STAT828

Data Mining

S1 Evening 2014

Statistics

Contents

General Information	2
Learning Outcomes	2
Assessment Tasks	3
Delivery and Resources	7
Unit Schedule	9
Policies and Procedures	10
Graduate Capabilities	11
Recommended Text Books	15
Software Packages	16

General Information

Unit convenor and teaching staff
Unit Convenor
Ayse Bilgin
ayse.bilgin@mq.edu.au
Contact via ayse.bilgin@mq.edu.au
E4A515

Credit points
4

Prerequisites
Admission to MAppStat or PGDipAppStat or PGCertAppStat

Corequisites

Co-badged status
STAT728: Data Mining

Unit description
Data mining is emerging as an important analytical tool as organisations deal with increasingly large data sets. It is about discovering patterns in the big data sets, and converting data into information or learning from data. Data mining uses techniques from different disciplines such as statistics, computing and machine learning. This unit introduces relevant data mining techniques using a white box approach to illuminate the underlying algorithms and statistical principles. This unit is designed to inform students about the data mining techniques by arming them with a deeper understanding of the algorithms and statistical principles underlying the techniques. At least two different software packages will be used to apply the different methods to discover information from different data sources such as health and biological data. The first part of the unit will cover descriptive data mining, which will concentrate on exploratory tools such as graphical displays and descriptive statistics by using R and IBM SPSS Modeler. The second part will introduce the model building and predictive data mining such as classification, market basket analysis, clustering and more.

Important Academic Dates

Information about important academic dates including deadlines for withdrawing from units are available at https://www.mq.edu.au/study/calendar-of-dates

Learning Outcomes

On successful completion of this unit, you will be able to:

have an extensive understanding of the principles and the concepts of data mining methods and their applications
ability to apply creative thinking to resolve complex problems or issues as well as summarising complex multivariate data and creating visual summaries of such data
explain the link between descriptive and predictive data mining to support good decision making
examine and compare the differences between different decision trees and interpret sophisticated decision tree models for decision makers by writing a professional data mining report
analyse data sets by applying classification and cluster analysis methods and use their results to create an action plan for the management
apply market basket analysis to the sales data of a company, synthesise the results for a professional data mining report
demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and modelling to possible users
high-level research, analytical and conceptual skills and ability to apply these skills in development of models and client profiling

Assessment Tasks

Name	Weighting	Due
Data Mining Project Plan	5%	TBA
Market Basket Analysis Report	15%	TBA
Data Mining Project Report	20%	TBA
Data Mining Project Poster	5%	TBA
Participation in Lab Exercises	5%	TBA
Final Exam	50%	Examination Period

Data Mining Project Plan

Due: **TBA**
Weighting: **5%**

A project plan template will be provide in iLearn.

Market Basket Analysis Report

Due: **TBA**
Weighting: **15%**

Market Basket Analysis Project is an individual assessment task.

If you work with another student, you need to acknowledge it in your report.

Students are allowed to bring in a data set from their work place to work on, however, they need to consult Dr Bilgin for approval of the suitability of the data set for the project.

A model format and the examples of earlier reports will be provided through iLearn with the issue of the projects.

On successful completion you will be able to:

- have an extensive understanding of the principles and the concepts of data mining methods and their applications
- ability to apply creative thinking to resolve complex problems or issues as well as summarising complex multivariate data and creating visual summaries of such data
- explain the link between descriptive and predictive data mining to support good decision making
- apply market basket analysis to the sales data of a company, synthesise the results for a professional data mining report
- demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and modelling to possible users
- high-level research, analytical and conceptual skills and ability to apply these skills in development of models and client profiling

Data Mining Project Report

Due: **TBA**
Weighting: **20%**

Data Mining Project is a group work project.

Students will be put into groups as soon as possible (i.e. by week three) and they will be given opportunity to work on their project during tutorials.

Students are allowed to bring in a data set from their work place to work on, however, they need to consult Dr Bilgin for approval of the suitability of the data set for the project.

A model format and the examples of earlier reports will be provided through iLearn with the issue of the projects.

On successful completion you will be able to:

- have an extensive understanding of the principles and the concepts of data mining methods and their applications
- ability to apply creative thinking to resolve complex problems or issues as well as summarising complex multivariate data and creating visual summaries of such data
- explain the link between descriptive and predictive data mining to support good decision making
- examine and compare the differences between different decision trees and interpret sophisticated decision tree models for decision makers by writing a professional data mining report
- analyse data sets by applying classification and cluster analysis methods and use their results to create an action plan for the management
- demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and modelling to possible users
- high-level research, analytical and conceptual skills and ability to apply these skills in development of models and client profiling

Data Mining Project Poster

Due: **TBA**
Weighting: **5%**

One poster per group on iLearn by due date (power point document or pdf) clearly stating the group members. Also include a summary handout (see iLearn) to your submission (possibly pdf document).

On successful completion you will be able to:

- have an extensive understanding of the principles and the concepts of data mining methods and their applications
- ability to apply creative thinking to resolve complex problems or issues as well as summarising complex multivariate data and creating visual summaries of such data
- examine and compare the differences between different decision trees and interpret sophisticated decision tree models for decision makers by writing a professional data mining report

- demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and modelling to possible users
- high-level research, analytical and conceptual skills and ability to apply these skills in development of models and client profiling

Participation in Lab Exercises

Due: **TBA**
Weighting: **5%**

Lab exercise submission and contribution to tutorial discussions will be taken into account when allocating the marks. For individual due dates of lab exercises see iLearn.

On successful completion you will be able to:

- have an extensive understanding of the principles and the concepts of data mining methods and their applications
- analyse data sets by applying classification and cluster analysis methods and use their results to create an action plan for the management
- apply market basket analysis to the sales data of a company, synthesise the results for a professional data mining report
- demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and modelling to possible users

Final Exam

Due: **Examination Period**
Weighting: **50%**

Final examination is 3 hours long with 10 minutes reading time and will be held during the exam period. You will be permitted to bring an A4 sheet of notes, handwritten or typed, on both sides, into the final examination. 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.The final examination will be timetabled in the official University examination timetable. The University Examination timetable will be available in draft form approximately eight weeks before the commencement of the examinations and in final form approximately four weeks before the commencement of the examinations at: http://www.exams.mq.edu.au/exam/

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.mq.edu.au/policy/docs/special consideration/policy.html

You can submit your special consideration request(s) through the following link https://ask.mq.edu.au/index.php

Your final grade in STAT828 will be based on your work during the semester and in the final examination. You need to achieve the same standards both during the semester assessments and the final exam to be awarded a particular grade as set out in the Grading Policy (http://www.mq.edu.au/policy/docs/grading/policy.html).

On successful completion you will be able to:

- have an extensive understanding of the principles and the concepts of data mining methods and their applications
- ability to apply creative thinking to resolve complex problems or issues as well as summarising complex multivariate data and creating visual summaries of such data
- explain the link between descriptive and predictive data mining to support good decision making
- examine and compare the differences between different decision trees and interpret sophisticated decision tree models for decision makers by writing a professional data mining report
- analyse data sets by applying classification and cluster analysis methods and use their results to create an action plan for the management
- apply market basket analysis to the sales data of a company, synthesise the results for a professional data mining report
- demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and modelling to possible users
- high-level research, analytical and conceptual skills and ability to apply these skills in development of models and client profiling

Delivery and Resources

Changes to Content

The initial course notes for this unit have been developed by Associate Prof Julian Leslie and Dr Ayse Bilgin in 2007. Ms Gillian Miller made changes and added new topics to the course notes in 2010 and 2011. In 2012 and 2013, the notes are revised based on the current developments in the data mining discipline and students' feedbacks to the unit.

Classes

Lectures

Lectures begin in Week 1. Students should attend **ONE** 2-hour session per week: Mondays between 6:00 and 8:00pm in **EMC-G220 Faculty PC Lab**.

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 asking as many questions as they need to improve their understanding. Tutors are the facilitators in the tutorial groups. They will assist students and they will create an environment for thinking process and discussion between the students. Tutorials will be held on Mondays in **EMC-G220 Faculty PC Lab** between 8:00pm and 10:00pm.

Teaching and Learning Strategy

- Students are expected to attend all the lectures and the tutorials.
- Additional readings will be provided through iLearn to provide opportunities for students to increase their knowledge.
- Weekly tutorial exercises are set for individual development and considered formative assessment, although participation (see assessments) will be based on the submitted lab exercises. These lab exercises are designed to be completed by each individual to achieve the best learning. Therefore, it is suggested that if students decide to work together, the final product should be written individually and group work should be acknowledged to draw attention of the lecturer. Students need to bring a hard or soft copy of their completed (or in progress) lab exercises to the class for discussion, as well as submitting a soft copy electronically through iLearn.
- Projects are extensions to the lab exercises. They require applying the learned techniques to unseen data sets and writing professional reports.

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 packages to develop models and private study are all parts of the work involved.

Weekly lab (tutorial) 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 submit them through iLearn and bring a copy (soft or printed) to the class. You will be given opportunity during the tutorial to discuss your solution with your peers. These discussions will form part of the feedback to your submitted (prepared) lab exercises.

In addition to group discussions, suggested solutions to lab exercises will be provided through iLearn in a timely manner. You are expected to submit at least 8 of the lab exercises. Failure to comply with this may result in exclusion from the unit. Instead of content marking for the weekly lab exercises, a participation mark will be given to each student at the end of the semester based on the quality of their submissions (which will be shared by all students – details will be provided in the first lecture and within each lab exercise).

See Assessment Section for other assessment tasks.

If for any reason, students cannot hand in their assessment tasks on time, they have to contact the teaching staff in advance. No extensions for the lab exercises will be granted unless satisfactory documentation outlining illness or misadventure is submitted.

The marked assessments (projects) will be distributed during the tutorials by the Lecturer. Only word format files will be accepted; each page should have the student ID and student name as footer to eliminate any problems. When naming files please adopt the following convention: StudentID-(Your Surname)(Initial of Your First Name) – Assessment Task (Lab 1 or Assignment 1) e.g., **40000000-BilginA-Project 1**. No other format of naming the assessment tasks will be accepted. If you are unable to submit you assessment through iLearn (due to technical problems); an electronic (word) file can be e-mailed to Dr Ayse Bilgin (ayse.bilgin@mq.edu.au).

Unit Schedule

Week 1:Introduction to Data Mining & Introduction to R

Week 2:Data Preprocessing, missing data, outliers & Further R

Week 3:Descriptive and exploratory data mining, concept hierarchies & graphical displays with R

Week 4: Graphics and data explorations & Introduction to IBM SPSS Modeler

Week 5: Market Basket Analysis

Week 6: Classification (1)

Week 7: Classification (2)

Week 8: Classification (3)

Week 9: Cluster Analysis (1)

Week 10: Classification (4)

Week 11: Classification (5)

Week 12: Cluster Analysis (2)

Week 13: Revision and Data Mining Project Poster Presentations

Note that the order of the lectures might change and all lab exercises are due by 5:30pm a week after they are issued

Policies and Procedures

Macquarie University policies and procedures are accessible from Policy Central. Students should be aware of the following policies in particular with regard to Learning and Teaching:

Academic Honesty Policy http://mq.edu.au/policy/docs/academic_honesty/policy.html

Assessment Policy http://mq.edu.au/policy/docs/assessment/policy.html

Grading Policy http://mq.edu.au/policy/docs/grading/policy.html

Grade Appeal Policy http://mq.edu.au/policy/docs/gradeappeal/policy.html

Grievance Management Policy http://mq.edu.au/policy/docs/grievance_management/policy.html

Disruption to Studies Policy http://www.mq.edu.au/policy/docs/disruption_studies/policy.html *The Disruption to Studies Policy is effective from March 3 2014 and replaces the Special Consideration Policy.*

In addition, a number of other policies can be found in the Learning and Teaching Category of Policy Central.

Student Code of Conduct

Macquarie University students have a responsibility to be familiar with the Student Code of Conduct: https://students.mq.edu.au/support/student_conduct/

Grading Policy http://www.mq.edu.au/policy/docs/grading/policy.html

Student Support

Macquarie University provides a range of support services for students. For details, visit http://students.mq.edu.au/support/

Learning Skills

Learning Skills (mq.edu.au/learningskills) provides academic writing resources and study strategies to improve your marks and take control of your study.

- Workshops
- StudyWise
- Academic Integrity Module for Students
- Ask a Learning Adviser

The Macquarie University offers various workshops for the postgraduate students which you might find useful. The overviews and timetables can be accessed at http://www.students.mq.edu.au/support/learning_skills/postgraduate/workshops_for_postgraduate_students/

There are specific workshops for international students that help them to integrate into Australian

Education System http://www.international.mq.edu.au/.

Student Services and Support

Students with a disability are encouraged to contact the Disability Service who can provide appropriate help with any issues that arise during their studies.

Student Enquiries

For all student enquiries, visit Student Connect at ask.mq.edu.au

IT Help

For help with University computer systems and technology, visit http://informatics.mq.edu.au/help/.

When using the University's IT, you must adhere to the Acceptable Use Policy. The policy applies to all who connect to the MQ network including students.

Graduate Capabilities

PG - Discipline Knowledge and Skills

Our postgraduates will be able to demonstrate a significantly enhanced depth and breadth of knowledge, scholarly understanding, and specific subject content knowledge in their chosen fields.

This graduate capability is supported by:

Learning outcomes

- have an extensive understanding of the principles and the concepts of data mining methods and their applications
- ability to apply creative thinking to resolve complex problems or issues as well as summarising complex multivariate data and creating visual summaries of such data
- explain the link between descriptive and predictive data mining to support good decision making
- examine and compare the differences between different decision trees and interpret sophisticated decision tree models for decision makers by writing a professional data mining report
- analyse data sets by applying classification and cluster analysis methods and use their results to create an action plan for the management
- apply market basket analysis to the sales data of a company, synthesise the results for a professional data mining report
- demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and

modelling to possible users
- high-level research, analytical and conceptual skills and ability to apply these skills in development of models and client profiling

PG - Critical, Analytical and Integrative Thinking

Our postgraduates will be capable of utilising and reflecting on prior knowledge and experience, of applying higher level critical thinking skills, and of integrating and synthesising learning and knowledge from a range of sources and environments. A characteristic of this form of thinking is the generation of new, professionally oriented knowledge through personal or group-based critique of practice and theory.

This graduate capability is supported by:

Learning outcomes

- have an extensive understanding of the principles and the concepts of data mining methods and their applications
- ability to apply creative thinking to resolve complex problems or issues as well as summarising complex multivariate data and creating visual summaries of such data
- explain the link between descriptive and predictive data mining to support good decision making
- examine and compare the differences between different decision trees and interpret sophisticated decision tree models for decision makers by writing a professional data mining report
- analyse data sets by applying classification and cluster analysis methods and use their results to create an action plan for the management
- apply market basket analysis to the sales data of a company, synthesise the results for a professional data mining report
- demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and modelling to possible users
- high-level research, analytical and conceptual skills and ability to apply these skills in development of models and client profiling

PG - Research and Problem Solving Capability

Our postgraduates will be capable of systematic enquiry; able to use research skills to create new knowledge that can be applied to real world issues, or contribute to a field of study or practice to enhance society. They will be capable of creative questioning, problem finding and problem solving.

This graduate capability is supported by:

Learning outcomes

- have an extensive understanding of the principles and the concepts of data mining methods and their applications
- ability to apply creative thinking to resolve complex problems or issues as well as summarising complex multivariate data and creating visual summaries of such data
- examine and compare the differences between different decision trees and interpret sophisticated decision tree models for decision makers by writing a professional data mining report
- analyse data sets by applying classification and cluster analysis methods and use their results to create an action plan for the management
- apply market basket analysis to the sales data of a company, synthesise the results for a professional data mining report
- demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and modelling to possible users
- high-level research, analytical and conceptual skills and ability to apply these skills in development of models and client profiling

PG - Effective Communication

Our postgraduates will be able to communicate effectively and convey their views to different social, cultural, and professional audiences. They will be able to use a variety of technologically supported media to communicate with empathy using a range of written, spoken or visual formats.

This graduate capability is supported by:

Learning outcomes

- ability to apply creative thinking to resolve complex problems or issues as well as summarising complex multivariate data and creating visual summaries of such data
- explain the link between descriptive and predictive data mining to support good decision making
- examine and compare the differences between different decision trees and interpret sophisticated decision tree models for decision makers by writing a professional data mining report
- analyse data sets by applying classification and cluster analysis methods and use their results to create an action plan for the management
- apply market basket analysis to the sales data of a company, synthesise the results for a

professional data mining report
- demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and modelling to possible users
- high-level research, analytical and conceptual skills and ability to apply these skills in development of models and client profiling

PG - Engaged and Responsible, Active and Ethical Citizens

Our postgraduates will be ethically aware and capable of confident transformative action in relation to their professional responsibilities and the wider community. They will have a sense of connectedness with others and country and have a sense of mutual obligation. They will be able to appreciate the impact of their professional roles for social justice and inclusion related to national and global issues

This graduate capability is supported by:

Learning outcomes

- ability to apply creative thinking to resolve complex problems or issues as well as summarising complex multivariate data and creating visual summaries of such data
- examine and compare the differences between different decision trees and interpret sophisticated decision tree models for decision makers by writing a professional data mining report
- analyse data sets by applying classification and cluster analysis methods and use their results to create an action plan for the management
- apply market basket analysis to the sales data of a company, synthesise the results for a professional data mining report
- demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and modelling to possible users
- high-level research, analytical and conceptual skills and ability to apply these skills in development of models and client profiling

PG - Capable of Professional and Personal Judgment and Initiative

Our postgraduates will demonstrate a high standard of discernment and common sense in their professional and personal judgment. They will have the ability to make informed choices and decisions that reflect both the nature of their professional work and their personal perspectives.

This graduate capability is supported by:

Learning outcomes

- ability to apply creative thinking to resolve complex problems or issues as well as summarising complex multivariate data and creating visual summaries of such data
- explain the link between descriptive and predictive data mining to support good decision making
- examine and compare the differences between different decision trees and interpret sophisticated decision tree models for decision makers by writing a professional data mining report
- analyse data sets by applying classification and cluster analysis methods and use their results to create an action plan for the management
- apply market basket analysis to the sales data of a company, synthesise the results for a professional data mining report
- demonstrated level of knowledge and technical expertise in data mining activities, including cleaning and transformation of data; presentation of results of mining and modelling to possible users
- high-level research, analytical and conceptual skills and ability to apply these skills in development of models and client profiling

Recommended Text Books

The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Trevor HASTIE, Robert TIBSHIRANI,and Jerome FRIEDMAN.New York: Springer-Verlag, 2009. ISBN 9780387848570. (library call number Q325.75 .H37 2009 or the first edition Q325.75.F75 2001) Please note that for 2009 edition limited preview is available from MQ library web site (as google books).

An Introduction to Statistical Learning with Applications in R, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer-Verlag, 2014. ISBN 978-1-4614-7137-0 (at http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf)

Data Mining: Concepts and techniques by Jiawei Han and Micheline Kamber, 2006, Morgan and Kaufmann (library call number QA76.9.D343 H36 2006 or earlier version QA76.9.D343.H36 2001) Please note that for 2006 edition limited preview is available from MQ library web site (as google books)

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)

Knowledge discovery with support vector machines by Lutz Hamel, 2009, Wiley, limited view is available from Google Books

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 (CRISP-DM)

http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

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

Software Packages

R We will use open source software called R. You can download and install a copy of the program from the developers' web page: http://cran.r-project.org/ or www.R-project.org

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 its disadvantages. The benefits include and 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.

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

Learning management system (LMS)

There is a iLearn (which is modified Moodle) site for this unit where the required course materials for the unit will be posted. In addition, the forums are created for each week will enable us to communicate within the unit without having the danger of spam filters. The lecturers might make announcements via the online unit page therefore you should make sure you log in and read the posts at least twice a week.

The web page for the LMS is https://ilearn.mq.edu.au/login/MQ/, use your **Macquarie OneID** to log in.