6.470 Final Submission

1. What is the name of your website?
ProtoM

2. What is the url for your website?
http://miscoe.csail.mit.edu

3. List the name, school (if not MIT), year, major, and number of years of web programming experience of everyone on your team.
Ying Yin, MIT, graduate student, computer science, 1 year of web programming experience
Jeremy Scott, MIT, graduate student, computer science, 3 months of experience
Andrew Correa, MIT, graduate student, computer science, 4 years experience.

4. Describe the problem your website aims to solve.
Machine learning is both a field of study and a tool, and because most courses focus on theory rather than applications, it can be difficult for students to learn how to use ML as a tool. Furthermore, when applying machine learning to some problem, it is difficult to visualize and understand how your data relates to the quality of your results.

5. Describe how your website solves this problem.
Protom brings practical machine learning into the browser, by allowing a user to upload and visualize data, experiment with features and classifiers, and save the results of their trials. For ML beginners, the website offers tutorials built upon the tool that interactively introduce basic machine learning concepts by way of simple, real-world examples, such as determining positive or negative sentiment of topics on Twitter.

If the reader of this document isn't familiar with machine learning, we believe our Machine Learning Basics tutorial will serve you well. Nevertheless, here is a short explanation of the premise behind machine learning:

A dataset called *weather* is included on the website. A dataset is made up of '*examples'* that consist of '*features'* and *'labels'*. In *weather,* each example's *features* represent weather information on a particular day: the outlook (sunny, overcast or rainy), temperature, humidity and whether or not it was windy. Each example has one of two *labels* or *classes (yes* and *no)* representing whether or not a soccer game could be played on that day. An example: **3,overcast,83,86,FALSE,yes**

It is then the task of the *classifier* to learn the correlation between the features and the labels: this is called *training the* classifier. Once the classifier has been trained, it should able to determine whether or not a soccer game can be played based on the weather. In order to evaluate how well the classifier works, one then *tests* it by having the classifier label new, unseen examples based only on the calculated features.

6. What backend technology did you use?
Ruby on Rails
Weka, for training and testing classifiers

7. What frontend technology did you use?
JavaScript and Protovis

8. List all of your site's features.

- Tutorials:
    1. Introduction to Machine Learning: basic concepts and terminology, gets right to the flow of data in practical ML and in Protom, using the real-world example of determining a person's contact lenses prescription
    2. Intermediate ML: more involved concepts, like how to choose features and interpret results, also using the contact lenses example
    3. Determining Twitter Sentiment: introduces users to basic natural language processing (NLP) and allows them to experiment with classifiers on actual, live tweets on a topic of their choice. Part of a series of application-specific tutorials that we would like to produce for our site; other topics might include speech recognition, hand-drawn digit recognition, object recognition in images
- Machine learning experimenter:
    - Prepare Data:
        - Choose from sample data, or upload your own training and test data
        - Visualize instances ('examples') and features in your data
            - Interactive histogram for viewing examples by class and endless pagination for fast loading of many examples
            - Histograms of class distribution across feature values
        - Select features to be used for training a classifier
    - Classifier:
        - Choose a classifier from a range of options (Bayesian classifiers, SVM, decision trees and other functions)
    - Training and Testing:
        - Train and test the classifier using your prepared data via cross-validation, or testing on unseen data
        - View results: accuracy and confusion matrix
            - Confusion matrix: interactive table of actual vs. predicted classification, where clicking cells displays corresponding examples below the matrix
    - The results of each **trial** (data -> classifier -> training and testing) are saved automatically in a side-pane to allow users to experiment with feature and classifier settings, and then compare results between experiments
    - All interactions implemented via JavaScript and AJAX to avoid reloading the whole page
- User sessions:
    - Newcomers without Protom accounts can...
        - reach the experimenter and the tutorials within two clicks of the website's home page
        - engage in hands-on learning of both machine learning concepts and Protom itself
        - save and review their trial results, as long as they have not left the website or refreshed the page
        - create an account!
    - Users with accounts can, upon login, ...

- access the experimenter and the same tutorials as the newcomer
- create projects and permanently save their trials
- upload and save their own training and testing data

9. Describe which feature(s) are most important and describe why/how that adds to the core value of your website.
The most important feature is the experimenter, and specifically, the flow of interactions and data through it. Users can produce results in roughly 5 clicks from login through an interaction sequence that gets to the core of practical machine learning: 1) preparing data, 2) choosing a classifier, 3) training and testing the classifier, and 4) viewing results. This is not only a simplified way for ML practitioners to produce results and iterate on their experiments, but also lends itself to the education of ML beginners. The tutorials rely upon the experimenter's simple flow for interactively teaching beginners about core concepts through actual examples.

10. Tell us what you think makes your site cool and/or unique.
Protom appears to be the first in-browser machine learning tool. While Weka, our machine learning back-end, is free and widely available, it requires installation and offers either a command line interface, or a messy, cluttered GUI. Protom lends itself better to teaching machine learning beginners who do not need advanced configuration settings. Furthermore, we believe the ability to save data online and quickly run experiments without needing to install software will be invaluable for ML practitioners.

11. List all 3rd party API's, 3rd party plugins, open-source code, etc that you used.
- Weka
- jQuery
- Protovis
- rubygems:
    - rails, mysql2, sass-rails, coffee-rails, uglifier, therubyracer, json, jquery-rails,
    - pwnsytles_rails, bourbon, authpwn_rails, fbgraph_rails, configvars_rails, enhanced_select, remotipart, tipsy-rails, twitter

12. Does your site fit within the MIT Utility category?
Yes. For countless undergraduate and graduate students, machine learning is becoming an important subject and tool. Courses will continue to focus on either the theory behind ML or strategies for specific applications (e.g. features that are useful for computer vision), and the core flow of practical machine learning will not be covered. Protom can serve as an educational supplement for MIT students, and a tool for class projects and research.

13. Does your site fit within the Philanthropy category?
No, other than the fact that our website is freely available and philanthropic possibilities with machine learning are endless.

14. Will at least one of your team members be able to attend judging on Wednesday and awards on Thursday if you are chosen as a semifinalist?
Yes