MLD 10-702 Statistical Machine Learning

Spring 2008 Syllabus

http://www.stat.cmu.edu/~larry/=sml2008

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. It is intended for students who want to practice the "art" of designing good learning algorithms, and also understand 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.

Schedule

Lectures	Mon. and Wed. 1:30-2:50	Wean Hall 4623
OFFICE HOURS	Tuesdays 4:00-5:00	Baker Hall 228a
TA OFFICE HOURS	ТВА	TBA

Contact Information

Professor: Larry Wasserman	Baker Hall 228A, 268-8727	larry@stat.cmu.edu
Teaching Assistan Jingrui He	t: TBA	jingruih@cs.cmu.edu
Secretary: Diane Stidle	Wean Hall 4609, 268-3431	diane@cs.cmu.edu

Prerequisites

You should have taken 10-701 and 36-705. I 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
- posterior distribution
 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 3:00 p.m.
- Midterm exam. There will be a midterm exam on Monday, March 3.
- **Project**. There will be a final project, similar to the project in 10-701. The project is described later in this syllabus and on the website.

Grading for the class will be as follows:

50% Assignments25% Midterm exam25% 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.

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, 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 2.
- 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. The proposal is due **February 15**. The length is 1 page. It should contain the following information: Project title, Team members, Description of the data, Precise description of the question you are trying to answer with the data, Preliminary plan for analysis, Reading list. (Papers you will need to read).

Progress Report. Due **April 4**. Three pages. Include: (i) a high quality introduction, (ii) what have you done so far and (iii) what remains to be done.

Final Report: **Due May 2.** 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

Week of		Mon	Wed	Friday
January	14			
	21			Homework 1
	28			
February	4			Homework 2
	11			Project Proposal
	18			Homework 3
	25			
March	3	Midterm Exam	No Class	Spring Break
	10	Spring Break	Spring Break	Spring Break
	17			Homework 4
	24			
	31			Progress report
April	7			Homework 5
	14			
	21			
	28		Last Class	Submit Project
May	5			Homework 6