10-702: Statistical Machine Learning

Syllabus, Spring 2010

http://www.cs.cmu.edu/~10702

Statistical Machine Learning is a second graduate level course in machine learning, assuming students have taken Machine Learning (10-701) and Intermediate Statistics (36-705). The term "statistical" in the title reflects the emphasis on statistical analysis and methodology, which is the predominant approach in modern machine learning.

The course combines methodology with theoretical foundations and computational aspects. It treats both the "art" of designing good learning algorithms and the "science" of analyzing an algorithm's statistical properties and performance guarantees. Theorems are presented together with practical aspects of methodology and intuition to help students develop tools for selecting appropriate methods and approaches to problems in their own research.

The course includes topics in statistical theory that are now becoming important for researchers in machine learning, including consistency, minimax estimation, and concentration of measure. It also presents topics in computation including elements of convex optimization, variational methods, randomized projection algorithms, and techniques for handling large data sets.

Schedule

Lectures	Tues. and Thurs. 1:30-2:50pm	GHC 4215
Office hours		
Xi Chen	Thurs. 3:00-4:00pm	GHC 8th floor
Mladen Kolar	Thurs. 4:30-5:30pm	GHC 8th floor

Contact Information

Instructors:

John Lafferty	GHC 8205, 268-6791	lafferty@cs.cmu.edu
Larry Wasserman	BH 228A, 268-8727	larry@stat.cmu.edu

Teaching Assistants:

Xi Chen	GHC 8219, 268-3536	xichen@cs.cmu.edu
Mladen Kolar	GHC 8223, 268-3653	mladenk@cs.cmu.edu

Course Secretary:

Sharon Cavlovich	GHC 8215, 268-5196	sharonw@cs.cmu.edu

Prerequisites

You should have taken 10-701 and 36-705. We will assume that you are familiar with the following concepts:

1. convergence in probability
2. central limit theorem
3. maximum likelihood
4. delta method
5. Fisher information
6. Bayesian inference
7. posterior distribution
8. bias, variance and mean squared error
9. determinants, eigenvalues, eigenvectors

It is essential that you know these topics.

Text and Reference Materials

There is no required text for the course; however, lecture notes will be regularly distributed. These are draft chapters and sections from a book in progress (also called "Statistical Machine Learning"). *Comments, corrections, and other input on the drafts are highly encouraged.*

The book is intended to be at a more advanced level than current texts such as *The Elements of Statistical Learning* by Hastie, Tibshirani and Freedman or *Pattern Recognition and Machine Learning* by Bishop. But these books are excellent references that may complement many parts of the course. Recommended texts include:

- Chris Bishop, *Pattern Recognition and Machine Learning*, Springer, Information Science and Statistics Series, 2006.
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Texts in Statistics, Springer-Verlag, New York, 2001.

- Larry Wasserman, *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics, Springer-Verlag, New York, 2004.
- Larry Wasserman, *All of Nonparametric Statistics*, Springer Texts in Statistics, Springer-Verlag, New York, 2005.

Assignments, Exams, and Grades

The course will have

- **Six (6) assignments**, which will include both problem solving and experimental components. The assignments will be given roughly every two weeks. They will be due on **Fridays at 5:00 p.m.**, in Sharon Cavlovich's office, GHC 8215.
- **Midterm exam**. There will be a midterm exam on Thursday, March 4.
- **Project**. There will be a final project, described later in this syllabus.

Grading for the class will be as follows:

50% Assignments
25% Midterm exam
25% Project

Programming Language

All computational problems for the course are to be completed using the R programming language. R is an excellent language for statistical computing, which has many advantages over Matlab and other scientific scripting languages. The underlying programming language is elegant and powerful. Students have found it useful, and not difficult, to learn this language even if they primarily use another language in their own research. Free downloads of the language, together with an extensive set of resources, can be found at `http://www.r-project.org`. For a recent news article on R, see `http://www.nytimes.com/2009/01/07/technology/-business-computing/07program.html`

Policy on Collaboration

Collaboration on homework assignments with fellow students is encouraged. However, such collaboration should be clearly acknowledged, by listing the names of the students with

whom you have had discussions concerning your solution. You may not, however, share written work or code—after discussing a problem with others, the solution should be written by yourself.

Topics

The course will follow the outline of the book manuscript, and will include topics from the following:

1. **Statistical Theory**: Maximum likelihood, Bayes, minimax, parametric versus nonparametric methods, Bayesian versus Non-Bayesian approaches, classification, regression, density estimation.

2. **Convexity and Optimization:** Convexity, conjugate functions, unconstrained and constrained optimization, KKT conditions.

3. **Parametric Methods**: Linear regression, model selection, generalized linear models, mixture models, classification, graphical models, structured prediction, hidden Markov models

4. **Sparsity**: High dimensional data and the role of sparsity, basis pursuit and the lasso revisited, sparsistency, consistency, persistency, greedy algorithms for sparse linear regression, sparsity in nonparametric regression. sparsity in graphical models, compressed sensing

5. **Nonparametric Methods**: Nonparametric regression and density estimation, nonparametric classification, clustering and dimension reduction, manifold methods, spectral methods, the bootstrap and subsampling, nonparametric Bayes.

6. **Advanced Theory**: Concentration of measure, covering numbers, learning theory, risk minimization, Tsybakov noise conditions, minimax rates for classification and regression, surrogate loss functions.

7. **Kernel Methods**: Mercel kernels, kernel classification, kernel PCA, kernel tests of independence.

8. **Computation**: The EM Algorithm, simulation, variational methods, regularization path algorithms, graph algorithms

9. **Other Learning Methods**: Semi-supervised learning, reinforcement learning, minimum description length, online learning, the PAC model, active learning

Final Project

The project is similar to the project in 10-701. Here are the rules:

1. You may work by yourself or in teams of two.
2. Choose an interesting dataset **that you have not analyzed before.** A good source of data is: `http://www.ics.uci.edu/~mlearn/MLRepository.html`
3. The goals are (i) to use the methods you have learned in class or, if you wish, to develop a new method and (ii) present a theoretical analysis of the methods.
4. You will provide: (i) a proposal, (ii) a progress report and (iii) and final report.
5. The reports should be well-written. This is a good time to buy a copy of *The Elements of Style* by Strunk and White.

Proposal. A one page proposal is due **Tuesday, February 16**. It should contain the following information: (1) project title, (2) team members, (3) description of the data, (4) precise description of the question you are trying to answer with the data, (5) preliminary plan for analysis, (6) reading list. (Papers you will need to read).

Progress Report. Due **Friday, April 9**. Three pages. Include: (i) a high quality introduction, (ii) what have you done so far and (iii) what remains to be done.

Project Ad. Due **Tuesday, April 27**. One pdf slide. An "advertisement" describing your project to the class. Include (i) brief description of your problem and results (ii) graphic (optional).

Final Report: **Due Tuesday, May 4.** The paper should be in NIPS format. However, it can be up to 20 pages long. You should submit a pdf file electronically. It should have the following format:

1. Introduction. A quick summary of the problem, methods and results.
2. Problem description. Detailed description of the problem. What question are you trying to address?
3. Methods. Description of methods used.
4. Results. The results of applying the methods to the data set.

5. Theory. This section should contain a cogent discussion of the theoretical properties of the method. It should also discuss under what assumptions the methods should work and under what conditions they will fail.

6. Simulation studies. Results of applying the method to simulated data sets.

7. Conclusions. What is the answer to the question? What did you learn about the methods?

Course Calendar

The course calendar is posted on the course website, `http://www.cs.cmu.edu/~10702`, and will be updated throughout the semester.