CSCI 1950-F: Introduction to Machine Learning

Brown University, Spring 2012

How can artificial systems learn from examples, and discover information buried in massive datasets? This course explores the theory and practice of statistical machine learning. Topics include parameter estimation, probabilistic graphical models, approximate inference, and kernel and nonparametric methods. Applications to regression, categorization, and clustering problems are illustrated by examples from vision, language, communications, and bioinformatics. *Prerequisites:* CSCI0160, CSCI0180 or CSCI0190, and comfort with basic probability, linear algebra, and calculus.

Introduction

The main goal of this class is to introduce you to the ideas and techniques of machine learning, and the probabilistic models that underlie behind them. These ideas have their origins in classical results from statisticians such as Laplace, Bayes, and Fisher. However, modern computing techniques now permit applications of a scale and diversity that was barely conceivable only a few decades ago.

As opposed to the traditional statistical focus on analysis of experiments, most problems we'll discuss involve some form of prediction. *Classification* algorithms predict a discrete value from a finite set of choices, while *regression* algorithms predict a continuous value. *Supervised learning* techniques can be used to design such predictors using training data that is labeled with the values you are trying to learn. *Unsupervised learning* techniques are instead used when such labels are unavailable, but you nevertheless hope to discover interesting structure within your data. These methods lead to effective algorithms for *clustering* and *dimensionality reduction*. This course will explore the conceptual relationships between these different learning problems, and introduce some of the most practically effective statistical models and computational methods.

Administrative Information

Lectures: Tuesdays and Thursdays from 1:00-2:20pm, CIT room 227, 115 Waterman St.

Recitations: Thursdays 5:00-6:00pm, location to be determined.

Instructor:

Erik Sudderth (sudderth@cs.brown.edu; 401-863-7660)
Office Hours: Mondays 11:00am-12:00pm, Tuesdays 2:30-3:30pm, CIT room 509.

Graduate Teaching Assistants:

Dae Il Kim (daeil@cs.brown.edu)
Office Hours: Wednesdays 10:00am-12:00pm, CIT room 409.

Ben Swanson (chonger@cs.brown.edu)
Office Hours: Tuesdays 12:00-1:00pm & 3:00-4:00pm, CIT room 411.

Undergraduate Teaching Assistants:

William Allen (william_allen@brown.edu)
Office Hours: Mondays 6:00-8:00pm, CIT room 227.

Soravit Changpinyo, Zachary Kahn, Paul Kernfeld, & Vazheh Moussavi
Office Hours: Sundays through Wednesdays, times to be announced, CIT room 227.

Grading, Assignments, and Readings

There will be at most ten homework assignments, each due at least one week after it is assigned. Homework problems will involve both mathematical derivations and Matlab implementation of learning algorithms. Detailed electronic formatting and submission instructions, and a formal collaboration policy, will be announced with the first assignment on February 2.

Each homework assignment will be worth 100 points. For every day or fraction of a day that an assignment is turned in late, 25 points will be subtracted from the earned score. Exceptions to this policy are only given in unusual circumstances, and any extensions must be requested by e-mail to the instructor well before the deadline.

In addition to homeworks, there will be one in-class midterm in mid-March, and a final exam. Overall grades will be assigned as follows: 50% homeworks, 25% final exam, 20% midterm exam, 5% class participation. Computer science graduate students may receive 2000-level degree credit by completing an additional project involving an application of the course material. To successfully complete the project, students must submit a brief proposal in late March, give a short oral presentation on their results during reading period, and prepare a written report (4 to 8 pages) by the end of the semester. Specific due dates will be announced later.

Course readings will be taken from a draft textbook by Kevin P. Murphy, entitled *Machine Learning: A Probabilistic Perspective*. We will provide supplemental readings from this book, and occasionally other sources, for each lecture. Hardcopies are available at the Metcalf copy center.

Tentative Syllabus

- Binary and multinomial categorization, ROC Curves, naive Bayes, K nearest neighbors
- Frequentist and Bayesian estimation, decision theory, loss functions
- Maximum likelihood estimation, asymptotics
- Bayes' rule, Bayesian MAP and MMSE estimation
- Cross-validation, model selection, regularization, sparsity
- Generalized linear models, exponential families, logistic regression
- Optimization, stochastic gradient descent
- Kernel methods, Gaussian processes, perceptron algorithm, support vector machines
- Directed graphical models (Bayesian networks)
- Multivariate Gaussian distributions, linear regression, logistic regression
- Clustering, K-means/medoids, expectation maximization (EM) algorithm
- Hidden Markov models (HMMs), Viterbi algorithm, EM parameter estimation
- Dimensionality reduction, PCA, factor analysis, manifold learning
- Monte Carlo estimators, importance sampling, Markov chain Monte Carlo (MCMC)