NYU Tandon - CS6923 Fall 2016
Machine Learning Professor Sandoval

# Final Project

Academic Honesty

Cheating and plagiarism **will not be tolerated**. If you have any questions about a specific case, *please ask me.*

NYU Poly's Policy on Academic Misconduct:
http://engineering.nyu.edu/academics/code-of-conduct/academic-misconduct

Final projects are to be done in groups of 2 or 3 people. This project is meant to be open ended. The goal here is for you to spend time thinking and working deeply with machine learning. To give you an idea of the scope, we are expecting you to spend ~40 hours (per person) between now and the end of the semester on the project.

## Milestones:

1. Project proposal – **Due Date: Nov 8 (10%)**
2. Midway Report - **Due Date Nov 29 ( 25%)**
3. Final Presentation – **Dec 12th** and **Dec 13th**. (**20%**)
4. Final Report – **Dec 12th**. (**45%**)

## Project Proposal:

A 1-page PDF of your project proposal is ***due by Nov 8***. We will send you feedback ~1 week later. Please note that your project proposal will factor into your overall project grade, so make sure that it is written well and follows the guidelines below

Include the following information:

- Project title
- Data set

- Project idea. This should be approximately two paragraphs.
- Software you will need to write.
- Papers to read. Include 1-3 relevant papers. You will probably want to read at least one of them before submitting your proposal
- Teammate (if any) and work division. We expect projects done in a group to be more substantial than projects done individually.
- Midterm milestone: What will you complete by Nov 29? Specifically, what baselines will you try on your datasets and what kind of results do you expect to report?

## Midway Report:

This should be a 5-8 pages report and it serves as a check-point. By this date you should have applied one or more commonly used learning methods on a specific problem. The report should include:

- Background section which describes the problem you are tackling,
- Related work section,
- Description of the method(s) you tried,
- Experiments section with detailed description of the dataset you experimented with and the results you obtained.

## Final Report:

The structure of the final report is similar to that of the midterm report. By that time you should extend the baseline(s) you tried in an interesting and creative way. That can include feature engineering, optimizing the training algorithm, modifications to the model ... etc.
For your final report you need to include a code section. Either a link to a github repo or a zip archive. Here are some examples of how to write up your project

## Final Presentation:

Your final presentation will be in class to your classmates, TAs, Professor and potentially other faculty members and guests. You may do a video for the presentation if you like for some extra credit. You can find some video presentations of students' final projects.

The final presentation will take place on the last day of class before finals Dec 12th. Aim for a **5-7 minute** talk that summarizes the results in an engaging way. The presentations should target a technical but not a Machine Learning class **audience** and serve the purpose of having students practice the highly sought after communication skills that data scientists need. You should think you are presenting to other CS students as your audience. Focus on creating an engaging, clear, and informative presentation that tells the story of your project.

Presentation contents:

- Overview of problem and hypothesis
- Overview of your data
- Modeling techniques used and why (save the details for your paper)
- Conclusions
- Further research and/or business opportunities

## Data sets:

Below are some links to some publicly available data sets. Your project may make use of these or other data sets.

Google Open Images Dataset
CitiBike
NYC Open Data
Kaggle
Datasets on Amazon's AWS cloud
Yelp Dataset Challenge
Data.gov
UN Data
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
Quora's meta list of datasets
NYTimes news article
ICPSR at the University of Michigan. ICPSR stands for Inter-university Consortium for Political and Social Research.
The Bureau of Labor Statistics
Data from the U.S. Census Bureau
Data from papers in the Journal of Applied Econometrics
NBER Macro History Database. See also their list of Business Cycle Dates.
PSID
Statistical Resources on the Web for Economics.