10-702/36-702: Statistical Machine Learning

Syllabus, Spring 2013

http://www.cs.cmu.edu/∼aarti/Class/10702_Spring13/

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, randomized projection algorithms, and techniques for handling large data sets.

Schedule

Lectures	Mon and Wed 1:30 - 2:50 pm	BH A51
Office Hours		
Akshay Krishnamurthy, Yifei Ma	Mondays 3:00-4:00 pm	GHC 8th floor Commons
Larry Wasserman	Wednesdays 3:00-4:00 pm	Baker Hall 228a
Aarti Singh	Tuesdays 1:30-2:30 pm	GHC 8207
Recitation		
Akshay Krishnamurthy, Yifei Ma	Thursdays 5:00-6:00 pm	Doherty Hall A302 (Jan 17) Porter Hall 125C (all other weeks)

Contact Information

Instructor:

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

Teaching Assistants:

Akshay Krishnamurthy	TBA	akshaykr@cs.cmu.edu
Yifei Ma	TBA	yifeim@cs.cmu.edu

Course Secretary:

Michelle Martin	GHC 8001	268-5527	mhaywood@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. Book chapters will be availbe at:

www.cmu.edu/blackboard

Please do not distribute these chapters. 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, write your name, the course number, date and time of submission on the homework and slide it under her door. Note: It is very important that you write the course number as she handles multiple classes.
2. **Midterm Exam**. The date is **Wednesday March 6**.
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. 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_Spring13/`
and will be updated throughout the semester.

Project

The project involves picking a topic of interest, reading the relevant results in the area and then writing a short paper (8 pages) summarizing the key theoretical results in the area. The paper should include background, statement of important results, and brief proof outlines for the results. Example of topics include, but are not limited to, low rank matrix completion, high-dimensional learning of graphical model structure, compressed sensing, entropy estimation, sparse PCA, sparse subspace clustering, etc. You are encouraged to discuss your topic with the instructors or TAs.

1. You may work by yourself or in teams of two.
2. The goals are (i) to summarize key results in literature on a particular topic **and** (ii) present a summary of the theoretical analysis (results and proof sketch) of the methods. You may develop new theory if you like but it is not required.
3. You will provide: (i) a proposal, (ii) a progress report and (iii) and final report.
4. The reports should be well-written.

Proposal. A one page proposal is due **February 15**. It should contain the following information: (1) project title, (2) team members, (3) precise description of the problem you are studying, (4) anticipated scope of the project, and (5) reading list. (Papers you will need to read).

Progress Report. Due **March 29**. Three pages. Include: (i) a high quality introduction, (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 Spotlight. On **April 29** and **May 1**, you will get 2 minutes to present your project to the class.

Final Report: Due **Monday, May 6**. The paper should be in NIPS format. **Maximum 8 pages**. No appendix is allowed. (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. Motivation and a quick summary of the area.
2. Notation and Assumptions.
3. Key Results.
4. Proof outlines for the results.
5. Conclusion. This includes comments on the meaning of the results and open questions.