CS4262/5262: Foundations of Machine Learning
Vanderbilt University, Fall 2021

Final project information

The final project is an opportunity to investigate machine learning approaches on a problem of your choice, and to gain more experience working with "real world" datasets. You are encouraged to identify a problem and domain that interests you! Please feel free to contact us anytime for feedback or assistance.

Teams

You may work in teams of up to 3 people.

Timeline

- Proposals: **October 5**
- Update (interim report): **November 7**
- Presentations: **December 7, 9**
- Final report due: **December 15**

All deadlines are at 11:59pm on the above dates, and no late-days may be used for the final report.

Guidelines

- Choose a dataset and the main question you will address, and develop a plan for your investigation. What is the primary approach you will take? How will you evaluate the performance of your approach against other candidate methods? You can experiment with elements such as feature selection / dimensionality reduction strategies / data pre-processing, etc., and examine the impact of these steps on your results. Your project may focus on supervised or unsupervised learning.
- Given the limited time-frame of the semester, we recommend choosing a dataset that is (a) readily accessible to you, and (b) doesn't require large efforts to prepare from its raw form before ML techniques can be applied (though we always encourage visualizing, quality-checking your data, and pre-processing the data in ways that can improve its quality!)
- Use of ML libraries: You are not required to implement ML algorithms (like SVM, logistic regression) from scratch. You may draw upon routines from scikit-learn and other toolboxes.
- Use and citation of resources: There are numerous online repositories with code / notebooks for mining certain public datasets. While you are welcome to read over existing code for inspiration, please do not copy outside code for your project, and try to do something a bit different from the examples you find. You must also acknowledge resources you have drawn upon for the project.
- If you would like work on a deep learning project, please make sure that you incorporate other material you learned in the class as well. For example, you might also try logistic regression or SVM, or unsupervised learning techniques such as k-means clustering, to look for interesting structure in the data.

- If you're conducting research in a lab, you are welcome to use data from your lab (provided you obtain permission from the principal investigator). Your class project can be aligned with the goals of your research, but should be something you haven't already tried.
- Graduate students (CS5262) will be expected to undertake a more in-depth literature review, and to motivate the project in the context of existing work. Graduate and undergraduate students can work together on the same team; in such cases, the report would be expected to have the quality of literature review and motivation that is expected of graduate students.

Project deliverables

All submissions should include the project title and names of all team members, and should be uploaded to Brightspace. Please submit all documents (except code) in PDF format. Only one person per group needs to submit, on behalf of the group.

Proposal (due: **October 5**)

Your proposal should be brief (between ½ - 1 page long) and include the following elements:
- Which dataset will you use, and what question(s) will you address?
- What method(s) will you apply, compare, or build upon?
- How will performance be evaluated?
- What are the anticipated steps and timeline for your work? If you are working as part of a team, how do you plan to collaborate and divide up the tasks?
- List 1-3 related references that you will read as background

Discussion of proposals: During class, each group should plan to briefly share their project idea with the class (5 minutes max, slides optional).

Update (due: **November 7**)

The aim of this report is to ensure that you're on track. This document should 1-2 pages (excluding references), and should contain:
- Data exploration: what dataset are you using, and how have you tried to explore / visualize / familiarize yourself with it? What pre-processing steps have been performed, and what features were (or will be) extracted?
- Preliminary results: describe the ML techniques you have attempted so far, and any results you may have obtained. You may provide 1-2 figures if helpful.
- Teamwork: if you are working as part of a team, how have you been collaborating and dividing the work?
- Next steps: based on your progress and any challenges you have encountered, what are the next steps you are considering?

Project presentations (week of December 7 and 9)

We will share and discuss the final projects during this week. Each team will give a brief presentation (approx. 10min).

Final report (due: **December 15**)

The final project report should be at most 5 pages long. References may be included on additional pages.
The report should include:

- Introduction: describe the dataset and problem you are addressing. Provide some background and discuss/reference previous literature.
    - **Graduate students (CS5262)** will be expected to conduct a more in-depth literature review into the background of the problem and machine learning techniques that have been previously applied, and describe how your work fits into this context. Your introduction does not need to exceed 1.5 pages.
- Methods, including a description of:
    - Dataset, pre-processing, and features extracted
    - Your approaches and experiments, e.g.: machine learning algorithms that were applied / compared, approaches for selecting hyperparameters and for assessing training and test-set performance (e.g. cross-validation), and performance metrics.
- Results: report your findings here. Figures should include captions.
- Discussion and Conclusions: Summarize and discuss your findings, bringing in concepts we have covered in this class. Mention challenges, unexpected findings, and possible avenues for future work.
- Contributions: If you are working as part of a team, please briefly describe the contributions of each team member. If you have sought advice from other researchers at Vanderbilt, please acknowledge their assistance.
- Resources: if you referred to any Kaggle write-ups or other code from the web, please list them here.
- Code: Provide a zip file or link to any code you have written for the project. You do not need to include external toolboxes/libraries with your code submission.

Clarification/concerns:
At any stage, please let us know if you would like to discuss your project, need any clarification regarding the expectations, or if any concerns arise.