# AMS Short Course

## *Aspects of Statistical Learning*

### New Orleans, Louisiana, January 3–4, 2007

### Organized by:
### Cynthia Rudin, New York University
### Miroslav Dudík, Princeton University

Statistical learning is a rapidly growing area of computer science. Its goal is to design computer algorithms that generalize or "learn" from examples. Applications of statistical learning have been extremely wide-ranging and include tasks in pattern and speech recognition, text processing, portfolio management, medical diagnosis, robotics, and bioinformatics.

The goal of this course is to introduce the field of statistical learning to a wider mathematical audience. Possible areas of interest to mathematicians include the design of algorithms and analysis of their convergence, probabilistic guarantees of generalization performance, and beating the "curse of dimensionality". The talks will span a variety of problem domains, techniques, and algorithm types, and touch upon many interesting applications. Each talk will be tutorial-style and accessible to a broad mathematical audience.

It is planned that lecture notes will be available to those who register for this course. Advance registration fees are: member of the AMS—US$90; nonmember—US$120; student, unemployed, emeritus—US$40. On-site fees are: member of the AMS—US$120; nonmember—US$151; student, unemployed, emeritus—US$60. Registration and housing information can be found in this issue of the *Notices*; see the section "Registering in Advance and Hotel Accommodations" in the announcement for meetings in New Orleans. The registration form is at the back of this issue.

A tentative list of speakers:

Robert E. Schapire (Princeton University)

Lisa Hellerstein (Polytechnic University, Brooklyn)

Adam Tauman Kalai (Weizmann Institute of Science and Toyota Technological Institute)

Lawrence Saul (University of California San Diego)

## Machine Learning Algorithms for Classification

*Robert E. Schapire, Princeton University*

Machine learning studies the design of computer algorithms that automatically make predictions about the unknown based on past observations. Often, the goal is to learn to categorize objects into one of a relatively small set of classes. This tutorial will introduce some of the main state-of-the-art machine learning techniques for solving such classification problems, possibly including decision trees, boosting and support-vector machines. The tutorial will also discuss some of the key issues in classifier design, including avoidance of overfitting.

### Reading List

[1] VLADIMIR N. VAPNIK, *Statistical Learning Theory*, Wiley, (1998).

[2] LEO BREIMAN, JEROME H. FRIEDMAN, RICHARD A. OLSHEN, and CHARLES J. STONE, *Classification and Regression Trees*, Wadsworth & Brooks, (1984).

[3] ROBERT E. SCHAPIRE, The boosting approach to machine learning: An overview, *Nonlinear Estimation and Classification*, Springer, (2003). http://www.cs.princeton.edu/~schapire/boost.html.

[4] NELLO CRISTIANINI and JOHN SHAWE-TAYLOR, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, (2000).

## Exact Learning of Boolean Functions and Finite Automata with Queries

*Lisa Hellerstein, Polytechnic University, Brooklyn*

Which Boolean functions can be efficiently learned? The answer to this question clearly depends on the type of information that is available to the learner. In query models of learning, we assume that the learner has access to oracles that provide information about the function being learned. Query models are meant to capture situations in which the learner has access to a teacher who can answer questions, or can interact with the environment through experience or experiments. In this tutorial, we will define the standard models of query learning, and discuss efficient algorithms for learning particular types of Boolean functions, functional representations, and automata. We will also give an overview of techniques for proving lower

bounds on the number of queries needed to learn, and complexity-theoretic hardness results for learning.

### Reading List

[1] N. Littlestone, Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, *Machine Learning* 2, pp. 285–318, (1988).

[2] D. Angluin, Learning regular sets from queries and counterexamples, *Information and Computation* 75, pp. 87–106, (1987).

[3] W. Maass and G. Turan, On the complexity of learning from counterexamples and membership queries, *Proc. 31st Annual IEEE Symposium on Foundations of Computer Science*, pp. 203–210, (1990).

[4] Tibor Hegedus, Generalized teaching dimensions and the query complexity of learning, *Proc. 8th Annual Conference on Computational Learning Theory*, pp. 108–117, (1995).

[5] N. H. Bshouty, Exact learning of Boolean functions via the monotone theory, *Information and Computation* 123, pp. 146–153, (1995).

[6] R. E. Schapire and L. M. Sellie, Learning sparse multivariate polynomials over a field with queries and counterexamples, *J. Comput. Syst. Sci.* 52(2), pp. 201–213, (1996).

[7] L. Hellerstein, K. Pillaipakkamnatt, V. Raghavan, and D. Wilkins, How many queries are needed to learn?, *J. Association for Computing Machinery* 43(4–6), pp. 840–862, (1996).

[8] A. Beimel, F. Bergadano, N. H. Bshouty, E. Kushilevitz, and S. Varricchio, Learning functions represented as multiplicity automata, *J. ACM* 47(3), pp. 506–530, (2000).

[9] A. Klivans and R. Servedio, Learning DNF in time $2^{O(n^{1/3})}$, *J. Computer and System Sciences* 68(2), pp. 303–318, (2004).

## Online Learning

### Adam Tauman Kalai, Weizmann Institute of Science and Toyota Technological Institute

In online learning, a learner makes sequential predictions about data one at a time, online. After each prediction is made, the learner is informed of the quality of its prediction. Elegant learning algorithms come with surprisingly strong guarantees for online learning of such sequential prediction problems. One striking aspect of these guarantees is that they hold for arbitrary sequences of data—no distributional assumptions (such as independence) are required. Nonetheless, online analysis is typically very simple and implies analogous results in more standard learning settings that involve probabilistic assumptions.

Related results in online learning span a number of fields including learning theory, game theory, and information theory.

### Reading List

[1] Nicolo Cesa-Bianchi and Gabor Lugosi, *Prediction, Learning and Games*, Cambridge University Press, (2006).

[2] Avrim Blum, On-line algorithms in machine learning, *Online Algorithms: The State of the Art*, Chapter 14, LNCS # 1442, Eds. Fiat and Woeginger (1998). `http://www.cs.cmu.edu/afs/cs/usr/avrim/www/Papers/survey.ps`.

[3] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire, Gambling in a rigged casino: The adversarial multi-armed bandit problem, *Proc.36th Annual Symposium on Foundations of Computer Science*, pp. 322–331, (1995).

`http://www.cs.ualberta.ca/~bowling/classes/cmput608/AuerEtAl95.pdf`.

[4] Nick Littlestone, Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm, *Machine Learning* 2, pp. 285–318, (1987). `http://www.cse.ucsc.edu/classes/cmps242/Fall02/paps/winnow.ps`.

## Spectral Methods for Visualization and Analysis of High Dimensional Data

### Lawrence Saul, University of California San Diego

How can we detect low-dimensional structure in high-dimensional data? If the data is mainly confined to a low-dimensional subspace, then simple linear methods can be used to discover the subspace and estimate its dimensionality. More generally, though, if the data lies on (or near) a low-dimensional submanifold, then its structure may be highly nonlinear, and linear methods are bound to fail.

Graph-based spectral methods have recently emerged as a powerful tool for high-dimensional data analysis. These methods are able to reveal low-dimensional structure in high-dimensional data from the top or bottom eigenvectors of specially constructed matrices. The matrices are constructed from sparse weighted graphs whose vertices represent input patterns and whose edges indicate neighborhood relations. The main computations in this framework are based on highly tractable optimizations, such as shortest path problems, least squares fits, semidefinite programming, and matrix diagonalization. In this talk, I will provide an overview of these methods.

### Reading List

[1] C. J. C. Burges, *Geometric Methods for Feature Extraction and Dimensional Reduction, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Researchers and Practitioners*, O. Maimon and L. Rokach (eds.), Kluwer Academic Publishers, (2005).

[2] L. K. Saul, K. Q. Weinberger, J. H. Ham, F. Sha, and D. D. Lee, Spectral methods for dimensionality reduction, to appear in *Semi-supervised Learning*, O. Chapelle, B. Schoelkopf, and A. Zien (eds.), MIT Press, (2006).