**Machine Learning Final Project Guidelines**

**Objectives and background**

The purpose of the final project is two-fold. First, it will give you the opportunity and incentive to study in-depth a topic that interests you. Second, it will test your ability to use machine learning to solve a problem within this topic. The project definition is purposefully open-ended. The goal is for you to be able to spend time thinking deeply about machine learning and how to best apply it in a real-life scenario.

Final projects are to be done in teams of 2-3. Your final deliveries include a proposal, final presentation, final paper, and final source code. 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.

**Project proposal**

A 1-2 page PDF of your project proposal is *due by Nov 19*.  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 provided below.

You are free to choose any topic that you are excited about, try to optimize for the world and then reduce it to something realistic. You must make sure that it's feasible (please check with us when in doubt). Your project proposal must detail the data that you plan to use, how you will pre-process it, the questions you will ask, the machine learning algorithms you expect to use (permitted to change), and what you expect to get out of your project. Most importantly, you should demonstrate in the proposal that you have gained familiarity with the data set.

You can draw inspiration from these great projects from previous years:

1. [A Concert-planning Tool for Independent Musicians by Machine Learning Models](http://ml.qa.shanghai.nyu.edu/projects/ConcertPlanning2018.pdf)
2. [Predicting Personality Using Textual Data](http://ml.qa.shanghai.nyu.edu/projects/Personality2018.pdf)

**Final presentations**

The final presentation is an opportunity for you to showcase your work. You will probably want to follow a structure similar to:

1. Problem situation: what is the problem? Reflect on why it's important.
2. Descriptive analytics: Which data did you get? Any interesting insights/visualizations?
3. Machine Learning: How did you tackle the problem?
4. Future work: If you had infinite time, what would you envision the next steps to take are?

**Final paper**

Your final paper should read like a mini-paper. It should contain the following sections:

* Abstract: a short description of the whole paper, highlighting important results.
* Introduction: introduction to the problem, potentially relating to prior work done by other people.
* The dataset and features: a concise explanation of the dataset and features, where did you get it from? How did you process the data? Did you combine multiple datasets, and if so, how?
* Explanation of the method used: don't explain the actual algorithm (do mention it), but rather how you used the algorithm to solve the problem. If you used a special loss function or deviated in any other way from the standard algorithm, this is where you mention it.
* Results: walk the reader through your main results, how well did your model perform? Highlight both positive results and shortcomings of your approach.
* Future work/Conclusion: reflect on your findings and think of how you would solve this problem given that you would be able to work on it for 5 more years. Which data would you try getting? Which method would you use? How would it help the problem?

**Less is more**, but you will probably want to reach about 4-5 pages. We recommend you to work with LaTeX to write the document, but you are free to choose whichever editor you like.

**Resources-data**

Below are some links to some publicly available data sets. We will be adding more links in the course of the semester. Your project may make use of these or other data sets. *Projects based on data sets emanating from PRC sources are highly desirable.*

* [Datasets on Amazon's AWS cloud](http://aws.amazon.com/datasets)
* [Yelp Dataset Challenge](http://www.yelp.com/dataset_challenge)
* [NYC Open Data](https://data.cityofnewyork.us/browse)
* [Data.gov](http://catalog.data.gov/dataset)
* [UN Data](http://data.un.org/)
* [Kaggle](https://www.kaggle.com/datasets)
* [Quandl financial, economic, social datasets](http://www.quandl.com/)
* [Face recognition, collaborative filtering, web ranking](http://sli.ics.uci.edu/Classes/2012W-178) (see bottom, under "Projects")
  + See [here](http://www.grouplens.org/node/12) for more collaborative filtering data
* [20 Newsgroups](http://qwone.com/~jason/20Newsgroups/)
* [Blogs](http://ebiquity.umbc.edu/resource/html/id/212/Splog-Blog-Dataset) (with spam labels)
* [Enron e-mail data set](http://www.cs.cmu.edu/~enron/) (see also [here](http://www.cs.cmu.edu/~einat/datasets.html))
* [Congress voting records](http://www.govtrack.us/developers/data)
* [NYTimes news article](http://archive.ics.uci.edu/ml/datasets/Bag+of+Words)
* [ICPSR](http://www.icpsr.umich.edu/) at the University of Michigan. ICPSR stands for Inter-university Consortium for Political and Social Research.
* The [Bureau of Labor Statistics](http://stats.bls.gov/)
* Data from the U.S. [Census Bureau](http://www.census.gov/)
* Data from papers in the [Journal of Applied Econometrics](http://qed.econ.queensu.ca/jae/)
* [NBER Macro History Database](http://www.nber.org/databases/macrohistory/contents/index.html). See also their list of [Business Cycle Dates](http://www.nber.org/cycles.html).
* [PSID](http://www.isr.umich.edu/src/psid/) Panel Study of Income Dynamics
* [US Gov Data Sourses](http://usgovxml.com/)
* [Federal Reserve Economic Data](https://fred.stlouisfed.org/)
* [Knoema](http://knoema.com/atlas):  Economic data about many countries.
* [Data of National Institutes of Health of United States](http://ftp.ncbi.nih.gov/):  Part of data supplied by NIH

Chinese Data Sets:

*The NYI Shanghai Center for Data Science has created a portal for Chinese data sets:*

[*Chinese Datasets Archive*](https://datascience.shanghai.nyu.edu/datasets)

**Resources-cloud computing**

For resource-heavy projects, you may want to use [Google Cloud](https://cloud.google.com/compute/) / [Colab](https://colab.research.google.com/), [Microsoft Azure](https://azure.microsoft.com/en-us/community/education/) or [Amazon AWS](https://aws.amazon.com/machine-learning/). Each of these has a free track that allows you to perform computations up to a certain point. It does take quite some time to set-up so we do not advocate for you using these cloud services. For computationally heavy theoretical projects, we can exceptionally allow students to work on [NYU Shanghai's HPC](https://shanghai.nyu.edu/page/computing-server).

**Grading Rubric**

|  |  |  |
| --- | --- | --- |
| **Component (Weight, total = 100 points)** | | **Grade** |
| **Proposal (20 points)** | |  |
|  | The proposal clearly specifies the problem settings and shows that you have done some preliminary research on this problem already. It motivates why you think it's important and what the role of machine learning will be to solve the problem. The data description includes at least one relevant dataset and potentially other sources where you think you may be able to get data. | A to B |
|  | The proposal is vague and unclear for those who have not been a priori exposed to it. It may be ill-motivated, contain grammar or spelling errors and the proposed data sources are not realistically obtainable or not well-connected to the problem. | B- to F |
| **Implementation (20 points)** | |  |
|  | The code is neatly written and solves the problem correctly. The logical flow of your project notebook makes sense and starts with an exploratory analysis. You correctly attribute other people's source code using references and comments when needed. Higher grades will be awarded to code that is well-documented, concise, vectorized (avoids for loops) and clear. A top grade is reserved for notebooks that are, in essence, flawless on all dimensions. | A to B+ |
|  | The code works but contains some problems which may vary in severity. It is not easily readable by other people because of a lack of good documentation/comments. There may be issues to the flow of the code execution or parts may be missing. You may also not have used vectorization to its fullest effect or left out other potential improvements to optimize your code in terms of speed or accuracy. | B to C |
|  | Grades in this range are reserved for notebooks containing fundamental problems in the analysis (such as confusing causation with correlation, or not using a test-set). The code may not run or is solving a different problem. Acts of plagiarism, cheating are considered the direst of fundamental flaws. | C- to F |
| **Final paper (30 points)** | |  |
|  | The final paper is appropriate for the intended audience and covers all important aspects of the analysis. Information regarding the setting is of outstanding quality (recent, useful, relevant) and based on well-research and competently used external sources which are properly attributed to their original sources in (Author, Date) format when needed. Higher grades will be awarded to documents that are concise, well-written, use LaTeX, are persuasive and clear. A top grade is reserved for documents that are, in essence, flawless on all dimensions. | A to B+ |
|  | Grades in this range will be awarded when the final paper leaves out important ideas, or when it demonstrates a lack of familiarity with the underlying problem or previous solutions to the problem. It may also be logically inconsistent, or problematic in any other way. Especially poorly written documents might earn grades in this range, particularly if the core problem and solution are not well connected or have not been clearly communicated. | B to C |
|  | Grades in this range will be awarded if there are serious issues with the final paper, including but not limited to acts of plagiarism, a failure to take into account concepts taught in class (such as making basic statistical errors), or writing quality that impedes understanding. | C- to F |
| **Final presentation (30 points)** | |  |
|  | The type of presentation is appropriate for the topic and audience, the introduction is attention-grabbing, lays out the problem well, and establishes a framework for the rest of the presentation. The information is presented in a logical sequence. The solution you came up with is logical, relevant and intuitively follows from the rest of the information. The speaker maintains good eye-contact, is appropriately animated and communicates information in a clear, controlled and smooth manner. The length of the presentation is within the assigned time limits. A top grade is reserved for presentations that are, in essence, flawless on all dimensions. | A to B+ |
|  | The presentation contains all the relevant information, but may be hard to follow because it is disorganized, or not communicated clearly by the speaker. The content may have been presented in an uninspired manner and, the speaker may not have been well prepared (or is holding flashcards). Furthermore, there may be issues with any of the following, which impede clarity: visual aids are hard to read or otherwise problematic (label your axes!), the speaker does not come across as knowledgeable or professional. | B to C |
|  | There are serious issues with the presentation which impeded comprehension of the topic: unfamiliarity with the problem setting, not being able to answer basic questions about your solution, acts of plagiarism, ... | C- to F |
|  |  |  |