Syllabus for Data Mining
CS738
Fall 2000

Prof. Dan Simovici
dsim@cs.umb.edu
Tel.: 617-287-6472

I. The Goals of the Course: Data mining (DM) is the process of secondary analysis aimed at finding unsuspected relationships which are of interest or value for decision making. In this offering of the course I shall take the view that DM is the application of machine learning techniques in knowledge discovery in data. Consequently, the course will have two parts: in the first part I will teach computational learning theory using [1]. In the second part, you will present a topic that you read from [2] including experiments on real or synthetic data. The topics you may choose will be selected from a list that I include below.

II. Course Outline: The first part of the course shall cover the following topics:

1. Introduction to Data Mining and Computational Learning Theory.
2. Concept-Learning and General-to-Specific Ordering.
3. Decision Tree Learning.
4. Neural Networks and Learning.
5. Evaluation Hypotheses.

The topics that must be presented in the second part of the course are:

1. Association Rules.
2. Decision Trees.
3. Filters.
5. Clustering.

You have to form five groups no later than our last meeting in September and select one of these topics. You will have to give two or three presentations in class (after the mid-term) involving these issues in the second part of the course and to write an individual term paper on the presentation. Your presentation will have to involve three elements:
1. theoretical basis of the issues presented;
2. discussion of the relevant algorithms of the weka package;
3. generation of testing data (real or synthetic) and application of the algorithms.

I strongly suggest that you start immediately reading the relevant portion from [2] and you come to my office to discuss what you read. You should see me at least three times before your presentation.

III. Methods of Evaluation: The grade will be based on:

1. homework (20%);
2. a mid-term examination (25%);
3. an one-week presentation of a data mining algorithm (20%);
4. a term paper related to your presentation (35%).

References