used for comparison, and how you will assess the validity of your model's predictions. It's also important in this section to carefully describe what features you will use and how you had to process the data to obtain them. Make sure to select a task and models that are *relevant to the course content*; if you want to try out models you've seen in other classes that's fine, but you should still implement models from this class as baselines / comparison points.

3. Describe your **model**. Explain and justify your decision to use the model you proposed. How will you optimize it? Did you run into any issues due to scalability, overfitting, etc.? What other models did you consider for comparison? What were your unsuccessful attempts along the way? What are the strengths and weaknesses of the different models being compared?

4. Describe **literature** related to the problem you are studying. If you are using an existing dataset, where did it come from and how was it used? What other similar datasets have been studied in the past and how? What are the state-of-the-art methods currently employed to study this type of data? Are the conclusions from existing work similar to or different from your own findings?

5. Describe your **results** and conclusions. How well does your model perform compared to alternatives, and what is the significance of the results? Which feature representations worked well and which do not? What is the interpretation of your model's parameters? Why did the proposed model succeed why others failed (or if it failed, why did it fail)?