CSI5387/COMP5706
DATA MINING AND CONCEPT LEARNING

Fall 2012

Instructor: Dr. Stan Matwin
Office: SITE 5100
Contact: stan@eecs.uottawa.ca
Office hours: on request

Data Mining (DM) and **Machine Learning** (ML) have as their goal automatic acquisition of knowledge from examples. Obtaining concepts (models, profiles, definitions) from examples is a typical learning task. The examples may be, e.g., articles selected by a user from a news website, and the system will learn user's profile in terms of their interest. Or genetic expression model can be 'learned' from time series of DNA chip results. A functional view of Machine Learning is that it performs data analysis, and extraction of knowledge from data. Many ML techniques are used in Data Mining. The course will present state of the art techniques, as well some of their theoretical underpinnings and selected interesting applications.

This course will give you the general knowledge of the Machine Learning/Data Mining area and the background necessary for further self-study of advanced topics. Students will gain hands-on experience with selected data mining algorithms and tools. The topics covered include (some may not be discussed in class):

- Learning decision trees
- Introduction to evaluating the performance of DM/ML systems
- Probably Approximately Correct learning
- Basic probabilistic learning (Naïve Bayes)
- Kernel methods and Support Vector Machines
- Feature selection and discretization
- Advanced performance evaluation of ML/DM systems
- Data Mining Concepts - Association rules
- Clustering
- Semi-supervised learning: co-training
- Advanced Bayesian methods: Expectation Maximization
- Privacy issues and data mining algorithms
- Applications: Internet, bioinformatics, software engineering, etc.

The hands-on component of the course will give you practice in use of the main open-source Machine Learning and Data Mining systems, R and WEKA.

The final mark will consist of: the final exam (50%) and the evaluation of the practice (50%). The practice will consist of two parts: the assignment (15%) and the project (35%). The assignment on decision tree induction will be the same for all students, to be announced in the second or third class. For the project, the students will be encouraged to define their own assignment based on an application, thesis work, work project, etc. they are interested in. For those that will prefer an assigned project, one will be defined. Both the assignment and the project will involve experimentation with R and WEKA.
We will rely on extensive **course notes**. Slides and other course material can be found at http://www.eecs.uottawa.ca/~stan/csi5387/

PREREQUISITE: an undergraduate Introduction to AI course, basic knowledge of statistics, OR permission of the instructor. In the latter case, course webpage lists the sections of the AI textbook you will need before taking CSI5387.

PROGRAMMING: the main programming tools for this course will be WEKA (see http://www.cs.waikato.ac.nz/~ml/) and R (http://www.r-project.org/). No major programming effort is expected of the students, but some small-scale programming in Java or Perl may be required.

RESEARCH: I lead a large research group in the area of text and data mining, the Text Analysis and Machine Learning lab TAMALE), see www.tamale.uottawa.ca. We have weekly (more or less) open seminars in which advanced students and researchers from outside uOttawa present their current work. The seminar is open to all, and this year it takes place Thu 3:30-5:00 in SITE 5084. We start on Sep. 13, All students are welcome.

The TAMALE group is involved in a number of applied projects in ML/DM, and from time to time I am happy to offer RA's to qualified and interested graduate students who competed this course.