

## Data Mining: Concepts and Algorithms

Instructor: Jiming Peng

Department of CAS, McMaster University.

ITB 107, Ext: 27746 Email: pengj@mcmaster.ca

Office hours: Thursday 14:00PM–15:00PM or by appointment.

**Time:** Monday, Wednesday, Thursday 10:30-11:20 AM.

**Location:** BSB/318 .

**TA:** Mr Huarong Chen ITB-208. Email: chenh4@mcmaster.ca. Ext: 27029.

<http://www.cas.mcmaster.ca/~cs4tf3>

**Background:** We expect a background in Data structure, Database system, a preliminary knowledge of statistics and some programming experience. Prerequisites are CS 4EB3 (or SE 3HO3), one of the introductory courses in statistics (3YO3, 1CC3, 1LO3, ECON 2BO3 in McMaster or an equivalent course in other universities), or permission of the instructor.

**Homework:** There will be 4 homeworks and a project. Assignments have to be delivered to your TA at his office time. Late homework will be marked with a late penalty of 20% per working day. The final project will be group work and each group consists of three or four students. There is also a lab time reserved for this course to help the students to grasp the basic concepts and techniques taught in the class. The free downloadable package Weka provides many data mining tools (see <http://www.mkp.com/datamining>).

You are permitted to discuss the general aspects of the course materials and assignments with your classmates. But the homework must be your individual effort. You are encouraged to consult other sources beyond the textbooks and the outside sources must be documented when you use them.

**Graduate Students** are required to give a half-hour presentation about some specific topics pertinent to the course.

There will be a two-hour mid-term exam.

Grading: You will be graded in the following way:

**Grad's:** Home works (30%), mid-term exam (30%), project (20%) and presentation (20%).

**Undergrad's:** Home works (30%), mid-term exam (40%), and project (30%).

### **Text Books:**

I.H. Witten and E. Frank: 1: Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementation. Morgan Kaufmann Publishers, 2000, ISBN 1-55860-552-5.

2: J. Han and M. Kamber, Data Mining, Concepts and Techniques. Morgan Kaufmann Publishers, 2001, ISBN 1-55860-489-8.

### **References:**

D. Hand, H. Mannila and P. Smyth, Principles of Data Mining, The MIT Press, 2001. ISBN 0-262-08290

TOM M. Mitchell, Machine Learning, The McGraw-Hill Companies INC, 1997, ISBN 0-07-042807-7.

**Course Description:** This course aims at giving a basic introduction to the newly-emerging multidisciplinary field: Data Mining. The course will present fundamental concepts and discuss main tasks in data mining. These include classification, association, prediction and clustering. Various algorithms based on decision tree, Bayes' model, instanced-based learning and numeric classifiers will be introduced. The preprocess and postprocess will be discussed as well.

## Tentative Schedule:

- Preliminaries (2 lectures)
  - What is Data Mining?
  - What is Machine Learning;
  - From Data Mining to Machine Learning.
- The Process of Learning (4 Lectures)
  - Concepts and Concept Descriptions in Learning
  - Instances and Attributes, Quantities;
  - Preparing the Input;
  - Output: Knowledge Representation;
- Basic Algorithms (6 Lectures)
  - Inferring rudimentary rules;
  - Statistical modelling;
  - Divide and Conquer for decision tree;
  - Covering algorithms;
  - Mining association rules;
  - Linear models and simple instances-based learning.
- Credibility (6 lectures)
  - Training and testing, Predicting performance;
  - Cross-validation and Other estimates;
  - Comparing mining schemes and predicting probabilities;
  - Counting the cost;
  - The minimum description length principle.
- Advanced Machine Learning Schemes (12 Lectures )
  - Extended decision trees;
  - Classification Rules;
  - Support vector machines;
  - Instance-based learning;
  - Numeric prediction;
  - Clustering.
- Input and Output: A revisit (4 lectures)
  - Attribute selection;
  - Discretizing numeric attributes;
  - Automatic data cleansing;
  - Combination of multiple models.
- Miscellaneous Topics in data Mining (2 lectures)
  - Learning from massive data set;
  - Text and web mining;