Proposal for an undergraduate minor in machine learning

Contact: William Cohen (wcohen@cs)

Background.

Several divisions of SCS allow undergraduate minors, including robotics, LTI, software engineering, and CSD. The statistics department also allows undergraduate minors. It has been proposed that MLD offer a similar program. This could be of interest to students that

- plan to do graduate work in machine learning;
- plan to do graduate work in a research area that uses machine learning (e.g., robotics, vision, etc);
- plan to work in industry after they complete their bachelor's degree, at a company that uses machine learning (e.g., Google, Facebook, quantitative trading firms, etc); or,

The proposal.

Briefly:

- The curriculum for 10-601 will be standardized to a greater extent – similarly to 10-701 now.
- One or two additional ML courses will be developed for undergraduates, or broadened to make them more appropriate for undergraduates.
- All majors will take 10-601; an intermediate data analysis course from statistics; and three electives, which could include
    - A year-long senior project (counting as two courses);
    - A pair of courses including one introductory course from an ML-related area, and one advanced course from the same discipline (e.g., 11-441 Natural Language Processing and a second grad-level LTI course, 16-311 Intro to Robotics and a second robotics course, etc)
    - Additional courses from the ML core curriculum, or a list of advanced data analysis courses in statistics.

Below is a less-brief breakdown.

Prerequisites (freshman & sophomore years):

- CS background: 15-122 Data Structures and Algorithms.
- One year calculus: 21-120 and 21-122 or equivalent
- One year of probability & statistics: 36-225 or 21-325, followed by 36-226. (This is required for the necessary statistics background: 36-226 is a prereq for 36-401, which is in turn a prereq for 36-46x, the advanced data analysis courses in stats.)

- Optionally, a semester of matrix algebra: 21-240, 21-241 or 21-242 or 21-341 would be recommended, but would be required only for students taking courses in MLD that go beyond 10-601—i.e. the "foundations" students.

Core material (junior year):

- 10-601 Machine Learning. Since this is now core for the minor, it will be more important to have it cover a defined core of technical material, such as the *proposed* list below (this is the core material for 701 with some deletions and **additions**):

    - Review of basic probability, MLE, MAP estimation
    - **Review of matrix algebra**
    - Decision tree learning
    - Naïve Bayes classifiers, Bayes rule, Conditional independence
    - ~~Logistic regression, regularization, and its equivalence to Naïve Bayes under certain assumptions~~
    - Discriminative versus generative classification
    - Linear regression and equivalence of minimizing sum squared error and maximizing likelihood under appropriate assumptions
    - Nearest-neighbor classifier
    - ~~Boosting~~
    - Gradient descent and artificial neural networks
    - Overfitting, cross validation, bias-variance tradeoff
    - Learning Theory: PAC learning bounds, finite hypothesis spaces~~, VC dimension~~
    - Margin-based methods, Support Vector Machines ~~and use of kernels~~
    - Bayes nets (DAGS) representation, inference, learning
    - Expectation Maximization
    - K-means, Mixture models and clustering
    - Hidden Markov Models
    - ~~Markov Decision Processes and Reinforcement learning~~
- 36-401 Modern Regression. This covers exploratory data analysis, linear regression and multiple regression, and more generally, how to work with, and thinking critically about, real datasets.

    Note: 360-401 is often over-subscribed. ML minors should be aware of this, and plan appropriately – eg, register immediately when slots are available.

Advanced material (junior & senior year):

Minors must also take a total of three courses from the options below.

- A year-long senior project, supervised or co-supervised by an MLD member
- Electives, chosen from either the list of core MLD courses, or a list of "advanced data analysis" courses from statistics (see list below).
- A combination of two courses (see list below) including:
    - A course that provides an introduction to a field that uses machine learning methods.
    - A second course in the same discipline with a significant machine-learning component.

Pre-approved "data analysis" courses are: 6-402 Advanced Data Analysis; 36-315 Statistical Graphics and Visualization; or one of the special topics courses – which currently include: 36-461 Statistical Methods for Epidemiology, 36-462 Chaos, Complexity and Inference, 36-463 Hierarchical Models, or 36-464 Multivariate Methods. (Additional courses, especially in the 36-46x series, may be allowed, but need to be approved by MLD's Head of Undergraduate Studies.)

Preapproved two-course pairs are listed below. (Again, additional pairs may be allowed by need to be approved by MLD's Head of Undergraduate Studies.)

- In Language Technologies:
    - 11-411 Natural Language Processing , plus one of
    - 11-755 ML for Signal Processing; 11-761 Language and Statistics; 11-762 Language and Statistics II; 11-731 Machine Translation; 11-741 Information Retrieval ; 11-748 Information Extraction ; 11-751 Speech Recognition; 11-763 Structured Prediction for Language and Other Discrete Data; 11-773 Text-Driven Forecasting
- In Robotics:
    - 16-311 Introduction to Robotics, plus one of
    - 16-831 Statistical Techniques in Robotics; 16-745 Dynamic Optimization; 16-899C Adaptive Control and Reinforcement Learning
- In Computer Vision:
    - 15-385 Computer Vision, plus one of
    - 16-725 Methods in Medical Image Analysis; 16-824 Physics-based Methods in Vision; 16-823 Physics based Methods in Computer Vision; 15-862 Computational Photography
- In Neural Cognition:
    - 15-883 Computational Models of Neural Systems or 85-419/719 Introduction to Parallel Distributed Processing; plus one of
    - 03-761 Neural plasticity; 18-699/42-590 Neural Signal Processing; 85-765 Cognitive neuroscience
- In Artificial Intelligence:
    - 15381 Artificial Intelligence Representation and Problem Solving, plus one of
    - 15-887 Planning, Execution, and Learning; 15-892 Foundations of Electronic Marketplaces; 15-589 Independent Study in Artificial Intelligence
- In Bioinformatics:
    - 02-510 – Computational Genomics , plus 03-511 – Computational Molecular Biology and Genomics; or
    - 02-530 – Cell and Systems Modeling, plus 03-512 – Computational Methods for Bio modeling and Simulation

Implementing the proposal.

The proposal doesn't absolutely require any new courses to be developed to be functional. However, minors will have to take at least one elective in MLD or stats in this model, and over time it may be desirable either (a) to develop new courses aimed more at undergrads or (b) to "broaden" existing courses to better support undergrads. As an example of (b), note that some other departments will have courses where undergrads and grads take the same lectures, but grads do additional work—e.g., additional assignments—and/or have additional meetings—e.g., "lab" meetings.

As an example, an assignment-oriented course like 601 could be "broadened" by making some assignments required for grads, but extra credit for undergrads; by allowing grad students to drop one low grade, and undergrads to drop three; or similar methods. (On the other hand, a course that is taught primarily from research papers and graded based on a single semester project might be hard to adapt in this way.)

Some candidate courses that be useful for undergraduates:

- Graphical models, possibly based on broadening 15-708.
- Machine learning theory, possibly based on broadening 15-859.
- A practicum on machine learning for large datasets, sort of like Kamal's Nigam's course "15-505 Internet Search Technologies" http://www.cs.cmu.edu/~knigam/15-505/schedule.html . William will prototype a course like this for spring 2012.
- A course on the structure of information networks, sort of like Jon Kleinberg's "Networks" course at Cornell, http://www.infosci.cornell.edu/courses/info2040/2010sp/ (which covers aspects of graph theory, game theory, economics, and dynamics of networks.).

Issues to consider going forward.

- Who is the MLD Head of Undergrad studies? (William Cohen is willing, but I guess it should be voted on by MLD. Alternately this could be another hat worn by the current curriculum committee.)
- Is this list of pairs of sister-discipline courses to allow as electives ok? Should others be added, or should some of these be dropped?
- Will there be too heavy a load on 36-401? How large is this program likely to be? Going forward we may need to consider this.