MATH5836 – Course Outline

Information about the course

Course Authority and Lecturer: Dr. Peter Straka, email p.straka@unsw.edu.au

Consultation: Please use email to arrange an appointment.

Credit: This course counts for 6 Units of Credit (6UOC).

Prerequisites: It is expected that students in this course have a solid understanding of three years worth of probability and statistics at an undergraduate level. In addition, it will be assumed that students have had first experiences with statistical software packages such as R or Matlab.

Tutorials: There will be no formal tutorials for this course, though there will be assignments which will require students to undertake computational laboratory work. In addition, students will be given exercises throughout semester to consider with solutions and example code, as well as the opportunity to present their solutions to fellow students during the course.

moodle: Further information, lecture notes, and other material will be provided via moodle.

Lectures: Fridays 5:30pm -8:00pm in RC-3085

Course aims

This course will aim to provide the student with a working knowledge of data mining and machine learning techniques. Particular focus will be on the fundamental statistical properties and analysis of many popular techniques for learning, classification and prediction. The topics covered in this course will include elements of the following: Probability Theory, Model Selection, Curse of Dimensionality and Information Theory; Supervised Learning and Statistical Decision Theory; Linear Models – Regression and Classification; Kernel Methods and Kernel Machines; Graphical Models; Approximate Inference and Variational Approaches; Sampling Methods.

Relation to other statistics courses

This course is related to Multivariate Analysis - MATH5855; Statistical Computations - MATH5856; Machine Learning and Data Mining - COMP9417.

Student Learning Outcomes

This course is expected to give students an understanding of the fundamentals of machine learning and basics of data mining, which is essential for anyone contemplating a career as a professional statistician or data analyst in industries reliant upon such expertise. The student should develop a working knowledge of the statistical and theoretical underpinnings of the topics covered. Given this fundamental statistical understanding of these methodologies this will allow the student to utilise these techniques with confidence on real world data sets and scenarios. As such the student is expected to develop applied working knowledge of the methodologies covered, largely through application in assignments during semester. It is stressed that this course is aimed at fundamental statistical properties of these methods, it is not a course on application of computer software.
Relation to graduate attributes

The problem-solving activities in assignments will improve your research, inquiry and analytical thinking abilities (Science Graduate Attribute 1) and your capacity and motivation for intellectual development (Science Graduate Attribute 2); Coursework assignments will provide you with timely feedback on your progress and improve your Communication skills (Science Graduate Attribute 4); Computing skills developed in this course will improve your Information Literacy (Science Graduate Attribute 6).

Teaching strategies underpinning the course

New ideas and skills are introduced and demonstrated in lectures and through recommended reading of supplementary material such as research papers, then students develop these skills by applying them to specific tasks in assessments.

Rationale for learning and teaching strategies

We believe that effective learning is best supported by a climate of inquiry, in which students are actively engaged in the learning process. Hence this course is structured with a strong emphasis on problem-solving tasks. Students are expected to devote the majority of their class and study time to the solving of such tasks. New ideas and skills are first introduced and demonstrated in lectures, and then students develop these skills by applying them to specific tasks in assessments. Computing skills are developed and practiced in assignment working groups. This course has a major focus on research, inquiry and analytical thinking as well as information literacy. We will also explore capacity and motivation for intellectual development through the solution of both simple and complex mathematical models of problems arising in the quantitative sciences and the interpretation and communication of the results.

Assessment

Assessment in this course will consist of a final exam, a mid-session test and two group assignments (maximum of 4 members per group). See table below.

<table>
<thead>
<tr>
<th>Assessment</th>
<th>Weight</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Final Exam</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td>Midsession Test</td>
<td>15%</td>
<td>Within class</td>
</tr>
<tr>
<td>Assignment 1</td>
<td>10%</td>
<td>week 5</td>
</tr>
<tr>
<td>Assignment 2</td>
<td>10%</td>
<td>week 11</td>
</tr>
<tr>
<td>Problem Presentations</td>
<td>5 %</td>
<td>Throughout semester</td>
</tr>
</tbody>
</table>

Knowledge and abilities assessed: All assessment tasks will assess the learning outcomes outlined above, specifically, the ability to derive logical and coherent proofs of relevant results, and the ability to solve a variety of problems, both theoretical and in practice. In addition, through group assignments the students form practical time and relationship management skills. A single report will be required per group.

Assessment criteria: The main criteria for marking all assessment tasks will be clear and logical presentation of correct solutions.
Examination

**Duration:** Three hours

**Rationale:** The final examination will assess student mastery of the material covered in the lectures and assignments.

Midsession Test

**Rationale:** This one hour test will give students feedback on their progress and mastery of the material.

Assignments

**Rationale:** Assignments will give an opportunity for students to try their hand at more difficult problems requiring more than one line of argument and also introduce them to aspects of the subject which are not explicitly covered in lectures. Assignments are also intended to give regular feedback on a students progress and mastery of the material, to identify as soon as possible any problems that students may have. Assessment in this course will use problem-solving tasks of a similar form to those discussed in lectures and found in recommended additional reading such as research papers, to encourage the development of the core skills underpinning this course and the development of analytical thinking Assignments may be completed in groups of up to four members, they must be each groups OWN WORK, or severe penalties will be incurred. Assignments will only be accepted with a signed plagiarism cover sheet (provided by each member of the group). It is the responsibility of the students to organise the group structures. Late assignments will not be accepted.

You should consult the University web page on plagiarism: [www.lc.unsw.edu.au/plagiarism](http://www.lc.unsw.edu.au/plagiarism)

Problem Presentations

**Rationale:** Students will be given the opportunity to present solutions to homework problems during the course. This shall assess the skill to orally communicate ideas in statistics and data mining.

Additional resources and support

**Computer laboratories**

Computer laboratories (RC-M020 and RC-G012) are open 9:00am – 5:00pm Monday – Friday on teaching days. RC-M020 has extended teaching hours (usually 8:30am – 9:00pm Monday – Friday, and 9:00am – 5:00pm Monday-Friday on non-teaching weeks).

**Lecture notes**

All lecture slides will be provided on moodle. These notes are *insufficient* to understand the course material, therefore there will be provided detailed repositories of additional recommended texts. This is expected to also be read to supplement the material covered in lectures.
Textbooks

The course is based on two recommended text books:


These books as well as other material are freely available on the internet.

Course Evaluation and Development

The School of Mathematics and Statistics evaluates each course each time it is run. We carefully consider the student responses and their implications for course development. It is common practice to discuss informally with students how the course and their mastery of it are progressing.

Administrative matters


Plagiarism and academic honesty: Plagiarism is the presentation of the thoughts or work of another as ones own. Issues you must be aware of regarding plagiarism and the universitys policies on academic honesty and plagiarism can be found at http://www.lc.unsw.edu.au/plagiarism

Tentative Course Schedule

Week 01 Overview; Bayesian vs. Frequentist Reasoning
Week 02 Linear Models: Regression, Shrinkage, Classification
Week 03 Basis Expansions; Kernel Smoothing; Bayesian model selection
Week 04 Nearest Neighbour Classification mixture models; Clustering
Week 05 Belief Networks; Graphical Models
Week 06 Tree-Based Methods
Week 07 Dimension Reduction; Latent Variables
Week 08 Neural Networks
Week 09 Support Vector Machines
Week 10 Association Rules
Week 11 Markov Chain Monte Carlo
Week 12 Gaussian Processes in Machine Learning (subject to time)