SCHOOL OF MATHEMATICS AND STATISTICS

MATH5836 DATA MINING

Term 3, 2021

Course Instructor: Dr Rohitash Chandra, Office: RC-4110
Email: rohitash.chandra@unsw.edu.au

2021 Units of Credit: 6, Complementary Courses: MATH5855; MATH5856; COMP9417

Prerequisites: A prerequisite for this course is MATH2801/2901. A recommended prerequisite is MATH2831/2931 (Higher Linear Models). If you have not done the recommended prerequisite courses, make sure that you know sufficient statistical theory, can integrate and differentiate, and have competency in at least a programming language such as R or Python.

Delivery Mode: The course features 4 hours of lectures per week with blended tutorial sessions that focus on practical (programming) examples. Tuesday 5-7 pm, Wednesday 7-9 pm, Trimester 3, 2021. The lectures will be given online, and a recording will also be provided.

Administrative Contacts: Please visit the School of Mathematics and Statistics website for a range of information on School Policies, Forms and Help for Students.

For information on Courses, please go to "Current Students" and either Undergraduate and/or Postgraduate", Course Homepage" for information on all course offerings,

The "Student Notice Board" can be located by going to the "Current Students" page; Notices are posted regularly for your information here. Please familiarise yourself with the information found in these locations. The school web page is: https://www.maths.unsw.edu.au/

If you cannot find the answer to your queries on the web you are welcome to contact the Student Services Office directly.

By email: Postgraduate   pg.mathsstats@unsw.edu.au
By phone: 9385 7053

Should we need to contact you, we will use your official UNSW email address of in the first instance. It is your responsibility to regularly check your university email account. Please state your student number in all emails.

DESCRIPTION

Increasingly, organisations need to analyse enormous data sets to determine useful structures in them. In response to this, a range of statistical and machine learning methods have been developed in recent times. This course covers the key techniques in data mining and machine learning with theoretical background and applications. The topics include methods such as linear and logistic regression, neural networks, Bayesian neural networks, clustering and dimensionality reduction, ensemble learning, and an introduction to deep learning. Emerging machine learning tools and libraries are used to illustrate the methods in programming environments that includes Python and R.

RATIONALE

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

AIM

This course is expected to give students an understanding of the fundamentals of machine learning and the 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 practical applications. In addition, students will undertake additional reading of a collection of associated research papers in each topic, to further add context to the methodologies presented during the course. This will enhance the student's ability to utilise these techniques to solve real-world problems. It is stressed that this course is aimed at fundamental statistical properties of these methods, it is not a course on the application of computer software.

ASSESSMENT

1. **Assessment 1** (5%, Week 2: Duration of 24 hours)
2. **Assessment 2** (20%, Week 5: Duration of 24 hours)
3. **Assessment 3 - Project:** (25%, Deadline: Friday, 10 pm, Week 10)
4. **Assessment 4 - Final Exam** (50%, Duration of 3 hours)

Note that Python is the designated language for the course and minimum support on R will be given. You can submit the assessments in either R or Python. You need to get a 40% minimum in the final exam to pass the course; i.e if the final exam has a total of 100 marks, you will need to get at least 40/100 to pass the course.

SCHEDULE

1. Week 0 Python and R Tutorials (no Lectures)
2. Week 1 Data Processing and Introduction to Data mining
3. Week 2 Logistic Regression and Evaluation
4. Week 3 Intro to Neural Networks
5. Week 4 Advances in Neural Networks
6. Week 5 Bayesian Neural Networks
7. Week 6 Break (no Lectures)
8. Week 7 Trees and Forests
9. Week 8 Ensemble Learning
10. Week 9 Unsupervised Learning and Dimensionality Reduction
11. Week 10 Emerging Topics in Data Mining

LEARNING OUTCOMES

1. Demonstrate an understanding of the fundamentals of machine learning and basics of data mining.
2. Demonstrate a working knowledge of the statistical and theoretical underpinnings of the methods.
3. Demonstrate an applied working knowledge of the methodologies covered with practical assignments.
4. Develop models for solving data mining problems that include clustering, regression, and classification.
5. Build models and apply them to real-world data sets and use evaluation metrics to compare their performance.

LATE SUBMISSION

Assessment 1, 2 and 4: No late submission accepted (apply for special consideration in special cases)
Assessment 3: A late penalty of 10% of the awarded mark will be applied per day. Any assessment task submitted 5 or more days late will be given zero.

RECOMMENDED BOOK

1. Géron. A, 2019, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems, O'Reilly, second edition: https://www.bookshop.unsw.edu.au/details.cgi?ITEMNO=9781492032649 (a copy would be handy but not required)

EXTRA READING MATERIALS (optional)

2. Mitchell. Tom, 1997, Machine Learning, McGraw-Hill. (additional textbook for reference): https://www.amazon.com.au/Machine-Learning-Thomas-Mitchell/dp/0070428077
3. Kroese, Botev, Tamire & Vaisman (2020), Data Science and Machine Learning, Chapman & Hall: https://www.bookshop.unsw.edu.au/details.cgi?ITEMNO=9781138492530

SPECIAL CONSIDERATION

You can apply for special consideration if illness or other circumstances beyond your control interfere with your assessment performance, to get an extra opportunity to demonstrate your level of performance.

You must make your application online, through the Special Consideration portal on myUNSW. Do not apply to your course teaching staff - they will be notified automatically.

You can read more about special consideration at: https://student.unsw.edu.au/special-consideration.

ACADEMIC INTEGRITY

For further information about academic integrity and plagiarism at UNSW go to: https://student.unsw.edu.au/plagiarism

For information about acknowledging your sources and referencing go to: https://student.unsw.edu.au/referencing. If you are not sure what referencing style to use in this course, you should ask your Lecturer.