DS-GA 1003: Machine Learning

Project Requirements

Throughout the project, groups will have to work collaboratively to manage expectations and meet goals.

Groups should contain 2-5 members. If you would like to determine your group, then please post to Piazza under the *Project* thread to contact classmates.

The project will be in a competition hosted through Codalab. Groups will be working on the same datasets. Specifically, we have provided two tasks (details below) and each group can choose one.

Project 1:
https://worksheets.codalab.org/worksheets/0x33171fbfe67049fd9b0d61962c1d05ff
Project 2:
https://worksheets.codalab.org/worksheets/0x0a35e4ca487b4892976188108704011c

The test set is hidden so each group must submit their code and model on Codalab so that we can evaluate on the test set. While we will maintain a ranking, we will not evaluate groups against each other according to rank (details below) .

While it's not required, we strongly encourage all reports to be typed in Latex. The typesetting platform Overleaf could be useful for sharing reports with teammates.

Timelines

March 27th[t]: Project Proposal (week after Spring break)
May 19th: Codalab submission and Project Report (week after the final exam)
All deadlines are at 11:59pm.

Project Proposal

Groups must upload a one page pdf file on Gradescope containing:

- Title
- Group
    - Name and NetID of each member.
    - Member responsible for uploading submissions.
- Summary of Plans
    - Which project you have chosen
    - Proposed Approach
    - Suggested Experiments

NOTE: Only the *member responsible for uploads* needs to upload the pdf file. In other words, each group should have only one pdf file uploaded on Gradescope.

Codalab Submission

To get you started on Codalab, we have prepared a tiny project so that you can get familiar with running and submitting models on Codalab. See details in the Github repo under the project/ directory.

You are responsible for ensuring that the model runs on Codalab so we suggest that you start on project 0 now and leave at least one week for model submission.

Project report

Groups must upload a pdf on Gradescope describing their approach and results.

1. Title
2. Group Members
    a. Name and NetID of each member.
    b. Member responsible for uploading submissions
3. Introduction
    a. Description of Problem
    b. Approach
    c. Summary of results/contribution
4. Approach
    a. Describes the details of your approach

5. Experiments
    a. Description of Datasets
    b. Baselines or other approaches for comparison
    c. Explanation of Results
    d. Error analysis
6. Discussion
    a. Evaluation of Findings
    b. Possible Next Steps

NOTE: Only the *member responsible for uploads* needs to upload the pdf file. In other words, each group should have only one pdf file uploaded on Gradescope.

Evaluation

The final project will be graded based on three main aspects:

1. adherence to guidelines
2. quality of the report and final results
3. efforts to model development, e.g. a novel algorithm will be recognized

A final report should:

1. clearly state the problem, pointing to hurdles and issues to solve it;
2. clearly present the methodology employed to solve the problem, pointing out:
    i. the data sets used
    ii. the methods employed to (if necessary) handle missing data, transform data, combine data, etc.
    iii. the algorithms involved in the solution, as for example, SVM for classification, DBScan for clustering, etc.
    iv. present and discuss the results, highlighting the strengths and weaknesses of the proposed methodology
    v. make some conclusion, emphasizing whether the chosen approach was success and, if not, why.

Template

Below are guidelines on how to write-up your report for the final project. Not all of the comments may not be relevant to every project. However, please use it as a general guide in structuring your final report. A “standard” experimental machine learning paper consists of the following sections:

1. Introduction

Motivate and abstractly describe the problem you are solving and how you are addressing it. What is the problem? Why is it important? What is your basic approach? A short discussion of how it fits into related work in the area is also desirable (optional for this assignment). Summarize the basic results and conclusions that you will present.

2. Related Work

This section is optional. If in working on your project you came across other papers tackling the same or a similar problem, cite and describe the related work: What is their problem and method? How is your problem and method different? Why might your approach be better? How does your work fit in the bigger picture?

3. Problem Definition and Algorithm

3.1 Task

Precisely define the problem you are addressing (i.e. formally specify the inputs and outputs).

3.2 Algorithm

Describe in reasonable detail the algorithm(s) you are using to address this problem. A pseudocode description of the algorithm(s) you are using is frequently useful. Trace through a concrete example, showing how your algorithm processes this example. The example should be complex enough to illustrate all of the important aspects of the problem but simple enough to be easily understood. If possible, an intuitively meaningful example is better than one with meaningless symbols. Your description of the algorithm should include what assumptions if any you are making about the data, and also what parameters or design choices need to be made (the consequences of these choices should then be explored in detail in the experimental evaluation).

4 Experimental Evaluation

4.1 Data

Describe the data sets that you use in your experimental evaluation. If you do any feature pre-processing, this is the place to describe it.

4.2 Methodology

Describe the experimental methodology that you used. What are the criteria that you are using to evaluate your method? What specific hypotheses does your experiment test? How did you do training/validate/test splits? Comparisons to competing methods that address the same problem are particularly useful.

4.3 Results

Present the quantitative results of your experiments. Graphical data presentation such as graphs and histograms are frequently better than tables. What are the basic differences revealed in the data? Are they statistically significant?

4.4 Discussion

Is your hypothesis supported? What conclusions do the results support about the strengths and weaknesses of your method compared to other methods? How can the results be explained in terms of the underlying properties of the algorithm and/or the data.

5 Conclusions

Briefly summarize the important results and conclusions presented in the paper. What are the most important points illustrated by your work? If you were to continue working on the project, what are the interesting areas for future work? What are the major shortcomings of your current method? For each shortcoming, propose additions or enhancements that would help overcome it.

6 Bilbiography

Be sure to include a standard, well-formated, comprehensive bibliography with citations from the text referring to previously published papers in the scientific literature, resources, or code that you utilized or referenced during your project.