6.867 Final Project
Fall 2017

Important Milestone Dates:

M0. 9/26: Grade 5%. Decide on *Team* of size 3.

M1. 10/5: Grade 5%. Decide on *Data Set* and *Goals* associated with it.

M2. 10/12: Grade 5%. Decide on *Methods* to use for data processing to achieve defined goals.

M3. 10/24: Grade 5%. Decide on *Implementation / Algorithmic* environment (packages, etc.) that will be utilized.

M4. 11/7: Grade 5%. Produce Initial Results.

M5. 11/21: Grade 5%. Reason about results, interpret and iterate till goals are met.

M6. 12/12: Grade 70%. Project report of length 8 pages not including references.

Process

The final project in 6.867 will be delivered at 7 milestones listed above.

Milestones 1-5: For each of these, you will submit a paragraph worth of material to your assigned TA. Your project TA will approve (full grade) or disapprove (half or 0 grade). Please include your assessment of what the risks to the project are: that is, what things do you think might turn out to be more difficult than planned, and what thoughts do you have about how to mitigate the risks? Are you sure you can obtain the proposed data sets, etc? Are any contingency plans needed? In addition, make the division of labor clear at each milestone.

For milestone 1, you must detail the data that you plan to use, how you will pre-process it, and a precise plan of action, including what questions you would like to ask/problems to solve. We strongly encourage you to download the data and explore it carefully prior to submitting milestone 1.

For milestone 2, you should describe the machine learning algorithm(s) you hope to apply and how you will perform your evaluation. For example, for supervised prediction, you might use cross validation, looking at accuracy; then you might analyze your false positives/negatives to understand where and why the algorithms succeed/fail.

For milestone 3, describe the software infrastucture you will use: what's in a library versus what are you coding from scratch. How is the work split amongst your team? What questions are open?

For milestone 4: Get an initial version of your system running end-to-end and produce initial results.

For milsestone 5: Reason about results, interpret and iterate till goals are met.

If you are going to do an empirical study, be sure that you think about what method to use as a "baseline". It might be running a simple off-the-shelf algorithm or comparing to what happens if you predict the most common class.

Remember that almost anything will turn out to be harder and more time-consuming than you expect. Try to arrange your project so that there are intermediate milestones that can serve as alternative finishing points, in case you don't get to the end. It will be much better to turn in a polished version of a small-scale project than to find yourself at the end of the term with a three-quarters implemented system of great depth and scope.

The Project Report (one per team) will be a written document of length 8 pages (not counting references) in double-column, conference format, including whatever graphs and tables are necessary to make your point. The 8 pages does not count references, and you have unlimited pages for references. Use 10-12 point type and 1.5 inch left and right margins. The report is the means by which you communicate the process and results of your project, so it should be clear, coherent, and well written. Do not dump out large quantities of data or code or uninterpreted charts. Emulate the expositional style of a technical conference paper. You do not need a detailed related work section, but be sure to cite and very quickly explain any technical work you referenced in formulating and carrying out your project.

The main goals are to make clear what your findings are, why you think they came out the way they did, and why that might be important. Be precise enough to allow someone to replicate your experiments (or verify your proofs).

The Project

You will be developing the project over the course of the semester. We expect it to take at least 30 hours total per person. You'll have to plan ahead to avoid getting behind and doing a bad rush job at the end.

Approach Take one or more of the methods that we have talked about in class, or that we are about to cover, and apply them to a problem. Compare their performance and elucidate why they perform differently, if they do. Do they do a good job on the problem?

This is most interesting if you can apply it to some other research question or problem you know about. A big issue here is being sure that you can get the data you need.

You don't necessarily have to implement all (or even any) of the algorithms you use. There are several toolkits available with many learning algorithms already implemented in them. However, if you don't do any implementation yourself, we would expect something much deeper in the way of problem formulation or modeling.

If you decide to implement a numerical algorithm, keep in mind that there may be numerical problems such as you've experienced in the homework: for instance, problems may be ill-conditioned or products of probabilities may go to zero (necessitating the use of logs for intermediate values). We can help you with these sorts of problems during office hours.

There are repositories of data available; links are on the Project Resources page (the link is on the Piazza Course Page, under Resources). **Running existing implementations of algorithms on standard data sets is a bare minimum project which cannot earn more than a C.**

Collaboration

Make completely clear in your paper which software you wrote and which software you used but did not write.

You must do your project in groups of size 3 (if you have a very strong proposal and a good reason, we would consider groups of larger size). You must:

- Make clear before you start what the division of labor will be.
- Make clear in the milestones what the division of labor actually was.
- Be sure that all participants understand all of the work.

Be sure to cite all papers and web sites consulted during the course of your project, as well as to acknowledge other students or TAs who helped you substantially. Citations do not count against your page limit.

Relationship to other classes You may use a single project for 6.867 and another class that you are taking concurrently, or a graduate or undergraduate research project. If it is one project for two classes you must: (1) produce a project that is twice as large in depth and content as would have been required for either class individually; (2) obtain permission from the instructor of the other class; and (3) make clear to us what other class this project is being used for. If it is for your graduate or undergraduate research, make the context clear, and delineate the part of the overall project that is to be considered your project for this class.

Grading

The grading will be broken down as follows:

- Milestones 0-5: (5% each)
- Project report: technical content (60%)
- Project report: presentation, writing, clarity (10%)

Resources

Some Data repositories

- `http://www.ics.uci.edu/~mlearn/MLRepository.html` The UC Irvine repository
- `http://yann.lecun.com/exdb/mnist/` Character recognition
- `http://host.robots.ox.ac.uk/pascal/VOC/` The PASCAL Visual Object Classification Challenge
- `http://www.cs.cmu.edu/~enron/` Enron: A Dataset for Email Classification
- `http://people.csail.mit.edu/dsontag/courses/ml16/assignments/projects.html` Dataset list from last year

- https://archive.org/download/stackexchange Stackexchange data (heres a small project someone built on this data: http://p.migdal.pl/tagoverflow/?site=stackoverflow&size=16)

Machine learning algorithm libraries

- http://www.shogun-toolbox.org/ Shogun large-scale machine learning matlab toolbox
- http://www.mathworks.com/products/statistics/ Matlab statistics toolbox
- http://scikit-learn.sourceforge.net/stable/ scikit.learn: Python machine learning modules
- http://mc-stan.org/ For groups interested in doing a Bayesian modeling project.

Journals and conferences Journal papers are usually easier to read, because they are longer and have better exposition and motivation.

- http://www.jmlr.org Journal of Machine Learning]
- http://www.springerlink.com.libproxy.mit.edu/content/100309/ Machine Learning Journal (URL provided is for MIT access only)
- http://books.nips.cc NIPS Conferences
- http://www.machinelearning.org/icml.html International Conference on Machine Learning