Machine Learning Methods in Bioinformatics II

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):

An Introduction to Support Vector Machines and Other Kernel-based Learning Methods by Nello Cristianini & John Shawe-Taylor. Cambridge University Press (2000). ISBN 0-521-78019-5.

Reference Books (optional):

Biological Sequence Analysis by R. Durbin *et al*. Cambridge University Press (1999). ISBN 0-521-62971-3.

Programming Perl (3rd ed.) by Larry Wall *et al*. O'Reilly Media (2000). ISBN 0-596-00027-8.

Background:

Machine Learning Methods for Classification and Clustering are introduced. This course delves further into the advanced *classification/clustering methods* along the lines introduced in CSCI 4567 or 4568. There is a large project component to the course with a wide selection of problems, from programming intensive informatics solutions to theoretical/computational explorations.

Course Abstract:

Last taught in Spring 2005 with focus on Support Vector Machines for general, non-parametric, classification and clustering, with applications in Bioinformatics & Cheminformatics. This is the precursor to the graduate-level course on classification, clustering, and rule-based knowledge discovery: CSCI 6588.

Course Objectives:

Task decomposition.

Students should understand how to decompose a complex informatics task into a collection of standard informatics tasks: feature identification and knowledge discovery, signal acquisition and filtering, feature extraction, classification, and data-rejection.

Method selection.

Students should understand how to analyze the general properties of their data and factor in their computational limitations in order to select the most efficient informatics method at each stage of the task decomposition.

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.

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 directions
4. Final project is mature enough for journal paper submission

Policies:

- Most of the assignments can done with others
- 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:

Perl	*week 1,2*
Information Theory	*week 3,4*
Unsupervised Learning Methods	*week 5*
Supervised Learning	*week 6*
Support Vector Machines	*week 7*
Kernels	*week 8*
Generalization	*week 9*
Bioinformatics applications	*week 10,11*
Project presentations	*week 12-14*