Advanced Machine Learning Methods in Bioinformatics II

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 SVM-based machine learning methods and applications (particularly for classification in cheminformatics and bioinformatics). Projects will generally start by implementing a Support Vector Machine (SVM), and then be applied towards computational genomics or biomolecular engineering problems, etc. The project areas include: classifier (algorithm) development, kernel development, development of acgt maps to reals, development of real maps to reals (pre-processing). For those interested in more theoretical projects, machine-learning projects exploring the strengths of the informatics approaches are also possible.

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
- Variational Calculus

II. Support Vector Machines (SVMs)
- Sequential Minimal Optimization (SMO)
- SMO Variants (WH-SMO, Kheerthi, etc.)
- Kernel Theory (entropic, prob. vectors); Kernel Testing (mercer conditions)
- **Internal and External multiclass implementations for classification and clustering**

III. Bioinformatics Applications
- Genome Structure Identification (HMM, SVM polarization)
- TFBS Identification using gIMM, SVM polarization
- Transcriptome Structure Identification (feature extraction, SVM clustering)

IV. Cheminformatics Applications
- SVM (Decision Tree)
- Blockade Structure Identification (SVM clustering) → antibody blockade classes

V. Immunoinformatics Applications
- Boosting: SVM Decision Forest → test on antibody study, compare to SVM (Monolithic)