Machine Learning for Bioinformatics (COMP-766-02)

Winter Session, 2004

Taught by: Theodore J. Perkins
Office: McGill Centre for Bioinformatics
Email: perkins@mcb.mcgill.ca
Phone: 398-7071 x 09317

Course web page: www.mcb.mcgill.ca/~perkins/COMP76602/COMP76602.html
Location: Arts Building 210
Time: 10:35 am - 11:25 am, MWF

The purpose of this course is to introduce students with background in bioinformatics to the major principles and techniques of machine learning, and to look at how these can be applied to problems in bioinformatics. This course is aimed at students who have not previously studied machine learning and who may have limited background in probability and statistics, but who do have a basic background in computer science and an understanding of the problems studied in bioinformatics. The goals of this course are to:

1. Provide students with a "toolbox" of practical machine learning techniques that are useful for bioinformatics data analysis and research.
2. Describe proper methodology for applying machine learning techniques, and common pitfalls.
3. Give students enough expertise to understand and evaluate bioinformatics research papers that involves machine learning.
4. Provide a sense of what can and what cannot be inferred from data.
5. To examine which machine learning approaches have been most successful in bioinformatics to date.

Format: Approximately half of the classes will be lectures taught by Dr. Perkins, and half will be discussions of bioinformatics research papers that use machine learning.

Evaluation:

- 25% -- Homework assignments, which may include written and programming exercises. Expect 4 or 5 assignments of moderate length.
- 25% -- Research paper critiques and discussion. For classes in which a research paper is the main topic of discussion, students will write a short (1-2 page) evaluation of the strengths and weaknesses of the paper(s), and discuss potential improvements, alternative solutions, etc.
- 50% -- Final "exam".

Credits: 4

Prerequisites:

- 308-462: Methods in Computational Biology

Readings: The official text for the course is

- Bioinformatics: A Machine Learning Approach. Baldi, Brunak. MIT Press, 1999.

Other references:

- Machine Learning. Mitchell, McGraw-Hill, 1997
- The Elements of Statistical Learning. Hastie, Tibshirani, Friedman. Springer-Verlag, 2001
- Pattern Classification (2nd Edition). Duda, Hart, Stork. Wiley-Interscience, 2000
- Neural Networks for Pattern Recognition. Bishop. Oxford University Press, 1997
- Probabilistic Reasoning in Intelligent Systems. Pearl. Morgan Kaufmann Publishers Inc., 1988
- Statistical Methods in Bioinformatics. Grant, Evans. Springer-Verlag, 2001
- ... and numerous research papers from Nature, Science, ISMB, RECOMB, PSB, etc.

Course outline:

Section 1: Unsupervised Learning (Dimensionality reduction, visualization, clustering -- approx 2 weeks)

- Principal components analysis, independent components analysis, multidimensional scaling, "flat" clustering (such as k-means), hierarchical clustering

Section 2: Supervised Learning (a.k.a. function approximation -- approx 4 weeks)

- Classification and regression problems, nearest neighbor, perceptron, decision trees and regression trees, linear and logistic regression, artificial neural networks, support vector machines.

Section 3: Probabilistic Modeling (including some more supervised learning -- approx 4 weeks)

- Maximum likelihood and maximum a posteriori principles, discrete models, Markov chains, parametric estimation, Bayesian networks.

Section 4: Modeling Dynamical Systems (approx 2 weeks)

- Linear and nonlinear differential equations, dynamical Bayesian networks