Final Project

CS412

Released: March 31st

Due: May 8, 11:30pm on Gradescope

1 Project Overview

This will be the final term project for CS412, Intro to Machine Learning. In it, I expect you to demonstrate a complete machine learning project from data to final report. The project can be completed in groups of up to three. This project should demonstrate an understanding of several topics in the course as well as an ability to find and process information from the discipline of Machine Learning as a whole

2 First submission

Your first submission **Due April 10, 11:30 pm, no late days** should be approximately one page and include the following.

- Your team members (Only one team member should submit on gradescope)
- Which dataset(s) you intend to use
- Which ML method from outside of class you intend to implement
- What you expect your principal conclusions to be.
- Three options for a 30 minute time frame where you and your group are all free to meet online during the week 4/13-4/17. I will be available those days from 10-6 (other than class) and time slots will be awarded in a first-come-first-served basis.

This submission is non-binding and as long as it is clear to me that you and your teammates have put some thought into your project and how you will proceed, you will receive full marks for this portion. I will also give some guiding remarks to help you with your project.

3 Team Meeting

Once the meeting time is set, you will receive an email over the weekend about which time you will meet, your team will meet with me to discuss your progress on the project. At this meeting, I expect that the team has completed

- Processing of the chosen data set
- Prinicpal statistical analysis of the data set
- A short presentation on the project proposed
- A literature review that shows you have read and understand relevant research documents that you may need for your project
- A timeline/project plan about which team members will do which elements of the final report and when.

4 Final Report

The primary grade for this assignment will be the written report. This should include all of your process from beginning to end in completing this project. As a general rule, I will expect the following page lengths, single spaced with figures, not counting the bibliography.

- Single Person: 7-10 pages
- Two people: 10-15 pages
- Three people 14-20 pages

While there is no strict page requirement, this is roughly the level of detail I expect for this report. I strongly encourage you to find a project that is interesting to you. For the following sections, each are required, however, do not feel the need that all sections be equally thorough. Focus on what interests you the most.

4.1 Dataset

Your project needs to include at least one dataset other than the digits dataset given in the course. It should also be some dataset that is not binary classification with numeric features. While only one new dataset is required, I encourage you to apply your analyses to multiple datasets as a way to make your final analyses more robust. As posted on the slides, there are several places to find interesting datasets and you should not hesitate to reach out to me if you are having trouble finding one.

For a more impressive report, you should select datasets that have unique properties. For example, missing data, regresssion, time-series and other datasets that have not been covered in class are likely to make for a more impactful final report.

4.2 Literature Review

This section is primarily focused toward graduate students. In your report, make sure that you identify which research papers you are incorporating into your project. This literature review should show that you read and understood the principal arguments of the paper and what exact elements you intend to extract for your implementation portion.

For undergraduates, this is a great place for you to introduce whichever machine learning principle from outside of the course you intend to implement. It does not need to be cutting edge, nor does it need to be complicated, however it should be very clear what the model or models actually are.

4.3 Machine learning applications

You will be required to perform some implementation of your models. I strongly encourage you to use SVM, Neural Network and Adaboost as comparisons, but you will also need some other approach. If you use code from anywhere other than the sklearn package, you should ask permission from the author(s) and cite their work appropriately in your bibliography.

While only one additional ML approach is required, including more modeling types will make for a more impressive final project. In this portion, you should include tables and figures as appropriate to provide data to the reader in an easily digestable fashion. There is no strict requirement for the number of figures, but make sure all of your figures are instructive and contribute to the ultimate argument of the paper.

4.4 Analysis and conclusions

After collecting your data, you should provide some analysis and principal conclusions of the work. This is the most important part of the project. In it, you should describe your approach, why you selected the model and data that you selected and and what conclusions you can draw about the model.

In this section, you should also include a concentration bound on the final reported error for your project for all datasets and models tested. A thorough discussion of the results and the trends is expected. While there are no particular questions you are required to answer, I expect some explanation of any unusual phenomenon that you observe. Figures and tables will be very helpful in your final demonstration

4.5 Bibliography

This is a university assignment. Therefore, I expect that all work is properly appropriated. You are free to use any code or concept so long as it is clearly cited in your bibliography. You are welcome to use any bibliography style that is prevalent in your field, but please make sure that you are consistent. **Any report that does not have a bibliography will receive a zero**

4.6 Grading

Because I want these reports to be as open ended as possible, it is difficult to assign a strict rubric to the work. As a result, I will assign one of four grades, along with feedback, to your final report.

- C: 20/30 or less. A C project fails to meet the minimum requirement of the assignment.
- B: 25/30: A B meets the minimum requirements of the assignment. Notably, a new dataset, a new machine learning model type and adequate analysis of the data collected. For graduate students, this also requires the implementation of a recent research concept.
- A: 30/30: An A project exceeds expectations in at least one area. Possible A projects will have multiple datasets, multiple models and a very thorough analysis. Figures and tables will be used effectively and the authors work should be sufficiently intensive.
- A+: 35/30. I will award up to 5 extra credit points for papers that are truly exceptional. The topic of the paper should focus on a more advanced topic of machine

learning discipline and should also demonstrate an advanced ability to synthesize material not presented in the course.

5 Individual assessments

As a final submission with your assignment on May 8th, each team member will submit an individual assessment. Details for this assessment will come out as the due date gets closer. However, results from these assessments may affect final grading if it is revealed that a project member did not fully contribute. Team work is a difficult skill to learn, but at this academic level, I would hope all teams work well together and this will not be a problem.

Making your report

When you submit, there will be two submissions on gradescope. One for your pdf report and another for your zip file containing all your code. Please include a list of any packages you downloaded in your report.