10-702/36-702: Statistical Machine Learning

Syllabus, Spring 2011

http://www.cs.cmu.edu/∼aarti/Class/10702

Statistical Machine Learning is a second graduate level course in **advanced 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	Mondays and Wednesdays 10:30 - 11:50 am	NSH 1305
Recitations	Thursdays 5:00 - 6:00 pm	NSH 1305
Office Hours		
T.K. Huang	Tuesday 5:00 - 6:00 pm	GHC 8014 Common Area
Min Xu	Tuesday 5:00 - 6:00 pm	GHC 8014 Common Area
Larry Wasserman	Mondays 12:00 - 1:00 pm	Baker Hall 228a
Aarti Singh	Wednesdays 1:30 - 2:30 pm	GHC 8207

Contact Information

Instructors:

Aarti Singh	GHC 8207	268-4266	aarti@cs.cmu.edu
Larry Wasserman	BH 228a	268-8727	larry@cmu.edu

Teaching Assistants:

T.K. Huang	GHC 8015	268-2960	tzukuoh@cs.cmu.edu
Min Xu	GHC 8013	268-2687	minx@cs.cmu.edu

Course Secretary:

Michelle Martin	GHC 8001	268-5527	michelle324@cs.cmu.edu

Prerequisites

You should have taken 10-701 and 36-705. If you did not take these courses, **it is your responsibility to do background reading to make sure you understand the concepts in those courses.** We will assume that you are familiar with the following concepts:

1. Convergence in probability and convergence in distribution.
2. The central limit theorem and the law of large numbers.
3. Maximum likelihood, Fisher information.
4. Bayesian inference.
5. Bias and variance.
6. Bayes classifiers; linear classifiers; support vector machines.
7. Determinants, eigenvalues and eigenvectors.

Text

The text is *Statistical Machine Learning* by Lafferty, Liu and Wasserman. However, the book is in preparation. We will hand out chapters as needed. Other useful reference are:

1. Trevor Hastie, Robert Tibshirani, Jerome Friedman (2001). *The Elements of Statistical Learning*, Available at http://www-stat.stanford.edu/∼tibs/ElemStatLearn/.
2. Chris Bishop (2006). *Pattern Recognition and Machine Learning.*
3. Luc Devroye, László Györfi, Gábor Lugosi. (1996). *A probabilistic theory of pattern recognition.*
4. Larry Wasserman (2004). *All of Statistics: A Concise Course in Statistical Inference.*
5. Larry Wasserman (2005). *All of Nonparametric Statistics.*

Grading

There will be:

1. **Six assignments**. They are due **Fridays at 3:00 p.m.**. Hand them in to Michelle Martin (GHC 8001). If she is not in office, please write your name, date and time of submission on the homework and slide it under her door.
2. **Midterm Exam**. The date is **Wednesday March 2**.
3. **Project**. There will be a final project, described later in the syllabus.

Grading will be as follows:

50% Assignments
25% Midterm
25% Project

Programming

You may use any programming language you want. However, we cannot provide help unless you use R or Matlab. If you don't know R you should strongly consider learning it. 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 easy to learn. **It is 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 solutions should be written by you.

Topics

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

1. Concentration of measure.
2. Minimax theory.
3. Convexity and optimization.
4. Parametric Methods.
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. 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 .
7. Advanced Theory: risk minimization, Tsybakov noise conditions, surrogate loss functions, random matrix theory.
8. Kernel Methods: Mercel kernels, kernel classification, kernel PCA, kernel tests of independence.
9. Computation: The EM Algorithm, simulation, variational methods, regularization path algorithms, graph algorithms.
10. Other Learning Methods: Semi-supervised learning, reinforcement learning, minimum description length, online learning, the PAC model, active learning.

Course Calendar

The course calendar is posted on the course website
http://www.cs.cmu.edu/∼aarti/Class/10702
and will be updated throughout the semester.

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.

Proposal. A one page proposal is due **Tuesday, February 15**. 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 8**. Three pages. Include: (i) a high quality intro- duction, (ii) what have you done so far, (iii) what remains to be done and (iv) a clear description of the division of work among teammates, if applicable.

Project Ad. Due **Wednesday, April 20**. One pdf slide. An "advertisement" describing your project to the class. Include (i) brief description of your problem and results (ii) graphic (optional).

Project Spotlight. On **Wed April 27**, you will get 2 minutes to present your project to the class.

Final Report: Due **Tuesday, May 3**. The paper should be in NIPS format. **Maximum 8 pages**. (You can have an appendix with extra material if needed. If working in groups of two, please include a clear description of the contribution of each person in the appendix.) 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. You do not need to develop new theory.
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? Mention any future directions of interest.

Note: You can also choose to do a purely theoretical project. In this case, you should choose an area of interest, read several key papers, and provide a **clear, unified summary** of the theoretical results in these papers.