Machine Learning – Statistical and Computational Foundations
CS489/698

Lecturer: Shai Ben-David
Intended term: Winter 08
Times: 11:30-12:50 TuTh
Location: MC4042

What is the course about?

Machine Learning (ML), and its neighboring subfields of *knowledge discovery in databases* (KDD, sometimes referred to simply as *data mining*) include on the one hand the automated analysis of large data sets using intelligent algorithms that are capable of extracting from the collected data hidden knowledge in order to produce models that can be used for prediction and decision making. On the other hand, they also include algorithms and systems that are capable of learning from experience and adapting to their environment or their users.

Given the enormous growth of collected and available data in companies, industry and science, techniques for analyzing such data are becoming ever more important. Consequently, Machine Learning is a fast growing topic both as an academic discipline and in practical research and development. It plays a central role in a wide range of important applications emerging from need to process data sets whose sizes and complexities are beyond the ability of humans to handle.
Research in knowledge discovery and machine learning combines classical questions of computer science (efficient algorithms, software systems, databases) with elements from artificial intelligence and statistics up to user oriented issues (visualization, interactive mining).

This course focuses on the theoretical backbones of machine learning, discussing the essential challenges, mathematical techniques and solutions upon which the present and future practical tools are based. This relatively young field draws from several established mathematical areas including statistics, geometry, combinatorics, and computational complexity.

Intended Audience:

CS and Math students who are interested in the interaction between mathematical and computational theory and real world applications. In the case of this course, we consider the application of tools and ideas of statistics and algorithms to data mining and AI issues.

Course objectives:

The course is aimed to familiarize the students with the theoretical foundations underlying some of the most useful machine learning techniques. For students interested

in future work in data mining, AI, and bioinformatics related areas it will provide an understanding of the principles behind some existing useful tools, as well as a basis for the development of further topic-specific applications. For theoretically oriented students my intention is to make them aware of the beauty and challenges of continuing studies and research in this intersection of CS, IT and statistics.

Related Courses:

Prerequisite: STAT 230 (or equivalent), CS 341

Marking Scheme:

Assignments (5) 30%

Final exam 70%

For interested students there will be an option to add a 30% for a project and reduce the weight of the final to 40%.

References:

Most of the material is covered in current machine learning textbooks, such as
John Shawe-Taylor & Nello Cristianini Kernel Methods for Pattern Analysis
Cambridge University Press, 2004,
Anthony and Bartlett "Neural Networks: Theoretical Foundations",
Or
Kearns and Vazirani "An Introduction to Computational Learning Theory" (MIT press, 1994).
However, some of the topics are based on more recent research and
will be based on some (reader friendly) recent research papers.

Syllabus Outline:

We shall cover the following topics, each will occupy roughly a week of teaching:

1. The statistical learning problem.
2. Basic pitfalls: Overfitting and the 'No Free Lunch' principle.
3. Traditional performance guarantees (based on Chernoff and Hoefding, inequalities).
4. Occam's Razor – Data-Compression based solutions.
5. Combinatorial tools: The Vapnik-Chervonenkis dimension and Sauer's Lemma.
6. The relationship between Statistics and Combinatorics: ε - Nets and ε - Approximations.
7. Applications of VC dimension and ε - Nets to Data Structures and other fields.
8. Generalization bounds based on the VC-dimension.
9. Regularization and Goodness-of-Fit/Complexity tradeoffs.
10. Algorithmic considerations – some basic learning algorithms and computational complexity lower bounds.

11. Query based learning models.
12. Online learning.
13. Relations between the different learning models.
14. Data mining applications, including clustering and change detection.