

Lecture 0: Welcome and Introduction

- Very short introduction to data mining (and why should I take this course...?)
- Course overview (introduction of all lecturers and modules)
- Administrative matters (course schedule, lecturing and tutorial times, assessment, etc)
- Course resources
- Questions and any other issues

Very short introduction to data mining (1)

- Many companies, organisations and research projects collect huge amounts of data
 - Ten largest decision support databases range from 10 to 29 Terabytes
 - Ten largest transaction-processing databases range from 3 to 18 Terabytes
 - Sizes have tripled / doubled between 2001 and end of 2003
 (Source: http://www.wintercorp.com/VLDB/2005_TopTen_Survey/TopTenProgram.html)
- Questions arise:
 - Is there any new, unexpected and potentially useful information contained in this data?
 - Can we use historical data to predict future outcomes (e.g. customer behaviour)

Very short introduction to data mining (2)

- Data mining involves
 - Database and data warehouse technologies
 - Machine learning and artificial intelligence
 - Statistics and numerical mathematics
 - Parallel and high-performance computing
 - Visualisation
- Data mining is being applied in many areas
 - Bioinformatics and health
 - Governments (statistics, census, taxation)
 - Credit card and insurance companies
 - Terror, crime and fraud detection
 - Networking and telecommunications
 - etc.

Very short introduction to data mining (3)

- Data mining techniques and applications
 - Cluster analysis
 - Rule discovery (association rules)
 - Outlier detection
 - Predictive modelling and classification
 - Data preprocessing, cleaning and integration
 - Spatial and temporal mining
 - Text and web mining
 - Sequence mining (e.g. DNA, proteins)
 - Multimedia mining (images, audio, video)

Why should you take this course...?

- Large data collections are widely available in many companies and organisations
(but often only used for transaction processing → Write-only data)
- Good job prospects
(as companies and organisations start to realise the potential wealth of their data)
- Data mining is still a young field
(it is multidisciplinary, and there are many open research issues)

Course overview

- Part 1 (before semester break): Lectures on data, statistics and data mining techniques
- Part 2 (after semester break): Lecture on computational aspects and special topics
- Part 3: Student presentations (weeks 12 and 13)

Details will be made available on the course Web site soon

Course lecturers

- John Maindonald (course coordinator) john.maindonald@anu.edu.au
Centre for Mathematics and its Applications, ANU Mathematical Sciences Institute
- Peter Christen peter.christen@anu.edu.au
Department of Computer Science, ANU Faculty of Engineering and Information Technology (FEIT)
- Stephen Roberts stephen.roberts@anu.edu.au
Department of Mathematics, ANU Mathematical Sciences Institute
- Markus Hegland markus.hegland@anu.edu.au
Centre for Mathematics and its Applications, ANU Mathematical Sciences Institute
- Alan Welsh alan.welsh@anu.edu.au
Statistical Science Program, ANU Mathematical Sciences Institute
- Graham Williams graham.williams@togaware.com
Australian Taxation Office / Togaware / Department of Computer Science, ANU FEIT

Course schedule

- Lecture times (all in John Dedman building, G35)
 - Tuesday 11-12
 - Thursday 10-11
 - Thursday 13-14
- Labs and tutorials: Using **R**
- Lab times and venue: Tuesday 2-4, in the Coombs Computer Laboratory 3005.

Proposed course assessment

- Four assignments, two worth 10% and two worth 15%
- Student paper presentation, worth 20%
(in week 12 and 13)
- Examination (take home), worth 30%
(after week 13, exam period)
- The final mark will be the sum of the assignments, presentation and exam mark

Course resources (1)

- Course web site:
http://datamining.anu.edu.au/student/math3346_2006.html
(online lecture slides, tutorials, contact details, Web resources, etc)
- Data mining links:
<http://datamining.anu.edu.au/links.html>
- Further Web resources will be given in the modules

Course resources (text books)

- **Data Mining: Concepts and Techniques**, 2nd edition
Jiawei Han and Micheline Kamber
Morgan Kaufmann Series in Data Management Systems,
Morgan Kaufmann Publishers, November 2005.
ISBN 1-55860-901-6
- **Data Mining: Practical Machine Learning Tools and Techniques**, 2nd edition
Ian Witten and Eibe Frank
Morgan Kaufmann Series in Data Management Systems,
Morgan Kaufmann Publishers, June 2005.
ISBN 0-12088-407-0