

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 on gradescope 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.

Note that you should exclude the name of your group members from the report itself if you want to be graded anonymously.

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)?

Peer grading

After the assignment is submitted, a **peer grading** assignment will be opened on gradescope where you will evaluate the work of your peers (and they will evaluate yours). Peer grades are due by Dec 11 (Monday week 11). Participating in peer grading is worth **4 marks**.

Instructions for peer grading are as follows:

1. Start the assignment by submitting an empty “responses.txt” file; the autograder will print out what submissions you need to grade and will tell you the format your responses should take.
2. Each question should be graded out of 5. Given that there’s some flexibility in these questions, you can assign up to 2 bonus points for any question (so that the score can be out of 7); if doing so use the word “bonus” in your evaluation for that question.

Although the assignment is fairly open-ended, I’d suggest using something like the following rubric for each of the five sections:

0 marks The assignment does not attempt this section.¹

1 mark Some minimal attempt has been made to include this section, though it fails to include any of components described above.

2 marks Includes some of the correct components, though is still fundamentally unsatisfactory in some way, or contains serious errors.

3 marks A minimally-viable solution to the task. This might mean, for example, an assignment that closely follows a homework task, but fails to explore any new dataset, feature design strategy, etc. Or, a task and models that have little relation to course materials.²

4 marks A solution that is essentially satisfactory, though falls short of better solutions you’ve seen, e.g. fails to explore a variety of features or modeling strategies, or only cites very basic literature, etc.

5 marks A solution that is strong in all components.

Historically, we’d expect the average mark for each section to be a little above 4/5.³ However a mark of 5/5 should be perfectly achievable by meeting all of the desired criteria, rather than requiring exceptional performance. Bonus marks may also be used for the sake of flexibility, acknowledging that some assignments may be trivial in some components while going above and beyond in others.

Please make lots of nice and constructive comments!

Receiving your grade

After peer grading is complete, the autograder will be modified to show grades from your peers. Most assignments should have been graded by 5-8 people (though this may vary depending on section). Your grade for each question will be the median of your peer grades for that section. The final grade is the sum of medians.

Note that even if you don’t have time to participate in peer grading, you’ll need to upload an (empty) responses.txt file in order to see your peer grades on gradescope.

¹Note that some assignments may rename or re-order sections; try to be flexible as long as the relevant material is included.

²Though again, be flexible; it is completely fine to use datasets very different from those in class; and it is fine to focus on regression, classification, or text rather than recommender systems.

³Though you are not expected to curve assignments or match a specific distribution of grades.