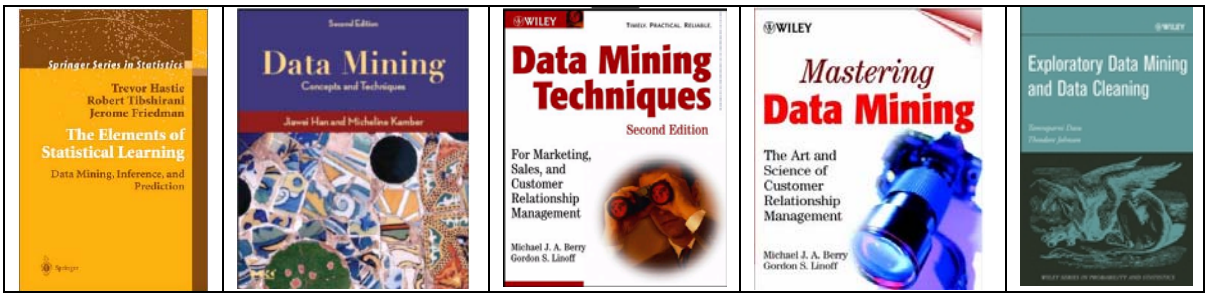
MACQUARIE UNIVERSITY ~ SYDNEY

STAT828 – Data Mining

First Semester 2007

Unit Outline



Students in this unit should read this unit outline carefully at the start of semester. It contains important information about the unit. If anything in it is unclear, please consult one of the teaching staff in the unit.

ABOUT THIS UNIT

Data Mining is about discovering patterns in the big data sets, and converting data into information or learning from data. The emphasis is on the data and the ways to convert data into information so that decision making is supported by facts.

Data mining uses techniques from different disciplines such as statistics, computing and machine learning.

This course is designed to introduce students to data mining techniques. At least two different software packages will be used to apply the different methods to discover information from different data sources such as business data, health data and biological data.

The first part of the course will cover descriptive data mining which will concentrate on exploratory tools such as graphical displays and descriptive statistics by using R and Clementine. The second part will introduce the model building and predictive data mining such as classification, regression, market basket analysis and clustering.

Software:

R We will use open source software called R. You can download and install a copy of the program from the developers' web page: www.R-project.org, it is freely available to interested users.

R is a command line software, it might be hard to learn if you are not used to this kind of environment, however, the benefits of learning to use this software overweight of its disadvantages. The benefits include but not limited to: it is free; it is very flexible; great support from R community through news groups and you can use it after you complete the course.

Clementine: This is graphical based data mining software owned by SPSS and widely used by business.

Prerequisite: Introductory undergraduate unit in statistics

UNIT WEB PAGE

Information relating to this unit can be found by visiting the Macquarie University Statistics Department web site. The URL for this unit is

http://www.stat.mq.edu.au/units/stat828/

WEBCT ACCESS

There is a WebCT site for this subject. Students are required to log into WebCT using their Student ID Number and myMQ Portal Password (note,

information about how to get hold of your password is provided in the weblink below).

The Web site for WebCT log in is: http://online.mq.edu.au
How to use WebCT is explained in the welcome message in the discussion board.

TEACHING STAFF

	
Lecturer In Charge:	**Lecturer**
Associate Professor Julian Leslie Room: E4A 554 Phone: 9850 8593 e-mail: jleslie@efs.mq.edu.au	Dr Ayse Bilgin Room: E4A 515 Phone: 9850 8509 e-mail: abilgin@efs.mq.edu.au

RECOMMENDED TEXT BOOKS – ONLINE

Data mining with R by Luís Torgo from
http://www.liacc.up.pt/~ltorgo/DataMiningWithR/

An Introduction to R – online manual http://www.r-project.org/

CRoss Industry Standard Process for Data Mining http://www.crisp-dm.org/download.htm

Introduction to Data Mining and Knowledge Discovery
http://www.twocrows.com/intro-dm.pdf

RECOMMENDED TEXT BOOKS

The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Trevor HASTIE, Robert TIBSHIRANI,and Jerome FRIEDMAN.New York: Springer-Verlag, 2001. ISBN 0-387-95284-5. (library call number Q325.75.F75 2001)

Data Mining: Concepts and techniques by Jiawei Han and Micheline Kamber, 2001, Morgan and Kaufmann (library call number QA76.9.D343.H36 2001) (There is also 2006 edition which is not available in the library)

Data mining techniques for marketing, sales and customer relationship management by Michael Berry and Gordon Linoff, 2004, John Wiley (library call number HF5415.125 .B47 2004)

Mastering Data Mining: The Art and Science of Customer Relationship Management by Michael J. A. Berry, Gordon S. Linoff, January 2000, John Wiley, ISBN: 978-0-471-33123-0 (library call number HF5415.125.B47/2000)

Exploratory Data Mining and Data Cleaning by Tamraparni Dasu, Theodore Johnson, May 2003 (library call number QA76.9.D343 D34 2003)

Statistics: An Introduction using R by Michael J. Crawley, March 2005, Wiley: ISBN: 0-470-02297-3 (library call number QA276.4 .C728)

Introductory Statistics with R by Peter Dalgaard, 2002, Springer (library call number QA276.4.D33 2002)

Other text books you might find useful for the course will be advised in the lectures.

CLASSES

Lectures

Lectures begin in Week 1. Students should attend **ONE** 2-hour session per week: Wednesdays between 6:00 and 8:00pm in **C5C 240**.

Tutorials

Tutorials also begin in Week 1. The aim of tutorials is to practise techniques learnt in lectures. They are designed so that students work through the exercises and ask as many questions as they need to improve their understanding. Tutors are the facilitators in the tutorial groups. They will assist students and instead of giving them straight answers for every question, they will create an environment for thinking process and discussion between the students.

Tutorials will be held in Statistical Computing Lab: **E4B 308**.

The timetable for classes can be found on the University web site at: http://www.timetables.mq.edu.au/

UNIT OBJECTIVES

- To introduce a variety of widely used data-mining techniques,
- to explain when and under what circumstances they can be used,
- to discuss the limitations of such methods,
- to demonstrate the use the freeware package “R” in carrying out some of these data-mining techniques,
- to demonstrate the use of the package SPSS “Clementine” in applying data-mining methods.

LEARNING OUTCOMES

By the end of this unit students will be able to:

- have an understanding of the principles and the concepts of the data mining
- summarise data and create visual summaries of data
- use market basket analysis to investigate the sales of a given company
- use classification and cluster analysis as data mining tools
- understand how the decision trees are developed and interpret decision trees
- to understand the link between descriptive and predictive data mining to support good decision making
- understand the role of logistic regression in data mining

GENERIC SKILLS

University study aims, not only to provide you with knowledge and skills in a particular academic discipline, but also to equip you with some generic skills. By studying this unit students will:

- improve their ability to work co-operatively as a team member
- have better problem solving skills
- improve their written communication skills, particularly report writing skills
- enhance their critical thinking skills through self assessment
- be confident in the use of different software packages for solving problems
- have opportunities to practice their presentation skills.

TEACHING AND LEARNING STRATEGY

- students are expected to attend all the lectures and the tutorials
- weekly tutorial exercises are set for individual development and considered formative assessment (no marks but suggestions to improve will be given each week to each student through mark lab exercises). Therefore, it is suggested that if students decide to work in

groups, the final product should be written individually and group work should be acknowledged to draw attention of the lecturer
- the group project(s) should be prepared in groups of four or less students and presented by every member of the group
- if for any reason, students can not hand in their assessment tasks on time, they have to contact one of the teaching staff in advance. It is strongly recommended that you use WebCT discussion or mail tool for communication
- students should hand in and collect their marked papers from ERIC (Economic Resource & Information Centre) E4B106. If you are unable to submit to ERIC, an electronic (word) file can be e-mailed to Ayse Bilgin through WebCT. Only word format files will be accepted; each page should have the student ID and student name as footer to eliminate any problems
- the solutions to lab exercises will not be given out however individual help for each student will be given during the consultation hours or through the WebCT discussion board

RELATIONSHIP BETWEEN ASSESSMENT AND LEARNING OUTCOMES

While attendance at classes is important, it is only a small proportion of the total workload for the unit: reading, research in the library, working with other students in groups, completing assignments, using the computer and private study are all parts of the work involved. At Macquarie it is expected that the average student should spend four hours per week per credit point.

Attendance at the examination is compulsory. The only exception to not sitting an examination at the designated time is because of documented illness or unavoidable disruption. In these circumstances you may wish to consider applying for Special Consideration. Information about unavoidable disruption and the special consideration process is available at http://www.reg.mq.edu.au/Forms/APSCon.pdf

If a Supplementary Examination is granted as a result of the Special Consideration process, the examination will be scheduled after the conclusion of the official examination period.

You are advised that it is Macquarie University policy not to set early examinations for individuals or groups of students. All students are expected to ensure that they are available until the end of the teaching semester that is the final day of the official examination period.

ASSESSMENTS

Weekly lab exercises are due at the BEGINNING of your lecture session on week following date of issue (e.g. Week 2 lab exercise solution is due in Week 3 before the lecture or by 6pm). You need to hand them into the appropriate box

in ERIC (E4B106). These weekly lab exercises will be corrected by your tutor/lecturer. There will not be any marks for these weekly exercises. It is important to collect the marked lab exercises from ERIC so that you can improve your learning. The quality of the submitted lab exercises will be used at the end of the term as guides for lecturers to decide on the grades. Solutions to at least 75% of lab exercises must be submitted. Failure to comply with this may result in exclusion from the unit.

The timetable for the two projects is given in the Unit Schedule at the end of this document.

Final examination (60%) is 3 hours long with 10 minutes reading time. It will examine any material covered throughout the course. The examination is 'closed book'. You may refer only to a single self-prepared A4 sheet of crib notes which may be written on both sides and must be easily readable. This summary must be submitted with your exam paper.
Calculators are permitted, but may be used only as calculators, and not as storage devices. No electronic devices (e.g. mobile phones, mp3 players) other than calculators are allowed during the exam.

NOTE: To obtain a passing grade, both coursework and exam performance must be satisfactory.

OVERALL ASSESSMENT

Students are expected to gain a reasonable level of proficiency in weekly topics.

The overall assessment for STAT828 is:

Project 1 (Individual)	15%
Poster for project 1	5%
Project 2	
Part A	
Market Basket Analysis Project (Individual)	5%
Part B	
Group Project	10%
PowerPoint Presentation of group project	5%
Final Exam	60%

The mark (SNG) recorded for this unit will be based on the weighted components above.

PLAGIARISM

The University defines plagiarism in its rules: "Plagiarism involves using the work of another person and presenting it as one's own." Plagiarism is a serious breach of the University's rules and carries significant penalties. You must read the University's practices and procedures on plagiarism. These can be found in the *Handbook of Undergraduate Studies* or on the web at: http://www.student.mq.edu.au/plagiarism/

The policies and procedures explain what plagiarism is, how to avoid it, the procedures that will be taken in cases of suspected plagiarism, and the penalties if you are found guilty. Penalties may include a deduction of marks, failure in the unit, and/or referral to the University Discipline Committee.

UNIVERSITY POLICY ON GRADING

Academic Senate has a set of guidelines on the distribution of grades across the range from fail to high distinction. Your final result will include one of these grades plus a standardised numerical grade (SNG).

On occasion your raw mark for a unit (i.e., the total of your marks for each assessment item) may not be the same as the SNG which you receive. Under the Senate guidelines, results may be scaled to ensure that there is a degree of comparability across the university, so that units with the same past performances of their students should achieve similar results.

It is important that you realise that the policy does not require that a minimum number of students are to be failed in any unit. In fact it does something like the opposite, in requiring examiners to explain their actions if more than 20% of students fail in a unit.

The process of scaling does not change the order of marks among students. A student who receives a higher raw mark than another will also receive a higher final scaled mark.

The grades and what they mean are given as below:

HD - High Distinction	Denotes a performance that meets all unit objectives in such an exceptional way and with such marked excellence that it deserves the highest level of recognition.
D - Distinction	Denotes performance that clearly deserves a very high level of recognition as an excellent achievement in the unit.
C -Credit	Denotes performance that is substantially better than would normally be expected of competent students in the unit.
P - Pass	Denotes performance that satisfies unit objectives.
PC - Conceded Pass	Denotes performance that meets unit objectives only marginally.
F - Fail	Denotes that a candidate has failed to complete a unit satisfactorily.

For further explanation of the policy see
http://www.mq.edu.au/senate/rules/Guidelines2003.doc or
http://www.mq.edu.au/senate/rules/detailedguidelines.doc.

ADVANCED STATISTICS COMPUTER LABS AND THEIR CONDITIONS OF USE

Obtaining User Account in these labs, each student will be given a user name and password for these labs once they are listed as enrolled in STAT828. After the first time logging into the server, the students need to change their password. The new (changed) password will expire in 30 days and needs to be changed again. If you do not change your password, you will not be able to login to the server again. If this happens, please talk to your tutor or the computer lab administrator:

Mr. Alfred Wong, awong@efs.mq.edu.au phone: 9850 6138

A time-table for the classes scheduled for each week will be displayed on the door of E4B 308. If there is a class in progress, students who are not enrolled in that class are not allowed to use the computers in the lab without permission from the tutor.

If you want to access to lab outside business hours, you need to apply for an access card. The card issued to you, can only be used by you. You can not pass the card for another person. You need to return your access card to your lecturer in the final exam since all cards will be disable after the exam you will not be able to use it. Ask your lecturer or tutor how you can get an access card.

PROBLEMS WITH LAB COMPUTERS?

Problems with lab computers should be reported as follows:

1. if the problem occurs during a class report problem to your tutor
2. if problem occurs outside class time, then report problem by phone or e-mail to the lab administrator

   Mr Alfred Wong awong@efs.mq.edu.au (ext 6138)

USING YOUR MU E-MAIL BROWSER ACCOUNT and no other (staff are instructed to ignore e-mails from Hotmail accounts, etc). BE SURE TO INCLUDE YOUR NAME AND CLASS, THE LAB AND PC NUMBER AND A BRIEF DESCRIPTION OF THE PROBLEM.

STUDENT SUPPORT SERVICES

Macquarie University provides a range of Academic Student Support Services. Details of these services can be accessed at
http://www.student.mq.edu.au.

STAT828 Data Mining
UNIT SCHEDULE

WEEK	LECTURE TOPIC		Assessment Given Out	Assessment Due
W1	Introduction to Data Mining Introduction to R		Lab Ex 1	
W2	Data Preprocessing, missing data, outliers Further R		Lab Ex 2	Lab Ex 1
W3	Descriptive and exploratory data mining Graphical displays with R		Lab Ex 3 **Project 1**	Lab Ex 2
W4	Introduction to Clementine Graphics and data exploration in Clementine		Lab Ex 4	Lab Ex 3
W5	Dimension reduction – concept hierarchies, PCA Techniques in R and in Clementine		Lab Ex 5	Lab Ex 4
W6	Classification: C 5.0 and C&R Trees, Neural Network, CHAID		Lab Ex 6	Lab Ex 5
SEMESTER BREAK				
W7	Anzac Day – Public Holiday			
W8	Classification continued		Lab Ex 7	Lab Ex 6 **Project 1 report and poster**
W9	Regression and Logistic Regression		Lab Ex 8 **Project 2**	Lab Ex 7
W10	Market Basket Analysis Methods in R and in Clementine		Lab Ex 9	Lab Ex 8
W11	Cluster Analysis K-means, Twostep, Hierarchical		Lab Ex 10	Lab Ex 9
W12	Cluster Analysis continued		Lab Ex 11	Lab Ex 10
W13	Real life examples (e.g. Guest Speaker) Revision			Lab Ex 11 **Project 2 report & presentation**

Note that all lab exercises are due by 6pm Wed in ERIC E4B106.