CptS 475/575: Data Science
Fall 2021
Semester Project, Description

Last Updated: October 4, 2021

Overview

The semester project constitutes a significant part of your work for this course (40% final grade). It offers a great opportunity to integrate and apply some of the things you learn in the course and to further explore a specific aspect. You are encouraged to work with one or two other students, forming a team of two or three. You have some degree of freedom to define the topic and the scope of the project as long as: the project clearly falls within the realm of the course, is likely to be doable within the remainder of the semester and is reasonably novel and interesting. I have posted a collection of project ideas. You can choose one from that collection as your project or propose your own.

Project types

Every project will involve some combination of the steps of the Data Science process (data collection, data processing, exploratory data analysis, model building, product building and/or communicating), but projects would vary in their relative emphasis on each of these aspects. Roughly speaking, a project could take one of the following general forms:

- analyze an interesting dataset and build a predictive model for it

- analyze an interesting dataset and draw insight, discover knowledge

- develop a novel method (for supervised or unsupervised learning)

- conduct a study comparing models/algorithms for a chosen problem

- build a data product aimed at a specific purpose

- create a visualization of a complex dataset

It is possible (and desirable) for a project to be a combination of some items among those listed above.

Submissions

There are three submissions associated with a project: Project Proposal (due date October 11), Progress Report (due date November 3), and Final Report (due date December 13).

Project Proposal

If you are proposing your own project, this would be a short (maximum 2 pages long) document that accomplishes the following:

- Describe what you intend to do

- Describe the methods you plan to use or develop

- Describe the data you will use and discuss how you plan to obtain it

- Discuss relevant background work

- Discuss your tentative plan

If you are picking one among the suggested project ideas, your proposal would only specify which project it is and the angle you have taken (but no need for a detailed description). In order to be able to get a good diversity of projects, in addition to your first pick, I'd like to get your second pick as well. I'll try my best to assign you your first choice, but I might need to go with the second.

In either case (whether you propose your own project or pick from the idea set), your "proposal" should list the names of the team members. If you haven't identified a team mate, indicate that and I will team you up with someone. There will be only one submission per team.

Progress Report

This will be a short (max two pages) document where you briefly report on how the project is going, and discuss if there are any course corrections you made.

Final Report

The outcome of the project will be a final report of about 8 to 12 pages long, an associated code/data product, and a short presentation in class. The nature and organization of the report will of course depend on the type of project undertaken, but its content should roughly map to the following rubric:

- Introduction/Motivation/Problem Definition
  (where you state what you are trying to solve/achieve and why it matters)

- Model/Algorithm/Method
  (where you give a detailed description of your work)

- Results and findings
  (where you interpret the results you obtain, discuss implications, make observations and draw conclusions)

- Related Work
  (where you cite (and briefly summarize) other work related to yours.)

I will later post a more detailed guideline on how to writeup a report.

Presentation

The last piece of the project will be a short (three minutes) presentation by each team summarizing their project, and an accompanying two or three minutes Q&A session in class for each team. The presentations are scheduled for Week 16.