Unit COMP5318
Knowledge Discovery and Data Mining
6 Credit Points
Unit of Study Outline & Assessment Details
First Semester, 2006
IMPORTANT: Policy relating to Academic Dishonesty and Plagiarism.

The School of Information Technologies\(^1\) views all forms of academic dishonesty, including plagiarism and recycling, very seriously.

**Plagiarism** means presenting another person’s ideas, findings or work as one’s own by copying or reproducing them without due acknowledgement of the source.

**Recycling** means the submission for assessment of one’s own work, or of work which is substantially the same, which has previously been counted towards the satisfactory completion of another unit of study, and credited towards a university degree, and where the examiner has not been informed that the student has already received credit for that work.

Students who submit work containing significant portions that have been copied from other sources, including published works, the internet, existing programs, work previously submitted for other awards or assessments, or the work of other students, without proper acknowledgement will be penalised. Decisions as to the penalty may include:

(a) counselling the student;
(b) issuing a written warning;
(c) requiring the student to resubmit the work for assessment; or to undertake other remedial work;
(d) requiring the student to undertake another form of assessment in lieu of the assignment in question, such as an unseen examination;
(e) applying a fail grade to the work, or part thereof, submitted for assessment;
(f) applying a fail grade overall in the unit of study; or
(g) referring the matter to the Registrar if the head of school considers there has been a breach of the University’s standards of academic honesty and the student continues in a denial, or, following the interview, the head of school considers that failing the unit of study is insufficient to deal with the matter.

Where there is doubt about which portions of work are contributed by a particular student he/she may be required to demonstrate knowledge of the relevant material by answering oral questions or by undertaking supplementary work, either written or in the laboratory, in order to arrive at the final assessment mark.

\(^{1}\) Refer to Academic Board policy: [http://policy.rms.usyd.edu.au/000003f.pdf](http://policy.rms.usyd.edu.au/000003f.pdf)
Unit COMP5316 Knowledge Discovery and Data Mining

1. Introduction
This course will offer a comprehensive coverage of well-known data mining topics including classification, clustering, and association rules. A number of specific algorithms and techniques under each category will be discussed. Methods for feature selection, dimensionality reduction and performance evaluation will also be covered. Students will learn and work with appropriate software tools and packages in the laboratory. They will be exposed to relevant research being done in the University of Sydney such as applications in text categorization, information retrieval, multimedia, spatio-temporal databases and bioinformatics.

2. Objectives
Students who successfully complete this unit will be able to:
1. explain the basic principles and understand the strengths, weaknesses and applicability of data mining algorithms for solving classification, clustering and association tasks,
2. gain practical experience in using data mining packages, and also in designing, implementing and evaluating such systems,
3. access relevant data mining research being done at the University of Sydney, and develop interest in the field sufficient to take more advanced studies.

3. Unit of Study Delivery
A variety of learning situations will be employed during the unit of study, including lectures, recommended reading, directed computer laboratory exercises, and assessed assignments. To benefit fully from this unit it is necessary to participate fully in all aspects of the unit of study.

4. Expectations
1. Students are expected to attend all scheduled lectures, and laboratory classes. You should expect to spend a minimum of 12 hours per week including scheduled lectures and laboratory times.
2. Students are expected to undertake recommended reading, to carry out exercises and laboratory tasks and to submit selected work for assessment as directed. It should be realised that some laboratory exercises are expected to take longer than just the time scheduled for classes.
3. Students are expected to be able to work independently and to make effective use of a range of resources including the library, the Internet and relevant on-line help facilities.

5. Assessment Package
The unit of study will be assessed by means of the following components:

<table>
<thead>
<tr>
<th>Component</th>
<th>% of Final Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Programming Assignment</td>
<td>25</td>
</tr>
<tr>
<td>Paper Presentation – application of data mining techniques</td>
<td>15</td>
</tr>
<tr>
<td>Written examination</td>
<td>60</td>
</tr>
</tbody>
</table>

*It is a policy of the School of Information Technologies that in order to pass this unit, a student must achieve at least 40% in the written examination as well as in the other components of assessment together. A student must also achieve an overall scaled final mark of 50 or more. Any student not meeting these requirements can achieve a maximum mark of no more than 45.*

Deadlines for assignments are set on the assumption that students may experience minor setbacks caused by sickness, computer breakdown etc. In this context, ‘minor’ means ‘causing a delay of up to
three working days’. Extensions will not be granted for minor setbacks. It is important to work steadily on assignments as soon as they are given.

Late work: In the interests of fairness to all students, the School of Information Technologies policy states that late work cannot be accepted. In exceptional cases late work must be submitted directly to the unit of study coordinator accompanied by an application for Special Consideration as outlined on page 16 of the School of Information Technologies Postgraduate Enrolment Guide.

Assessment results will be published on the course web page. Students are required to check their results.

Any errors or omissions must be reported to the unit coordinator, with appropriate evidence, within ten (10) days of being published. Ten days after being published, marks are considered to have been confirmed and will not subsequently be altered.

The total marks for COMP5318 are 100 with the breakdown as described above. The total will then be scaled by the School examiners’ meeting, to keep final results comparable between courses, to take account of academic judgement about the appropriate Pass line, and also to adhere to Faculty of Science policy on the number of merit grades awarded. Scaling may lead to students’ marks moving up or down. Scaling will not alter the relative order between two students who are enrolled in the same unit of study.

6. Details of Assessment Components

6.1 Assignments

There will be one programming assignment. Students will be required to design and implement Data Mining methods and evaluate their performance. Students will be allowed to work in pairs but not in groups of more than 2 people. In addition, students will be required to work in pairs and present a research paper describing an application of Data Mining techniques.

6.2 Written Examination

The written examination will cover all aspects of the unit of study. It will test the candidates’ ability to understand and apply the knowledge learnt during the course to specific problems. The duration of the examination will be two hours.

7. Teaching team

<table>
<thead>
<tr>
<th>Name</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irena Koprinska (course coordinator and lecturer)</td>
<td><a href="mailto:irena@it.usyd.edu.au">irena@it.usyd.edu.au</a></td>
</tr>
</tbody>
</table>

8. Textbook and Readings

Textbook:
Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar
Pearson Education (Addison Wesley), 0-321-32136-7, 2006

Other recommended books:
Data mining - practical machine learning tools and techniques with Java implementations
Ian H. Witten, Eibe Frank, Morgan Kaufmann. 1-55860-552-5, 2001

Data Mining: Introductory and Advanced Topics
Margaret Dunham, Prentice Hall, 0-13088892-3, 2003

Data Mining Concepts and Techniques.
J. Han and M. Kamber, Morgan Kaufmann, 2001, 1-55860-489-8

Principles of Data Mining
9. Tentative weekly schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6Mar</td>
<td>Introduction to data mining. Data: types, quality, pre-processing, measures of similarity.</td>
</tr>
<tr>
<td>2</td>
<td>13Mar</td>
<td>Dimensionality reduction. Exploring data.</td>
</tr>
<tr>
<td>4</td>
<td>27Mar</td>
<td>Decision trees. [ass1 out]</td>
</tr>
<tr>
<td>6</td>
<td>10Apr</td>
<td>Neural network-based algorithms.</td>
</tr>
<tr>
<td>7</td>
<td>17Apr</td>
<td>Mid-semester break</td>
</tr>
<tr>
<td>8</td>
<td>24Apr</td>
<td>Support vector machines. [ass1 due]</td>
</tr>
<tr>
<td>9</td>
<td>1May</td>
<td>Clustering.</td>
</tr>
<tr>
<td>10</td>
<td>8May</td>
<td>Association rule mining</td>
</tr>
<tr>
<td>11</td>
<td>15May</td>
<td>Sequential pattern analysis</td>
</tr>
<tr>
<td>12</td>
<td>22May</td>
<td>Guest lecture</td>
</tr>
<tr>
<td>13</td>
<td>29May</td>
<td>[ass2: Student presentations]</td>
</tr>
<tr>
<td>13</td>
<td>5June</td>
<td>[ass2: Student presentations]</td>
</tr>
</tbody>
</table>

10. Course home web page