

MONASH UNIVERSITY
Faculty of Information Technology
CSE5230 Data Mining
Semester 2, 2004

Lectures

The lectures will be held in lecture room B2.15 from 4:00 p.m. to 6:00 p.m. on Thursdays.

Lecturer

The lecturer in charge of the unit is

David Squire

Room 5.23A, Building B, Caulfield Campus

Room 134, Building 75, Clayton Campus (Tuesdays and Wednesdays during second semester)

Email: David.Squire@csse.monash.edu.au

Phone: 9903 1033 (Caulfield), 9905 8307 (Clayton)

Tutorials

Tutorials will begin in week 2. Tutorials are currently scheduled for:

Thursday 6:00pm to 8:00pm – C2.03, B3.42, B3.45, B5.37

Thursday 8:00pm to 10:00pm – C2.03

The number of tutorials may be revised when the final enrolment numbers for the unit are known. Check the unit website for updates.

Objectives

To develop student knowledge of techniques and methods for data mining in large databases, including both those currently being used and those which are presently being researched; for students to become familiar with the currently available techniques for the extraction of knowledge from large databases. At the end of the unit the student should be able to describe the algorithms underlying the most common state-of-the-art data mining tools, and make an informed choice of data mining tool for a given problem. The student should have sufficient understanding to implement at least one fundamental data mining algorithm.

Unit website

Information and resources for this unit are available from the unit website:

<http://www.csse.monash.edu.au/courseware/cse5230/>

This site will be updated regularly. Students should check it at least once a week.

Recommended Reading

There is no set text for this unit. Students may find the following books useful:

Hand, D., Mannila, H. and Smyth, P.; **Principles of Data Mining**; MIT Press, 2001

Berry J.A. and Linoff G.; **Data Mining Techniques: For Marketing, Sales, and Customer Support**; John Wiley & Sons, Inc.; 1997

Cabena P., Hadjinian P., Stadler R., Verhees J., and Zanasi A.; **Discovering Data Mining: From Concept to Implementation**; Prentice Hall PTR, 1998

Fayyad U., Piatetsky-Shapiro G., Smyth P., and Uthurusamy R. (eds); **Advances in Knowledge Discovery and Data Mining**; AAAI Press, 1996

Kennedy R.L., Lee Y., Van Roy B., Reed C.D., and Lippman R.P.; **Solving Data Mining Problems Through Pattern Recognition**; Prentice Hall PTR, 1997

Witten I. H. and Frank, E.; **Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**; Morgan Kaufmann, 1999

Students will also have to read extensively in journals and conference proceedings to prepare their research papers. Many links to these resources are provided at the unit website. In addition, set reading will be provided, as described in the section below.

Assessment

Students will form groups of four or five (depending on final enrolment numbers). Each group will prepare a paper on a particular data mining technique and its applications.

A list of topics will be provided by the lecturer. Each group will express their preferences for topics on this list, and topics will then be allocated to groups by an algorithm that ensures that as many groups as possible get one of their top preferences. Details of the possible topics, along with set reading for these topics, will be available in week 2.

Individual literature survey document and tutorial sheets	15%
Individual implementation of a data mining algorithm	20%
A group paper on an agreed topic of approximately 5000 words	50%
Group presentation of the paper to the class	15%

Individual Literature Survey and Tutorial Tasks (15%): due week 6

The literature survey is due in week 6 and should consist of a **discussion** of the papers read, including the problems addressed, the techniques used, and their advantages and disadvantages. Students must discuss at least five (preferably more) articles covering the topic of their research paper. These papers must include the set reading from the lecturer, as well as papers located by the students themselves. The majority of papers surveyed must be academic papers, published in peer-reviewed conferences or journals, not magazine articles. There will also be tutorial tasks related to the literature survey.

Individual Implementation of a Data Mining Algorithm (20%): due week 10

Students will choose one of the fundamental data mining algorithms (e.g. *k*-means clustering, Naïve Bayes classification, ID3 decision tree, etc.) from a list provided by the lecturer in week 2. Students are to implement this algorithm using the language and platform of their choice. A sample data set will be provided to students for use when developing their algorithms. Assessment will be via the demonstration of their code to a tutor using a newly provided test data set, and the explanation of the code to the tutor. Students must demonstrate their understanding of the algorithms and data structures they have used.

Group Paper (50%): due week 12

Papers are to be approximately 5,000 words in length. A list of allowed paper topics will be available from the unit website. The group paper will be marked as follows:

- Understanding of technique/algorithm (or issue) 20
- Case studies 20
- Organization and clarity 5
- Accuracy of referencing 5

Students should make use of the Faculty Guide to Writing Assignments

<http://www.csse.monash.edu.au/~ajh/adt/studentguide/studentguide.html>, paying particular attention to section 4, "Citations", and section 5, "Quotations and Paraphrases".

Group Presentation (15%): weeks 12 and 13

Content	10%
Structure	3%
Presentation	2%

The presentation will last at least 20 minutes with 5 minutes for questions. Groups should provide copies of their overheads. All group members must participate in the presentation. Depending on the final number of groups, time for presentations may be extended.

University Policy on Cheating and Plagiarism

Students should consult university materials on cheating, in particular:

1. Statute 4.1 on Discipline at <http://www.monash.edu.au/pubs/calendar/statutes/statutes04.html> - [Heading102](#)
2. Student Resource Guide at <http://www.monash.edu.au/pubs/handbooks/srg/>, particularly the section on Cheating at <http://www.monash.edu.au/pubs/handbooks/srg/srg0072.htm>.
3. Student Resource Guide - section on Student Rights and Responsibilities at <http://www.monash.edu.au/pubs/handbooks/srg/srg0059.htm>.
4. Faculty policy at <http://www.csse.monash.edu.au/~ajh/adt/policies/cheating.html>.

It is the student's responsibility to make themselves familiar with the contents of these documents. All work submitted in this unit will be exhaustively checked for plagiarism and cheating, both manually and using automated text mining tools, such as Damocles (see <http://viper.csse.monash.edu.au/~damocles/about/>).

Syllabus Outline

Week	Lecture Content
1	Unit Outline and Introduction
2	Introduction to Machine Learning, Data Mining and Statistics;
3	Pre-processing for Data Mining
4	Market Basket Analysis; Clustering Techniques
5	Neural Networks 1: MLPs
6	Neural Networks 2: SOMs
7	Decision Trees
8	Information Visualization
9	Bayesian Classification and Bayesian Networks
10	Hidden Markov Models
11	Web Mining
12-13	Student presentations

(note: This is a provisional timetable. It is subject to revision)