MAST30034 Final Project

Semester 2, 2020

Contents

1 Introduction

A major goal of this subject is to help you develop skills that will assist you in your future career. The final group project gives you a concrete opportunity to hone your communication, co-operation and problem solving skills.

2 Project components

The project consists of four components. Here we walk you through what each component entails.

2.1 Proposal

For this first milestone, you will form a team and pick a project idea. The proposal must describe:

- Project title: keep it succinct.
- Team: include the full names of all team members. To make administration easier, you must pick your group members from your lab session. (No exceptions to this rule unfortunately.)
- Dataset: find a challenging dataset (see FAQ on what might be considered challenging). UCI machine learning repository and Kaggle are excellent sources.
- Identify task(s): state the goals of your project, e.g., develop algorithm to predict whether a chest X-ray is healthy or not for a dataset of X-ray images.
- Method: state the statistical and machine learning techniques you plan to apply to solve your task.

Submission Submit your proposal as a PDF document on Canvas.

What we're looking for With this first milestone, we want to check you are on the right track. You will receive full marks as long as you follow the instructions above.

2.2 Meeting minutes

This project component will give you an extra incentive to schedule regular meetings with your team members. We have set aside certain lab sessions throughout the semester for your team to meet easily. Members should take turns to write down the meeting minutes, noting items discussed and actionable to-dos.

Submission Collate all meeting minutes into a PDF document and upload to canvas. The formatting is at your team's discretion.

What we're looking for We want to see that you were indeed working regularly throughout the semester with your team, the exact frequency (weekly, bi-weekly) is at your discretion. There is no correct or incorrect meeting minutes. Don't make this hard on yourself by waiting until the end of the semester and fabricating the meeting minutes.

2.3 4-minute video recording

In this component, each team will prepare a *pre-recorded* 4-minute presentation. You may consult past NeurIPS video presentations for examples of recordings.

Submission Submit the link to your recording to Canvas.

What we're looking for Slides that are well written with a logical ordering. Voice-over narration that complements the corresponding slides.

2.4 Final report

The biggest component of the group project is the final report. You must use the NeurIPS 2020 LaTeX style file with the **preprint** option. Submissions are limited to **eight** content pages, including all figures and tables. Additional pages may include references only. We recommend the following structure for your report:

- Abstract: one paragraph describing the problem you are trying to address, the method you applied, and the results you obtained.
- Introduction: give us the context of your problem
- Dataset: describe your dataset, do exploratory analysis, produce visualization
- Methods: describe your statistical models and learning algorithms. Remember the instructor, tutors and other students are your audience, i.e., do not spend 1 page telling us what linear regression is. If you end up using some cutting-edge algorithm, e.g., Transformer for NLP, please explain the technique in detail as you may no longer assume it is common knowledge.
- Experiments and results: detail the results you obtain from applying the method to your dataset
- Discussion: reflect upon the lessons you have learned. Did the super cool algorithm you were itching to try end up being disappointing? If you had more time, more team members, or more computational resources, what else would you have explored?

Submission First, you must submit your report as a PDF following the NeurIPS 2020 style. Next, to ensure we can reproduce your results, you must also submit the code that produces figures/tables in the report. To fulfill this reproducibility requirement, you can either 1) link to your Github repository in the PDF report, or 2) submit the code as a zip file on Canvas.

What we're looking for We take very seriously that data science work should be reproducible and transparent. As such, we are looking for honest exploration of the data, honest application of methods, and honest reporting of the results. For example, it's perfectly acceptable if your method doesn't beat the other Kaggle submissions on predictive accuracy. In fact, it's completely acceptable for you to write about methods you applied that disappointingly failed, as long as you critically examine what may have produced the failure.

3 Critical dates

The final project is worth 50% of your course mark. The following table summarizes how much each project component contributes to the final mark, as well as the submission format plus due date. *Please designate one person on your team to do all of the submissions.*

Project component	Total mark	Submission format	Due date
Proposal	5	PDF to Canvas	End of Week 7: 17:00 Friday September 18th
Meeting minutes	5	PDF to Canvas	End of Week 12: 17:00 Friday October 30th
4-minute video recording	15	Video link to Canvas	End of Week 11: 17:00 Friday October 23rd
Final report	25	Canvas	End of Week 12: 17:00 Friday October 30th

4 FAQs

1. What do I have to turn in?

   The project has four deliverables: a) proposal, b) meeting minutes, c) 4-minute video recording, d) final report.

2. What methods can I use to analyze my data?

   Anything you want! You are not restricted to only use methods reviewed in this class.

3. Where can I find datasets?

   Kaggle and UCI machine learning repository are good places to look for publicly available datasets.

4. What type of datasets are acceptable?

   In order for your data analysis to be interesting, the data should exhibit some challenging characteristics. Here are some examples.

   - Data with large sample size, i.e., large n
   - High dimensional data, i.e., large p
   - Spatial structure, e.g., images
   - Temporal structure, e.g., time series
   - Missing data
   - Censoring
   - Extremes

   It is only acceptable to choose a simple toy dataset if you devise new methodology in your work, in which case the toy dataset can be used to illustrate aspects of your novel methodology.

5. What are acceptable team sizes and is grading a function of the team size?

   We strongly recommend teams of 3-4 students. If you have a very large group, we will have higher expectations of all the components.

6. Are we required to use <insert programming language> for the project?

   Use anything you want. We highly recommend either R or Python.

7. Can we use existing libraries such as scikit-learn in Python, or must we implement everything from scratch?

   You can use any library you fancy. Please cite the library in the final report if the library asks you to. If you are copying and pasting large chunks of code, you must acknowledge the source in the comments of your code.

8. What do we do if we get stuck with our analysis?

   Each team is housed within a lab session. Your tutor is a great resource to consult if you need help with the analysis.

9. Is my team required to communicate via Slack?

   No, you are not required to use slack but it is a popular communication tool to familiarize yourselves with. You can easily share new ideas and leads on Slack, post meeting minutes at the end of your meetings to make sure everyone is on the same page, etc.

10. Are we required to use GitHub?

    No you are not required to use Github, but we can't imagine how else your team will easily collaborate on code. (The answer is not Dropbox!)

11. What is the late policy for the final project?

    Each team will be allotted *two* days of grace to allocate as they see fit. For example, if you use a grace day for the proposal and then a grace day for the oral presentation, you have no more grace days for the final report.