MONASH University

CSE5230
Data mining

Unit guide

Semester 2, 2008

Last updated : 31 Jul 2008

Table of Contents

CSE5230 Data mining - Semester 2 , 2008

Unit leader :

Geoff Webb

Lecturer(s) :

Clayton

- Geoff Webb

Tutors(s) :

Clayton

- Minh Viet LE

Introduction

Unit synopsis

The unit explores various fundamental ``data mining'' techniques and their application areas. Supporting techniques like data pre-processing and statistics are also covered. Mention is made of the important relationships with each of machine learning, econometrics and inductive inference and with general over-arching techniques such as Minimum Message Length (MML).

Learning outcomes

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.

Workload

Unit relationships

Prerequisites

Basic mathematical and statistical skills and competency in at least one programming language. Some knowledge of database would be advantageous.

Relationships

CSE5230 is an elective unit in the Masters in Information Technology degree.

Continuous improvement

Monash is committed to ‘Excellence in education' and strives for the highest possible quality in teaching and learning. To monitor how successful we are in providing quality teaching and learning Monash regularly seeks feedback from students, employers and staff. Two of the formal ways that you are invited to provide feedback are through Unit Evaluations and through Monquest Teaching Evaluations.

One of the key formal ways students have to provide feedback is through Unit Evaluation Surveys. It is Monash policy for every unit offered to be evaluated each year. Students are strongly encouraged to complete the surveys as they are an important avenue for students to "have their say". The feedback is anonymous and provides the Faculty with evidence of aspects that students are satisfied and areas for improvement.

Student Evaluations

The Faculty of IT administers the Unit Evaluation surveys online through the my.monash portal, although for some smaller classes there may be alternative evaluations conducted in class.

If you wish to view how previous students rated this unit, please go to
http://www.monash.edu.au/unit-evaluation-reports/

Over the past few years the Faculty of Information Technology has made a number of improvements to its courses as a result of unit evaluation feedback. Some of these include systematic analysis and planning of unit improvements, and consistent assignment return guidelines.

Monquest Teaching Evaluation surveys may be used by some of your academic staff this semester. They are administered by the Centre for Higher Education Quality (CHEQ) and may be completed in class with a facilitator or on-line through the my.monash portal. The data provided to lecturers is completely anonymous.
Monquest surveys provide academic staff with evidence of the effectiveness of their teaching and identify areas for improvement. Individual Monquest reports are confidential, however, you can see the summary results of Monquest evaluations for 2006 at http://www.adm.monash.edu.au/cheq/evaluations/monquest/profiles/index.html

Unit staff - contact details

Unit leader

Professor Geoff Webb
Professor
Phone +61 3 990 53296
Fax +61 3 990 55157

Lecturer(s) :

Professor Geoff Webb
Professor
Phone +61 3 990 53296
Fax +61 3 990 55157

Tutor(s) :

Minh Viet LE

Teaching and learning method

Communication, participation and feedback

Monash aims to provide a learning environment in which students receive a range of ongoing feedback throughout their studies. You will receive feedback on your work and progress in this unit. This may take the form of group feedback, individual feedback, peer feedback, self-comparison, verbal and written feedback, discussions (on line and in class) as well as more formal feedback related to assignment marks and grades. You are encouraged to draw on a variety of feedback to enhance your learning.

It is essential that you take action immediately if you realise that you have a problem that is affecting your study. Semesters are short, so we can help you best if you let us know as soon as problems arise. Regardless of whether the problem is related directly to your progress in the unit, if it is likely to interfere with your progress you should discuss it with your lecturer or a Community Service counsellor as soon as possible.

Unit Schedule

Week	Topic	Key dates
1	Unit Outline and Introduction	
2	Introduction to Machine Learning, Data Mining and Statistics	
3	Pre-processing for Data Mining	
4	Clustering Techniques: Association Rule Discovery	
5	Classifiers 1: Bayesian Classification and Bayesian Networks	
6	Classifiers 2: Decision Trees	Literature Review Assignment Due
7	Neural Networks 1: MLPs	
8	Neural Networks 2: SOMs	
9	Genetic Algorithm	
10	Hidden Markov Models	Algorithm Implementation Assignment Due
11	Information Visualization	
Mid semester break		
12	Student presentations	Group Research Paper Due
13	Student presentations	

Unit Resources

Prescribed text(s) and readings

Nil. See a list of relevant texts and papers on the unit website.
Text books are available from the Monash University Book Shops and the Monash library.

Recommended text(s) and readings

See the reading list on the unit website.

Required software and/or hardware

Nil.

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.

Study resources

Study resources we will provide for your study are:

- Weekly lecture notes;
- Tutorial tasks and exercises;
- Assignment specifications;
- This Unit Guide outlining the administrative information for the unit;
- The unit web site on MUSO, where resources outlined above will be made available

Library access

The Monash University Library site contains details about borrowing rights and catalogue searching. To learn more about the library and the various resources available, please go to http://www.lib.monash.edu.au. Be sure to obtain a copy of the Library Guide, and if necessary, the instructions for remote access from the library website.

Monash University Studies Online (MUSO)

All unit and lecture materials are available through MUSO (Monash University Studies Online). Blackboard is the primary application used to deliver your unit resources. Some units will be piloted in Moodle. If your unit is piloted in Moodle, you will see a link from your Blackboard unit to Moodle (http://moodle.monash.edu.au) and can bookmark this link to access directly. In Moodle, from the Faculty of Information Technology category, click on the link for your unit.

You can access MUSO and Blackboard via the portal: http://my.monash.edu.au

Click on the Study and enrolment tab, then Blackboard under the MUSO learning systems.

In order for your Blackboard unit(s) to function correctly, your computer needs to be correctly configured.

For example:

- Blackboard supported browser
- Supported Java runtime environment

For more information, please visit: http://www.monash.edu.au/muso/support/students/downloadables-student.html

You can contact the MUSO Support by: Phone: (+61 3) 9903 1268

For further contact information including operational hours, please visit:
http://www.monash.edu.au/muso/support/students/contact.html

Further information can be obtained from the MUSO support site:
http://www.monash.edu.au/muso/support/index.html

Assessment

Unit assessment policy

Get a total of 50% of the total marks of all of the assignments.

Assignment tasks

- **Assignment Task**

  Title : Individual literature survey document and tutorial sheets

  Description :

  The literature survey should consist of a discussion of the papers read, including the problems addressed, the techniques used, and their advantages and disadvantages. The focus of the literature survey will be on a problem domain in which researchers have tried different data mining techniques to solve it. 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.

  Weighting : 15%

  Criteria for assessment :

  Understanding of techniques/algorithms (or issues) and their advantages and disadvantages

  Organization and clarity

  Accuracy of referencing

  Due date : Aug 18, 9:30am

  Remarks (optional - leave blank for none) :

  A soft copy of the literature survey must be submitted by the due date. A hard copy must be submitted to the tutor within 10 minutes of the start of the student's tutorial in week 6.

  The tutorial sheets must be submitted in the tutorials in weeks 4 and 5.

- **Assignment Task**

  Title : Individual implementation of a data mining algorithm

  Description :

  Aim

  The aim of this assignment is for you to demonstrate your detailed understanding of a simple data mining algorithm. You will do this by writing a program that implements the algorithm, and explaining how the code works to your tutor. You will also demonstrate the algorithm on a test data set provided by the lecturer.

Choice of Algorithms

You may choose one to implement any one of the follow algorithms:

- k-means clustering

- The Naive Bayes classifier

- ID3 decision tree

Handouts explaining each of these algorithms will be available from the unit web site during week 3.

Platform for Implementation

You can implement the algorithm using the language and platform of your choice. The only constraint is that you must be able to demonstrate your code during tutorials. This means that you must use either a language available in the labs, or bring in a lap-top to do the demonstration.

Weighting : 20%

Criteria for assessment :

Assessment will be via the demonstration of their code to a tutor using a newly provided test dataset, and the explanation of the code to the tutor. Students must demonstrate their understanding of the algorithms and data structures they have used.

Detailed assessment criteria will be issued along with the assignment. However, some broad guideline are:
1. All programs must run and compile correctly. Evidence of testing is required.
2. Programs must meet the problem specification
3. Code should be readable and maintable and follow the style recommended in the prescribed text book.
4. Programs should be documented
5. Students should be able to answer questions about their own work

Due date : Sept 15, 9:30am

Remarks (optional - leave blank for none) :

This assignment will be assessed in the tutorial classes on Monday Sept 15.

- **Assignment Task**

Title : A group paper on an agreed topic of approximately 5000 words

Description :

Each group will prepare a research paper on a particular data mining technique and its applications. At the end of the semester, each group will present their findings to the whole class.

Students will form groups of two or three (depending on final enrolment numbers), and email the following details to their tutor by the end of the tute in Week 3:

- The name of their group
- The list of the ids and names of the group members
- The topics they wish to work on in order of preference

Students having difficulty finding group members are encouraged to use the Feeback Forum on the unit website to seek others in the same situation. Each group will be assigned a data mining technique or issue as a topic from a list provided by the lecturer. The number of groups assigned to each topic will be minimised (for most topics, there will be two groups).

The group members must take responsibility for researching and writing each of these parts of the paper:

- A literature survey giving the research background for the technique and brief accounts of how and where it is applied.
- An explanation of how the technique and the algorithms implementing it actually work, preferably with a worked example.
- Two or more detailed case studies showing how the technique has been applied in business, industrial or scientific applications.

Papers are to be approximately 5,000 words. A list of allowed paper topics will be available from the unit web site. Here is a preliminary list of possible topics:

- Association Rule Discovery
- Back-propagation Neural Networks
- Self-Organising Maps
- Decision Trees
- Clustering
- Bayesian Networks
- Hidden Markov Models
- Visualisation for Data Mining
- Ethics and Data Mining

Weighting : 50%

Criteria for assessment :

Understanding of technique/algorithm (or issue)

Case studies

Organization and clarity

Accuracy of referencing

Due date : Oct 7

- **Assignment Task**

Title : Group presentation of the paper to the class

Description :

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.

Weighting : 15%

Criteria for assessment :

Content

Structure

Presentation

Due date : Weeks 12 and 13

Examinations

- **Examination**

  Weighting : 0%

  Length :

  Type (open/closed book) :

Assignment submission

Assignments will be submitted by paper and CD submission with the appropriate cover sheet correctly filled out and attached. Assignment submissions will be given to the students' respective tutor within the first 10 minutes of their respective tutes/pracs in the week of each assignment's due date. The literature review and group research paper must also be submitted to at least one of the Damocles online submission systems and/or MUSO. This will be discussed in class, and the relevant link(s) will be available on the unit web site.

Do not e-mail submissions. The due date is both the date by which the submission must be received and also the date by which the the submission is to be posted.

University and Faculty policy on assessment

Due dates and extensions

The due dates for the submission of assignments are given in the previous section. 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 seldom regarded as appropriate reasons for granting extensions. Students are advised to NOT assume that granting of an extension is a matter of course.

Late assignment

Assignments received after the due date will be subject to a penalty of **10% per day the assignment is late. Assignments received later than one week after the due date will not normally be accepted.**

Return dates

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

Assessment for the unit as a whole is in accordance with the provisions of the Monash University Education Policy at http://www.policy.monash.edu/policy-bank/academic/education/assessment/

We will aim to have assignment results made available to you within three weeks after assignment receipt.

Plagiarism, cheating and collusion

Plagiarism and cheating are regarded as very serious offences. In cases where cheating has been confirmed, students have been severely penalised, from losing all marks for an assignment, to facing disciplinary action at the Faculty level. While we would wish that all our students adhere to sound ethical conduct and honesty, I will ask you to acquaint yourself with Student Rights and Responsibilities (http://www.infotech.monash.edu.au/about/committees-groups/facboard/policies/studrights.html) and the Faculty regulations that apply to students detected cheating as these will be applied in all detected cases.

In this University, cheating means seeking to obtain an unfair advantage in any examination or any other written or practical work to be submitted or completed by a student for assessment. It includes the use, or attempted use, of any means to gain an unfair advantage for any assessable work in the unit, where the means is contrary to the instructions for such work.

When you submit an individual assessment item, such as a program, a report, an essay, assignment or other piece of work, under your name you are understood to be stating that this is your own work. If a submission is identical with, or similar to, someone else's work, an assumption of cheating may arise. If you are planning on working with another student, it is acceptable to undertake research together, and discuss problems, but it is not acceptable to jointly develop or share solutions unless this is specified by your lecturer.

Intentionally providing students with your solutions to assignments is classified as "assisting to cheat" and students who do this may be subject to disciplinary action. You should take reasonable care that your solution is not accidentally or deliberately obtained by other students. For example, do not leave copies of your work in progress on the hard drives of shared computers, and do not show your work to other students. If you believe this may have happened, please be sure to contact your lecturer as soon as possible.

Cheating also includes taking into an examination any material contrary to the regulations, including any bilingual dictionary, whether or not with the intention of using it to obtain an advantage.

Plagiarism involves the false representation of another person's ideas, or findings, as your own by either copying material or paraphrasing without citing sources. It is both professional and ethical to reference clearly the ideas and information that you have used from another writer. If the source is not identified, then you have plagiarised work of the other author. Plagiarism is a form of dishonesty that is insulting to the reader and grossly unfair to your student colleagues.

Register of counselling about plagiarism

The university requires faculties to keep a simple and confidential register to record counselling to students about plagiarism (e.g. warnings). The register is accessible to Associate Deans Teaching (or nominees) and, where requested, students concerned have access to their own details in the register. The register is to serve as a record of counselling about the nature of plagiarism, not as a record of allegations; and no provision of appeals in relation to the register is necessary or applicable.

Non-discriminatory language

The Faculty of Information Technology is committed to the use of non-discriminatory language in all forms of communication. Discriminatory language is that which refers in abusive terms to gender, race, age, sexual orientation, citizenship or nationality, ethnic or language background, physical or mental ability, or political or religious views, or which stereotypes groups in an adverse manner. This is not meant to preclude or inhibit legitimate academic debate on any issue; however, the language used in such debate should be non-discriminatory and sensitive to these matters. It is important to avoid the use of discriminatory language in your communications and written work. The most common form of discriminatory language in academic work tends to be in the area of gender inclusiveness. You are, therefore, requested to check for this and to ensure your work and communications are non-discriminatory in all respects.

Students with disabilities

Students with disabilities that may disadvantage them in assessment should seek advice from one of the following before completing assessment tasks and examinations:

- Faculty of Information Technology Student Service staff, and / or
- your Unit Coordinator, or
- Disabilities Liaison Unit

Deferred assessment and special consideration

Deferred assessment (not to be confused with an extension for submission of an assignment) may be granted in cases of extenuating personal circumstances such as serious personal illness or bereavement. Information and forms for Special Consideration and deferred assessment applications are available at http://www.monash.edu.au/exams/special-consideration.html. Contact the Faculty's Student Services staff at your campus for further information and advice.