6.891 Machine Learning: Project Proposal

1-Page Proposal Due: **Thursday, November 16**
Project Due: Wednesday, December 13

As a part of the assigned work for this course, we are requiring you to complete a project of your own choosing that is based on the material of this course. The premise of the project must be closely related to some aspect of the material but may explore an avenue that was left unaddressed in class.

Project type and policies

There are various types of projects you can consider:

1. The project may be very practical in terms of applying techniques you have learned in the course to a real problem such as classification of email messages.

2. The project may involve designing or adapting existing algorithms to a novel class of problems. For example, how might we solve multiple related classification tasks? How can we improve document clustering by designing a new clustering metric?

3. The project may consist of a theoretical analysis of a method we have discussed. For example, this may be in terms of complexity, convergence, etc.

4. The project can be a theoretical or more applied survey of a branch of machine learning that we didn't go through in detail. For example, you may write about the use of machine learning in understanding neural systems or sample complexity of machine learning algorithms.

The project can be related to your research area (if you have one).

You can collaborate with other students. If you do, we ask that you outline the role of each person in the project. Projects involving more than one person have to scale in "size" with the number of people.

Project proposal:

In order to help guide your choice of a project, we are requiring you to submit a brief proposal (at most one-page, 12-point font, single spacing, 1 inch margins) that describes the idea for a project, the work you intend to perform, and all the people involved in the project. In particular, it should identify the project type, the problem you plan to address, the motivation for why you find the problem important or interesting, any previous work you already know about, and a rough tentative approach to solving the problem (if applicable).

Project size and the final report:

We expect that the "size" of your project should be equal to about the amount of work required for $1\frac{1}{2}$ homework assignments. The project, however, should be in some sense "complete". By this we mean that you cannot ignore relevant machine learning issues. In the final report you shouldn't just say what you did but also why it was a reasonable thing to do given the course material.

The final report should include about four (4) pages of text per person (not including figures) in the same format as the proposal. You shouldn't worry about getting "great" results. The idea and your understanding of the machine learning issues involved are much more important than getting "great" results.

Some examples:

There are many avenues that you may pursue for this project and we encourage you to be creative even if you don't think you'll necessarily get "great" results. Here are some ideas:

1. *Comparison of algorithms:* Throughout the course, we've been discussing various algorithms and their properties, but only on occasion have we dealt with these algorithms with real sets of data. Often times, algorithms don't work like expected and algorithms may need to be adapted or modified to better fit the assumptions inherent in the data. What work needs to be done to adapt a model to an interesting set of data that you've found? How do various algorithms perform on the same set of data? What are the properties of the various algorithms that exhibit such performance?

2. *Missing information:* Various real world classification problems involve missing components in the input vectors. How can you deal with such missing information? Do you expect your method to degrade rapidly if more information is missing?

3. *Clustering metric:* How do we cluster various types of examples such as sequences? Can you devise a clustering metric or a clustering algorithm that is appropriate in such cases? What if we know that the examples can be transformed in various ways (e.g., translation of images) without changing their "essence". How can we incorporate such prior knowledge into a clustering algorithm?

4. *The iid assumption:* For the problem of classification, we've made a number of assumptions, the greatest and most important of which is that the data is generated i.i.d. from an underlying distribution. Can one still perform classification reasonably well if this is not the case? What if the data for one class is drawn in a reasonable fashion, but not for the other?

5. *The choice of the kernel function in SVMs:* The kernel function in SVMs defines how examples are to be compared. How do we choose the kernel function? How could we adjust the kernel function if we thought it should have a particular form? Can you adapt/design a kernel function to a specific problem we are interested in solving?

Some data repositories you might find useful:

UCI ML Repository (Various) — http://www.ics.uci.edu/~mlearn/MLRepository.html
UCI KDD Repository (Various) — http://kdd.ics.uci.edu/
Protein data bank (Genome) — http://www.rcsb.org/pdb/
Protein structural database (Genome) — http://scop.mrc-lmb.cam.ac.uk/scop/
Cancer classification data (Medical) — http://waldo.wi.mit.edu/MPR/data_set_ALL_AML.html
20 Newsgroups (Text) — http://www.ai.mit.edu/people/jrennie/20_newsgroups/
Reuters Documents (Text) — http://www.research.att.com/~lewis/reuters21578.html
4 Universities (Text) — http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/