MONASH University
Information Technology

FIT5045
Knowledge discovery and data mining

Unit Guide

Semester 1, 2010

Last updated: 19 Feb 2010

Table of Contents

FIT5045 Knowledge discovery and data mining - Semester 1, 2010

Chief Examiner:

Dr Grace Rumantir
Lecturer
Phone: +61 3 990 31965
Fax: +61 3 990 31077

Lecturer(s) / Leader(s):

Gippsland

Associate Professor Kai Ming Ting
Director of Undergraduate Studies
Phone: +61 3 990 26241

Introduction

Welcome to FIT5045 Knowledge Discovery and Data Mining. This 6 point unit is an elective unit to all of the masters by coursework programs in the Faculty of IT. The unit has been designed to provide you with the fundamental principles of data mining and how it can be used to extract hidden patterns from data. It explores various data mining methods and its practical applications using a data mining tool.

Unit synopsis

Modern methods of discovering patterns in large-scale databases are introduced, including classification, clustering and association rules analysis. These are contrasted with more traditional methods of finding information from data, such as data queries. Data pre-processing methods for dealing with noisy and missing data and with dimensionality reduction are reviewed. Hands-on case studies in building data mining models are performed using a popular software package.

Learning outcomes

At the completion of this unit students will:

- be able to differentiate between supervised and unsupervised learning;
- know how to apply the main techniques for supervised and unsupervised learning;
- know how to use statistical methods for evaluating data mining models;
- be able to perform data pre-processing for data with outliers, incomplete and noisy data;
- be able to extract and analyse patterns from data using a data mining tool;
- have an understanding of the difference between discovery of hidden patterns and simple query extractions in a dataset;
- have an understanding of the different methods available to facilitate discovery of hidden patterns in a dataset;
- have developed the ability to preprocess data in preparation for data mining experiments;
- have developed the ability to evaluate the quality of data mining models;
- be able to appreciate the need to have representative sample input data to enable learning of patterns embedded in population data;
- be able to appreciate the need to provide quality input data to produce useful data mining models;
- have acquired the skill to use the common features in data mining tools;
- have acquired the skill to use the visualisation features in a data mining tools to facilitate knowledge discovery from a data set;
- have acquired the skill to compare data mining models based on the results on a set of performance criteria;
- be able to work in a team to extract knowledge from a common data set using different data mining methods and techniques.

Contact hours

2 hrs lectures/wk, 2 hrs laboratories/wk

Workload

Students are expected to commit to:

- two-hour lecture and

- two-hour tutorial (or laboratory) (requiring advance preparation)
- a minimum of 2-3 hours of personal study per one hour of contact time in order to satisfy the reading and assignment expectations.
- You will need to allocate up to 5 hours per week in some weeks, for use of a computer, including time for newsgroups/discussion groups.

Off-campus students generally do not attend lecture and tutorial sessions, however, you should plan to spend equivalent time working through the relevant resources and participating in discussion groups each week.

Unit relationships

Prerequisites

Sound fundamental knowledge in maths and statistics. Basic database and computer programming knowledge.

Prohibitions

CSE5230, FIT5024

Teaching and learning method

Teaching approach

This unit will be delivered via a weekly two-hour lecture. Lecturers may go through specific examples, give demonstrations and present slides that contain theoretical concepts.

In tutorials/practicals students will discuss in-depth fundamental and interesting aspects about data mining and have handons experience using data mining tools. The tutorials/practicals are particularly useful in helping students consolidate concepts and practise their problem solving skills.

Off-campus students will use the online forums to ask questions and to discuss with other students.

Timetable information

For information on timetabling for on-campus classes please refer to MUTTS, http://mutts.monash.edu.au/MUTTS/

Tutorial allocation

On-campus students should register for tutorials/laboratories using the Allocate+ system: http://allocate.its.monash.edu.au/

Off-Campus Learning or flexible delivery

Off-Campus students should treat the Unit Book (consisting of 12 modules) as their primary source for self-directed study. The modules contain text which is directed to leading you through the learning for each week. Also refer to the Unit Study Plan on the unit web page for further detail.

Online Discussion Forums are provided for the primary purpose of enabling off-campus students as well as on-campus students to engage with each other and the lecturer in Australia. The lecturer will expect all students to read these forums at least twice per week. In the forums, you may ask questions about the topics or exercises of each module, or to clarify interpretation of assignment tasks and marking criteria.

Unit Schedule

Week	Date*	Topic	Key dates
1	01/03/10	Unit Adminstration and Introduction to Data Mining	
2	08/03/10	Model Building	
3	15/03/10	Model Evaluation	
4	22/03/10	Data Preprocessing (1)	
5	29/03/10	Data Preprocessing (2)	
Mid semester break			
6	12/04/10	Classification	Assignment 01 due
7	19/04/10	Clustering	
8	26/04/10	Anomaly Detection	
9	03/05/10	Association Rules Mining (1)	

10	10/05/10	Association Rules Mining (2)	Assignment 02 due
11	17/05/10	Web Mining	
12	24/05/10	Data Mining and Information Visualization	
13	31/05/10	Revision	

*Please note that these dates may only apply to Australian campuses of Monash University. Off-shore students need to check the dates with their unit leader.

Unit Resources

Prescribed text(s) and readings

There is no one prescribed textbook for this unit. Students are expected to access the relevant chapters of the books on the recommended reading lists. Two of the books are available as online e-books in the library.

Text books are available from the Monash library and Monash University Book Shops. Availability from other suppliers cannot be assured. The Bookshop orders texts in specifically for this unit. You are advised to purchase your text book early.

Recommended text(s) and readings

Online e-books in the library:

J. Han & M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006I.H. Witten & E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd edition, Morgan Kaufmann, 2005

Other books:

R. Roiger and M. Geatz, Data Mining A Tutorial-based Primer, Pearson Education, Inc., 2003P. Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, Pearson Education, Inc., 2006G. Gupta, Introduction to Data Mining and Case Studies, Prentice-Hall, New Delhi, 2006A.B.M. Shawakat Ali and S. A. Wasimi, Data Mining: Methods and Techniques, Thomson Learning, 2007

Required software and/or hardware

You will need to download the data mining tool WEKA version 3.6 from
http://www.cs.waikato.ac.nz/ml/weka/

You will need to have Java http://www.java.com/ installed to run WEKA on your computer.

Equipment and consumables required or provided

Students studying off-campus are required to have the minimum system configuration specified by the faculty as a condition of accepting admission, and regular Internet access. On-campus students, and those studying at supported study locations may use the facilities available in the computing labs. Information about computer use for students is available from the ITS Student Resource Guide in the Monash University Handbook. You will need to allocate up to 6 hours per week for use of a computer, including time for newsgroups/discussion groups.

Study resources

Study resources we will provide for your study are:

all are available on Moodle.

Assessment

Overview

Examination (3 hours): 60%; In-semester assessment: 40%

Faculty assessment policy

To pass a unit which includes an examination as part of the assessment a student must obtain:

- 40% or more in the unit's examination, and
- 40% or more in the unit's total non-examination assessment, and
- an overall unit mark of 50% or more.

If a student does not achieve 40% or more in the unit examination or the unit non-examination total assessment, and the total mark for the unit is greater than 50% then a mark of no greater than 49-N will be recorded for the unit.

Assignment tasks

Assignment coversheets

Assignment coversheets are available via "Student Forms" on the Faculty website:
http://www.infotech.monash.edu.au/resources/student/forms/
You MUST submit a completed coversheet with all assignments, ensuring that the plagiarism declaration section is signed.

Assignment submission and return procedures, and assessment criteria will be specified with each assignment.

- **Assignment task 1**

  Title:
  Assignment 1
  Description:
  Students are to (i) use WEKA to perform various data mining methods on a data set and find a best performing for the task; and (i) answer conceptual questions
  Weighting:
  20%
  Due date:
  16 April 2010

- **Assignment task 2**

  Title:
  Assignment 2
  Description:
  This assignment requires students to use the data mining tool, WEKA, to explore several models and then choose one that will likely to produce the largest profit within the budgetary constraint for a mass mailing campaign.
  Weighting:
  20%

Due date:
5 May 2010

Examination

- **Weighting:** 60%
  Length: 3 hours
  Type (open/closed book): Closed book
  Remarks:

  to be conducted in the formal examination period

See Appendix for End of semester special consideration / deferred exams process.

Due dates and extensions

Please make every effort to submit work by the due dates. It is your responsibility to structure your study program around assignment deadlines, family, work and other commitments. Factors such as normal work pressures, vacations, etc. are not regarded as appropriate reasons for granting extensions. Students are advised to NOT assume that granting of an extension is a matter of course.

Students requesting an extension for any assessment during semester (eg. Assignments, tests or presentations) are required to submit a Special Consideration application form (in-semester exam/assessment task), along with original copies of supporting documentation, directly to their lecturer within two working days before the assessment submission deadline. Lecturers will provide specific outcomes directly to students via email within 2 working days. The lecturer reserves the right to refuse late applications.

A copy of the email or other written communication of an extension must be attached to the assignment submission.

Refer to the Faculty Special consideration webpage or further details and to access application forms: http://www.infotech.monash.edu.au/resources/student/equity/special-consideration.html

Late assignment

Assignments received after the due date will be subject to a penalty of 5% per day, including weekends.

Return dates

Students can expect assignments to be returned within two weeks of the submission date or after receipt, whichever is later.

Appendix

Please visit the following URL: http://www.infotech.monash.edu.au/units/appendix.html for further information about:

- Continuous improvement
- Unit evaluations
- Communication, participation and feedback
- Library access
- Monash University Studies Online (MUSO)
- Plagiarism, cheating and collusion
- Register of counselling about plagiarism
- Non-discriminatory language
- Students with disability
- End of semester special consideration / deferred exams