DS-GA 3001.007: Introduction to Machine Learning

Final Project Requirements

Throughout the project, groups will have to work collaboratively to manage expectations and meet goals. As discussed in class, the typesetting platform Overleaf will be useful for sharing reports with teammates.

Groups should contain 1, 2, or 3 members. If you would like to be assigned at random to a group, then please contact the instructors. If you would like to determine your group, then please post to Forums under the *Project* thread to contact classmates.

Timelines

October 31st: Project Proposal
November 28th: Project Milestone
December 15th: Project Report

October 31st: Project Proposal

By October 31 groups must upload a one page pdf file on Gradescope containing:

- Title
- Summary of Plans
  - Description of Problem
  - General Approach
  - Suggested Experiments
- Group
  - Name and NetID of each member.
  - Member responsible for uploading submissions.

Groups could check the resources below for possible datasets.

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.

November 28th: Project Milestone

By November 28th groups must upload on Gradescope a two page pdf file containing:

- Title
- Group Members
  - Name and NetID of each member.
  - Member responsible for uploading submissions.
- Background
  - Description of Problem
  - Motivation for Problem
  - References
- Plans
  - Description of Methodology
  - Proposed Experiments
  - Some Relevant Datasets

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.

December 14th: Project Report

Groups will not be responsible for a presentation. By December 14th groups must upload a notebook (.ipynb file format) and pdf on Gradescope.

The notebook should describe the problem, the methodology and experiments used to understand the problem, evaluation of results, and possible next steps. More specifically, the notebook should be structured as follows:

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. Motivation for Problem
4. References
5. Model
   a. Description of Methodology
   b. Explanation of Algorithm
6. Experiments
   a. Description of Datasets
   b. Explanation of Results

7. Discussion
    a. Evaluation of Findings
    b. Possible Next Steps

See the template below for more information.

The pdf should summarize the notebook. The summary should motivate the problem, explain some aspects of the approach and implementation, and describe the outcomes of the experiments. The pdf should be limited to one page in poster format. Groups can share their posters with the class by electing to upload pdf's to https://wp.nyu.edu/imlf19/

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.

Final Project Evaluation:

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

1. adherence to guidelines
2. quality of the report
3. efforts to model the data

While projects will not be assessed on their technicality, we will recognize work on

1. size and "cleanliness" of the datasets
2. sophistication of the algorithms
3. relevance of the problem to applications

A final report should:

1. clearly state the problem, pointing which are the hurdles and issues to solve it;
2. clearly present the methodology employed to solve the problem, pointing out:
a. the data sets used
b. the methods employed to (if necessary) handle missing data, transform data, combine data, etc.
c. the algorithms involved in the solution, as for example, SVM for classification, DBScan for clustering, etc.
d. present and discuss the results, highlighting the strengths and weaknesses of the proposed methodology
e. 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.

Resources

- Datasets on Amazon's AWS cloud
- Yelp Dataset Challenge
- NYC Open Data
- Data.gov
- UN Data
- Kaggle
- Quandl financial, economic, social datasets
- Face recognition, collaborative filtering, web ranking (see bottom, under "Projects")
    - See here for more collaborative filtering data
- 20 Newsgroups
- Blogs (with spam labels)
- Enron e-mail data set (see also here)
- Congress voting records
- Quota's meta list of datasets
- NYTimes news articles