Advanced Machine Learning Methods in Bioinformatics I

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

Time/Location:
Lecture Hours: MW 6:00-7:15pm
Location: Math Bldg, rm 105

Lab/Office Hours: MW 4 – 6pm
Lab Location: Math 342
Office Location: Math 312D
Admin Office Location: Math 312

Prerequisites: CSCI 2125, or permission of instructor; either CSCI 4567 or CSCI 4569, or permission of instructor; and either CSCI 4589 or 4590, or permission of instructor.

Textbooks (required):
(1) An Intro. to Support Vector Machines by Nello Cristianini
Reference Books (optional):
(1) Biological Sequence Analysis by Richard Durbin et al. (ISBN 0-521-62971-3)
(2) Programming Perl, 3rd edition, by Larry Wall et al.

Abstract & Course Objectives:
This will be a programming-intensive course focusing on project efforts in bioinformatics and cheminformatics. The project areas include: DNA analysis for gene finding (a hands on approach), data mining, and electrical signal analysis for biomedical informatics. For those interested in more theoretical projects, machine-learning projects exploring the strengths of the informatics approaches are also available. Special attention is given to statistical methods for identifying motifs in biosequences and identifying features in channel currents. One such method involves hidden Markov models for identifying structure in stochastic sequential data (for gene finding and for feature extraction from protein-channel ionic current measurements). One objective is to incorporate discriminative methods into the highly successful dynamic programming table based approaches, particularly via Support Vector Machine discrimination side-information at the table's cell-level.

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	10%
Midterm	10%
Final/Project	80%

Policies:

- Final Projects must be done individually
- Homework is due in class on due date specified
- Omit documentation in your code at your own risk

Topics Covered:

I. Introduction -- Informatics
- Bayesian Statistics (Graphical Models)
- Information Theory
- Stochastic Sequential Analysis
- Tools: Perl, C, Linux

II. Bioinformatics
- Markov Chains; HMM; gIMM
- pairwise alignment
- Genome Structure Identification (HMM, SVM polarization)
- TFBS Ident. (gIMM, SVM polarization)
- Transcriptome Structure Identification (feature extraction, SVM clustering)

III. Cheminformatics
- HMM/EM
- HMM Projection

IV. Immunoinformatics
- Neural Nets for feature extraction
- immunological screening