ENEE459M: Topics in Computer Engineering: Machine Learning and Data Mining

Instructor: Joseph JaJa

Spring 2010 Course Syllabus

Course Objectives: The course will introduce practical machine learning tools and techniques with applications to data mining using business, scientific, and web data sets. Techniques to be covered include: decision trees; learning rules; neural networks; Bayesian classification; support vector machines, association rules; and clustering. Students will acquire practical knowledge of these techniques through the use of the Weka software environment.

Course prerequisites: Senior level standing in ECE or consent of instructor

Prerequisite topics: good programming background, multivariable calculus, basic data structures and algorithms. Familiarity with basic probability concepts.

Textbook: Data Mining, Practical Machine Learning Tools and Techniques, Second Edition, I. H. Witten and E. Frank, Morgan-Kaufmann, 2005.

http://www.cs.waikato.ac.nz/ml/weka/book.html

Weka Software: Download Weka3.6 and the datasets UCI repository and regression datasets from the book web site.

Lecture Notes: will be available at
http://www.umiacs.umd.edu/~joseph//classes/459M/year2010/index.htm

Core Topics

1. **Introduction (Chapter 1)**
   - Machine Learning and Data Mining
   - Basic Terminology and Sample Applications

2. **Machine Learning Framework (Chapters 2 and 3)**
   - Data Representation
   - Concepts
   - Knowledge Representation
   - Examples of Decision Trees, Rules, Regression, and Clustering

3. **Introduction to Basic Machine Learning Models (Chapter 4)**
   - Rule-Based Classifiers

- Decision Trees
- Association Rules
- Linear Regression
- Nearest-Neighbor Classifiers
- Clustering

4. **Testing and Evaluation Methodologies (Chapter 5)**
   - Training and Testing
   - Validation Techniques
   - Performance Evaluation
   - Minimum Description Length Principle

5. **Detailed Description of Core Techniques (Chapter 6)**
   - Decision Trees
   - Neural Networks
   - Support Vector Machines
   - Regression
   - Clustering
   - Bayesian Networks

6. **Data Transformations (Sections 7.1-7.3)**
   - Discretization of Numerical Attributes
   - Principal Component Analysis
   - Handling Text

7. **Improved Methods (Sections 7.4-7.6)**
   - Automatic Data Cleansing
   - Combining Multiple Methods

Tentative Scheme for Course Grade:
- Three Midterms (Feb. 18, March 25, and April 22) – each worth 20%
- Final (May 17, 10:30-12:30) worth 30%
- Homeworks worth 10%

Meeting Times: Tu, Th 2-3:15 (EGR 3106)

Office Hours: Tu, Th 4-5:15 or by appointment

Contact Information: joseph@umiacs.umd.edu; 301-405-1925.