Knowledge Discovery and Data Mining

CSC 492 Sec 01 / CSC 592 Sec 03

Syllabus – Fall 2005

Course Description

Our ability to collect and store data vastly outstrips out current ability to analyze data and derive useful information from it. The field of data mining holds some promise to provide relief in this area.

This course provides an overview of Knowledge Discovery and Data Mining (KDD). KDD deals with data integration techniques and with the discovery, interpretation and visualization of patterns in large collections of data. Topics covered in this course include data mining methods such as rule-based learning, decision trees, association rules and neural-networks; data visualization; and the cross industry standard process for data mining (CRISP-DM). The work discussed originates in the fields of artificial intelligence, machine learning, statistical data analysis, data visualization, databases, and information retrieval. Several scientific and industrial applications of KDD will be described. In particular, current applications to bioinformatics, e-commerce, and web mining will be studied.

In addition to the course work described above, students will also be required to complete several projects using the Weka data mining tool set (http://www.cs.waikato.ac.nz/ml/weka/). For this course we will use their latest version 3.4 available at the course website (http://homepage.cs.uri.edu/faculty/hamel/courses/fall2005/csc492).

Students taking this course for graduate credit (CSC 592) are also required to either write a paper on a data mining topic of their choice or implement a non-trivial piece of software; either an data mining algorithm or preprocessing tool for data mining.

Prerequisites

Strong computing skills are required. Undergraduate students must have completed a data structure course (CSC212) or equivalent. It is also required that students have taken one of the following: linear algebra (MTH215), discrete mathematics (CSC447), or introductory statistics (STA308). A background in artificial intelligence and/or databases is helpful but is not required.

Meeting Time and Instructor

MWF 10-11 Quinn Hall Rm 209.
Dr. Lutz Hamel, Tyler Hall, Room 251, email: hamel@cs.uri.edu

Required Text

"Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", 2nd Edition, Ian Witten and Eibe Frank, Morgan Kaufmann Publishers, 2005.

Grading

Midterm	25%
Final	25%
Projects/Homework	50%

Policies

- Check the website (often)! I will try to keep the website as up-to-date as possible.
- Class attendance, promptness, participation, and adequate preparation for each class are expected. If you are absent, it is your responsibility to find out what you missed (e.g. handouts, announcements, assignments, new material, etc.)

- **Late assignments:** Late assignments will be accepted with a penalty of 5% per day late. No assignments will be accepted if they are more than 7 days late.
- Make-up quizzes and exams will not be given without a valid excuse, such as illness. If you are unable to attend a scheduled examination due to valid reasons, please inform myself, or the department office in Tyler Hall, prior to the exam time. Under such circumstances, you are not to discuss the exam with any other class member until after a make-up exam has been completed.
- All work is to be the result of your own individual efforts unless explicitly stated otherwise. Plagiarism, unauthorized cooperation or any form of cheating will be brought to the attention of the Dean for disciplinary action. See the appropriate sections (8.27) of the University Manual.
- Software piracy will be dealt with exactly like stealing of university or departmental property. Any abuse of computer or software equipment will subject to disciplinary action.

Website
http://homepage.cs.uri.edu/faculty/hamel/courses/fall2005/csc492

Syllabus Outline
Week 1:

Introduction KDD and Data Mining (Chap 1)

- Data Mining and Machine Learning
- Machine Learning and Statistics
- Generalization as Search
- Data Mining and Ethics

Week 2:

Input: Concepts, Instances, Attributes (Chap 2)

- What is a Concept?
- What is an Example?
- What is in an Attribute?
- Preparing the Input
- CRISP-DM

Output: Knowledge Representation (Chap 3)

- Decision Tables
- Decision Trees
- Classification Rules
- Association Rules
- Rules involving Relations
- Trees for Numeric Predictions

- Neural Networks (Mitchell, Chap 4.1, 4.2, 4.3)
- Clusters

Week 3:

Decision Trees (Chap 4.3, 6.1)

- Divide and Conquer
- Calculating Information, Entropy
- Pruning
- Estimating Error Rates
- The C4.5 Algorithm

Week 4:

Evaluation of Learned Results (Chap 5.1, 5.2, 5.3)

- Training and Testing
- Predicting Performance
- Cross-Validation

Week 5:

Classification Rules (Chap 4.1, 4.4, 6.2)

- Inferring rudimentary Rules
- Covering Algorithms for Rule Construction
- Probability Measure for Rule Evaluation

Association Rules (Chap 4.5)

- Item Sets
- Rule Efficiency

Week 6:

Numeric Predictions (Chap 4.6, 5.8, 6.5)

- Linear Models for Classification and Numeric Predictions
- Numeric Predictions with Regression Trees
- Evaluating Numeric Predictions

Week 7:

Artificial Neural Networks (Chap 6.3; Mitchell, Chap 4.4, 4.5, 4.6, 4.7)

- Perceptrons
- Multilayer Networks
- The Backpropagation Algorithm
- Convergence and Local Minima

Week 8:

Clustering (Chap 6.6)

- Iterative Distance-based Clustering
- Incremental Clustering
- The EM Algorithm

Week 9:

Transformations (Chap 7)

- Attribute Selection
- Discretizing Numeric Attributes
- Combining Multiple Models

Week 10:

Moving On (Chap 8)

- Text Mining / Web Mining
- DNA Sequences and Data Mining
- Learning from Massive Data Sets