BINF 5112 Computer Science Seminar for Bioinformatics
Machine Learning in Bioinformatics
Spring 2010

Instructor:
Olac Fuentes
ofuentes@utep.edu
www.cs.utep.edu/ofuentes
(915) 747-6956
Office hours: Tuesdays and Thursdays 11:00-12:00, or by appointment, in CSB 208 (feel free to drop by at other times if my door is open).

Meeting Times: Wed. 10:30 - 11:20 a.m in CSB 221.

Introduction:
Machine Learning studies the development of programs that can improve in the performance of a task with experience. For many difficult problems, solutions based on machine learning outperform all other solutions proposed to date. Examples of these problems include speech recognition, classification of objects in images, weather prediction, fraud detection, robot navigation, and many others.

In this course we will discuss several of the most commonly used machine learning algorithms and their application to problems in bioinformatics. Most meetings will consist of a presentation of either a paper from the scientific literature or a chapter from a textbook, followed by a discussion session.

Course Outcomes:
Upon completion of the course, students will:

- Have a basic understanding of the most commonly-used machine learning algorithms.
- Be able to use standard machine learning and data mining software tools to solve bioinformatics problems

Course Contents:

1) Introduction
   a) What is Machine Learning?
   b) Learning algorithms
2) Learning Algorithms
   a) Neural Networks
      i) Feed forward neural networks
   b) Decision trees
      i) ID3
      ii) C4.5
   c) Graphical Models
      i) Hidden Markov models
      ii) Bayes nets
      iii) Conditional random fields
   d) Kernel Methods
      i) Support Vector Machines
   e) Evolutionary Algorithms

        i) Genetic algorithms
    f) Instance-based learning
        i) k-nearest neighbors
        ii) Locally-weighted regression
    g) Probabilistic Methods
        i) Expectation maximization
        ii) Fisher's linear discriminant
    h) Ensembles of classifiers
        i) Boosting
        ii) Bagging
        iii) Randomization
        iv) Stacking
        v) Error-correcting output coding
3) Applications
    a) Folding prediction
        i) RNA folding
        ii) DNA folding
        iii) Protein folding.
    b) Mining the biological literature
    c) Gene regulatory networks
    d) Finding replication origins in DNA
    e) Gene finding
4) Tools
    a) WEKA
    b) Matlab

Pre-requisites:
There are no formal prerequisites, but knowledge of biology, programming, elementary calculus, linear algebra, probability, and statistics is useful.

Grading:
Class participation and presentations 100%

Bibliography:

Books:

- Pierre Baldi and Soren Brunak, *Bioinformatics – The Machine Learning Approach*, MIT Press, 2001.
- Tom Mitchell, Machine Learning, McGraw-Hill, 1997.
- Bernhard Scholkopf, Koji Tsuda and Jean-Philippe Vert, Kernel Methods in Computational Biology, MIT Press, 2001.
- Edward Keedwell and Ajit Narayanan, Intelligent Bioinformatics, Wiley, 2005.

Papers:

- E. Birney, Hidden Markov Models in Biological Sequence Analysis, IBM J. RES. DEV, Vol. 45(3), 2001.
- Jianlin Cheng and Pierre Baldi, A machine learning information retrieval approach to protein fold recognition, Bioinformatics Vol. 22 no. 12 2006, pages 1456-1463.
- C.B. Do, D.A. Woods, and S. Batzoglou, CONTRAfold: RNA secondary structure prediction without physics-based models, Bioinformatics, Vol. 22(14), 2006.

- J. Hu, Y.D. Yang, and D. Kihara, EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences, BMC Bioinformatics, Vol. 7(1), 2006.
- Z. Kou, W. W. Cohen, W.W. and R. F. Murphy, R.F., High-recall protein entity recognition using a dictionary, Bioinformatics, Vol. 21(1), 2005.
- L. K. Matukumalli, J. J. Grefenstette, D. L. Hyten, and I. Y. Choi, Application of machine learning in SNP discovery, BMC Bioinformatics,Vol. 7(4), 2006.
- I. Melvin, E. Ie, R. Kuang, J. Weston, W. Noble Stafford and C. Leslie, SVM-Fold: a tool for discriminative multi-class protein fold and superfamily recognition. BMC Bioinformatics 2007, 8 (Suppl 4).
- Jakob Skou Pedersen and Jotun Hein, Gene finding with a hidden Markov model of genome structure and evolution, Bioinformatics Vol. 19 no. 2, 2003, pages 219–227
- Yvan Saeys, Iñaki Inza and Pedro Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics Vol. 23 no. 19 2007, pages 2507–2517
- B.A. Shapiro, W. Kasprzak, C. Grunewald, and J. Aman, Graphical exploratory data analysis of RNA secondary structure dynamics predicted by the massively parallel genetic algorithm, Journal of Molecular Graphics and Modelling, Vol. 25(4), 2006.
- J. Yu and X. W. Chen, Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data, Bioinformatics, Vol 21(1), 2005.

Attendance Policy:

Students missing more than two class meeting will receive an "F" for the course. Two tardies will count as one absence. A tardy will be recorded each time a student shows up five minutes after the start of class.

Cell Phone Policy:

Cellular telephones are prohibited. Students are required to turn off their cellular telephones before entering the classroom. If your cell phone rings while you are at a lecture we will mark you as absent

Military Statement:

If you are a military student with the potential of being called to military service and/or training during the course of the semester, you are encouraged to contact as soon as possible.

Standards of Conduct and Academic Dishonesty:

You are expected to conduct yourself in a professional and courteous manner, as prescribed by the UTEP Standards of Conduct: http://studentaffairs.utep.edu/Default.aspx?tabid=4386

Academic dishonesty includes but is not limited to cheating, plagiarism and collusion. Cheating may involve copying from or providing information to another student, possessing unauthorized materials during a test, or falsifying data (for example program outputs) in laboratory reports. Plagiarism occurs when someone represents the work or ideas of another person as his/her own. Collusion involves collaborating with another person to commit an academically dishonest act.

Professors are required to - and will - report academic dishonesty and any other violation of the Standards of Conduct to the Dean of Students.