Topics in Bioinformatics

Instructor:
Asst. Prof. Stephen Winters-Hilt
Phone: (504) 896-2761; (504) 280-2407
Cell: (985) 789-2258
E-mail: winters@cs.uno.edu

Time/Location:
Lecture Hours: MW 1:00-2:15
Location: Math 229
Office Location: CERM 217
Office Hours: MW 10:30-12:00

Prerequisites: CSCI 4567 or CSCI 4568, or permission of instructor.

Textbooks (required):
The Elements of Statistical Learning by Trevor Hastie *et al*. Springer-Verlag (2003). ISBN 0-387-95284-5.
Instructor notes

Reference Books (optional):
Biological Sequence Analysis by R. Durbin *et al*. Cambridge University Press (1999). ISBN 0-521-62971-3.
Hidden Markov Models for Bioinformatics by Timo Koski. Springer (2001). ISBN 1-402-00136-3.
Programming Perl (3rd ed.) by Larry Wall *et al*. O'Reilly Media (2000). ISBN 0-596-00027-8.

Abstract & Course Objectives:

This is an upper-level course that builds on the programming-intensive applications of machine learning research from CSCI 4567, 4568, 4587 and 4588, as well as interdisciplinary (biophysics/biochemistry/EE) applications of the above. The class projects will typically lay the foundation for the advanced projects the student can undertake in the 6000-level project courses (which often result in peer-reviewed journal publications). Last taught in Fall 2005 and Spring 2006, with interdisciplinary focus on nanopore detector cheminformatics.

(more)

General Machine Learning & Bioinformatics Project Objectives:

Real-world deployment.

Students should be familiar with training and testing in a real computational environment (including simple distributed computational arrangements on a networked cluster of computers to the extent that time permits).

Performance optimization.

Students should understand how to obtain statistically valid (objective) scores of performance and how to use that information for performance optimization.

Peer-reviewed Publication.

Some students are expected to have projects sufficiently mature that they will be asked, for their Final Project, to communicate their results as a paper submission.

Grading:

(A) 90-100; (B) 75-89; (C) 65-74; (D) 55-64; (F) below 55.

Homework assignments	30%
Midterm	10%
Final Project	60%

Students will learn to do the following:

1. Follow the most recent research in the field covered in the course
2. Solve the real-world informatics problem using techniques covered in the class
3. Provide incisive critiques to current research and point out some potential research direction
4. Final project is mature enough for journal paper submission

Policies:

- Omit documentation in your code at your own risk
- Academic honesty and fairness is expected at all times

Topics Potentially Covered:

I. Channel Current Cheminformatics

Time-domain Finite State Automata (FSAs) for signal acquisition (single-pass "scan" mode)
tFSAs for "spike" feature extraction
ad hoc preprocessing methods in other applications (ab initio gene finding)

II. Advanced Hidden Markov Model (HMM) Projects

Segmented-stationary signal analysis and kinetic feature extraction
Genome Structure Identification (generalized HMMs, with duration, possible SVM polarization)
TFBS Identification (using gIMM, SVM polarization)
Transcriptome Analysis (HMM feature extraction, SVM clustering)

(more)

III. Advanced Support Vector Machines (SVMs) Projects

SVM heuristic variants (alpha-selection, etc.)

Kernel Variants, Theory, and Mercer testing

SVM Variants for Multiclass Classification: Decomposition and Monolithic

SVM-based Clustering: Internal and External Formulations

IV. Boosting

Elimination of "hands-on" tuning of multiclass classifiers by boosting classifier ensembles

SVM boosting to eliminate tuning and/or explicit kernel selection

V. Bayesian networks, Dynamic Bayesian networks and applications

Discrete and continuous Bayesian network models and applications

VI. Optimization Methods, Deterministic and Stochastic optimization methods and applicatons to Bioinformatics (i.e. Neural Networks, Genetic Algorithms, Monte-Carlo Methods, Simplex methods, etc.)