Introduction to Machine Learning and Data Mining

Instructor: Jiming Peng

Department of Computing and Software, McMaster University.
ITB 107, 525-9140 ext. 27746.
Email: pengj@mcmail.mcmaster.ca
Office hours: Tuesday 10:30AM–11:30AM.
Time: Tuesday, Thursday and Friday 8:30-9:20 AM.
Location: ABB/163 (on Campus).

Background:
We expect a background in Database system, a preliminary knowledge of statistics and some programming experience. Prerequisites are the same as for level 4.

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.
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 an one and half 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:

Course Description: The objective of this course is to develop fundamental skills in data mining and explore their applications. Data Mining is a recently emerging discipline that interacts with many areas such as database system, artificial intelligence, machine learning and statistics etc. Among others, machine learning provides the technical basis of data mining. This course presents some fundamental concepts involved in data mining and machine learning. We will learn basic methods for machine learning that are used in data mining. These algorithms provide approaches for some classical problems in data mining including classification, prediction and clustering. The applications of the methods and their implementation will be discussed as well.

Most course material will be presented by lectures and tutorials.
Tentative Schedule:

- Preliminaries (3 lectures)
  - What is Data Mining?
  - What is Machine Learning;
  - From Data Mining to Machine Learning.
- The Process of Learning (5 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 (5 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.
- Real Machine Learning Schemes (6 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.