10-702/36-702
Statistical Machine Learning

Syllabus, Spring 2016

http://www.stat.cmu.edu/∼larry/=sml
Lectures: Tuesday and Thursday 1:30 - 2:50 pm (HH B103)

Statistical Machine Learning is a second graduate level course in **advanced machine learning**, assuming students have taken Machine Learning (10-715) and Intermediate Statistics (36-705). The term "statistical" in the title reflects the emphasis on statistical theory and methodology. The course combines methodology with theoretical foundations. 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 important for researchers in machine learning, including nonparametric theory, consistency, minimax estimation, and concentration of measure.

Contact Information

Instructor:
Larry Wasserman BH 132F 412-268-8727 larry@cmu.edu

Teaching Assistants:
Jisu Kim jisuk1@andrew.cmu.edu
Bryan Hooi bhooi@andrew.cmu.edu
Adams Yu weiyu@andrew.cmu.edu.
Their office hours are on the website.

Course Assistant:
Mallory Deptola GHC 8001 412-268-5527 mdeptola@cs.cmu.edu

Office Hours
Larry Wasserman Tuesdays 3:00-4:00 pm Baker Hall 132F

Prerequisites

You should have taken 10-701 and 36-715. 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. Regression.
6. The Bias-Variance tradeoff.
7. Bayes classifiers; linear classifiers; support vector machines.
8. Determinants, eigenvalues and eigenvectors.

Text

There is no text but course notes will be posted. 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. **Four or five assignments**. They are due **Fridays at 3:00 p.m.**. Hand them in to Mallory Deptola (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.
2. **Midterm Exam**. The date is **TUESDAY MARCH 1**.
3. **Project and Course Conference**. There will be a final project, described later in the syllabus.

Grading will be as follows:

50% Assignments
25% Midterm
25% Project

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

1. Review: probability, bias/variance, mle, regression, classification.
2. Theoretical Foundations
   (a) Function Spaces: Holder spaces, Sobolev spaces, reproducing kernel Hilbert spaces (RKHS)
   (b) Concentration of Measure
   (c) Minimax Theory
3. Supervised Learning
   (a) Linear Regression: low dimensional, ridge regression, lasso, greedy regression
   (b) Nonpar Regression: kernel regression, local polynomials, additive, RKHS regression
   (c) Linear Classification: linear, logistic, SVM, sparse logistic
   (d) Nonpar Classification: NN, naive Bayes, plug-in, kernelized SVM
   (e) Conformal Prediction
   (f) Cross Validation
4. Unsupervised Learning
   (a) Nonpar Density Estimation
   (b) Clustering: k-means, mixtures, single-linkage, density clustering, spectral clustering
   (c) Measures of Dependence
   (d) Graphical Models: correlation graphs, partial correlation graphs, cond. indep. graphs
5. Other Topics
   (a) Nonparametric Bayesian Inference
   (b) Bootstrap and subsampling
   (c) Interactive Data Analysis
   (d) Robustness
   (e) Active Learning
   (f) Differential Privacy
   (g) Deep Learning
   (h) Distributed Learning
   (i) Streaming

Course Calendar

The course calendar is posted on the course website 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 emphasis should be on theory.** Your are NOT required to do new research.

The paper should include background, statement of important results, and brief proof outlines for the results. 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 12**. 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 18**. 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.

First Draft of Final Report is due **April 18**. Tis will be used for refereeing.

Course Conference. On **April 26** and **April 28**, there will be a course conference. You will get to vote on your favorite projects and the winners will present their projects.

Final Report: Due **Monday, May 2**. The paper should be in NIPS format. (pdf only). **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.