CS 439 : Data Science Course Project Guidelines

General Instruction

The purpose of this project is for you to pose a question on some subject, find data to answer that question, and answer it using data science techniques. The nature of the data and the subject are deliberately open-ended – the process and analysis are what is being emphasized and what is graded. I encourage you to pick a problem you are excited about and will be flexible if the project is relevant to topics and research papers in lectures. A project can be done individually (not recommended) or by a group of max 3 students (recommended, but no more than 3).

Group Formation and Submit the Project Topic -

You can form a project group (of max 3 people) and decide/submit the project topic by Nov 13th

Project Guidelines:

Formulate a question you would like an answer, or a problem you would like to solve, in a subject you are interested in. Then, find a dataset that you think can help you answer this question or solve that problem. A few basic kinds of projects are as follows:

1. Unsupervised learning project: This kind of project is based on answering open-ended questions about a dataset. Do data exploration, draw conclusions about what features of the dataset matter and what don't, note surprising features of the data, and visualize lots of things to make these conclusions legible. These projects will be graded on your correct implementation and use of the statistical techniques you learned, as well as how clearly your visuals convey insights about your data
2. Supervised learning project: This kind of project is built around producing a predictive model that solves some problem, using the data set as an input. Use data exploration and visualization to make decisions about what kind of model to use, and then train a supervised model on your data. These projects are graded on your correct implementation and use of machine learning techniques, as well as the performance of the model you trained.

3. Other: If there is a specific project you would like to do that doesn't fall into either of the other categories, ask and we'll be happy to discuss it with you.

For all projects, discuss shortcomings of your methods and provide suggestions for improving your analysis. Issues pertaining to the reliability and validity of your data, and appropriateness of the statistical analysis should be discussed here.

Extra Notes:
- Start Early: Start thinking about your project early and spend enough time developing it.

Deliverables

1. Project Proposal

Due: 23:59pm, Feb 25th, 2025 (encouraged to submit early)

Purpose: The written proposal should define your project/research question and explain what you are planning to do.

Length: 2 pages (no more than 3), typed, single-space.

Your proposal should have the following sections, each of which answers a number of questions about the project

Introduction:

For supervised learning projects:

What problem are you solving?

How do you plan to solve it?

For unsupervised learning projects:

What question are you trying to answer

How do you plan to answer it?

How does this approach relate to the lectures/papers we discussed?

Motivation:

Why is your project important? Why are you excited about it?

What are some existing questions in the area?

Are there any prior related works? Provide a brief summary.

Method:

What dataset do you plan to use?

What form does this data have? Is it images, raw text, tabular, etc? What are the features?

For unsupervised learning projects:

- What analysis do you plan to do?
- How do you plan to visualize your results?

For supervised learning projects:

- What kind of model do you plan to use?
- How will this model's predictions help you solve the problem?
- Why do you expect this to work better than existing methods?

What would be your implementation steps? How will you evaluate your method?

How will you test and measure success?

Discussion:

What outcome do you expect from your results?

Are there any potential problems you foresee with your approach? What assumptions are you making about the problem?

2. Final Report

Due: 23:59pm, May 8th, 2025

Length: Normally, a well-explained project would take 6-8 pages, typed, single-space.

Introduction:

For supervised learning projects:

- What problem are you solving?
- How do you plan to solve it?

For unsupervised learning projects:

- What question are you trying to answer
- How do you plan to answer it?

How does this approach relate to the lectures/papers we discussed?

Motivation:

Why is your project important? Why are you excited about it?

What are some existing questions in the area?

Are there any prior related works? Provide a brief summary.

Method:

What dataset did you use?

What form does this data have? Is it images, raw text, tabular, etc? What are the features?

For unsupervised learning projects:

What analysis did you do?

For supervised learning projects:

What kind of model did you use?

How did you define the problem/feature space?

What would be your implementation steps? How will you evaluate your method?

How will you test and measure success?

Results:

For unsupervised learning projects:

What results did your analysis show? Visualize them if possible

What new questions do these results raise, and how can they be addressed by further analysis?

Repeat as necessary

For supervised learning projects:

How did you model perform?

Analyze important performance metrics such as accuracy, recall, false positive/false negative, MSE, etc as appropriate

How does this method compare to existing methods?

Visualize your results

Discussion:

What outcome did you expect from your results?

How did your actual results differ from your expected results?

If your final report differs from your proposed project, discuss the differences, why you made certain changes, and the bottlenecks that prevented you from proceeding with the proposed project.

3. Code submission

Due: 23:59pm, May 8th, 2025, with the final submission

Your code should be submitted as a single Jupyter notebook. The code should open your dataset, perform any analyses you use in your paper, and produce all visualizations. Your final submission should do all of these in a single run, so that I can see all your results from running your notebook once.

Note to all: You may use tools to help with your writing but do not use generated contents directly. Please cite any tools, web sources, papers, and textbooks you consult/use. You are responsible for the content of your writing, including its originality and correctness. Plagiarism is not allowed.

Sample Rubric

Project proposal (20 points): should be turned in on time and student should have discussed with me before submission

Project implementation (80 points):

Project Design (25):

Methods and analyses used are appropriate and answer the questions they are mean to answer (15)

Student shows understanding of the techniques and their applications (10)

Code: 35 pts

Methods used are correctly implemented, and data is not mishandled in a way that invalidates the analysis (25)

Code is readable, well formatted, and reproducible (10)

Paper: 30 pts

Answers all questions in the guidelines (15)

Visualizations accurately convey the points they are meant to (10)

Paper is well-written and easy to follow (5)