Subject Outline

Subject Name	Data Mining and Machine Learning
Subject Code	MA5832
Study Period	SP83-2020
Study Mode	External
Campus	JCU Online
Subject Coordinator	Dr Kelly Trinh

We acknowledge the Traditional Owners of the lands and waters where our University is located and actively seek to contribute and support the JCU Reconciliation Statement, which exemplifies respect for Australian Aboriginal and Torres Strait cultures, heritage, knowledge and the valuing of justice and equity for all Australians.

Pre-requisites

MA5810 and 24 CP of Postgraduate subjects

Staff contact details

Teaching team	Staff member	Email
Subject Coordinator	Dr Kelly Trinh	kelly.trinh@jcu.edu.au
Tutor	Hongbin Liu	hongbin.liu@my.jcu.edu.au

Contents

1 Subject at a glance

1.1 Student participation requirements

The JCU (4.3) indicates that, "a **3 credit point subject** will require a **130 hour work load** of study-related participation including class attendance over the duration of the study period, **irrespective of mode of delivery"**. This work load comprises **timetabled hours** and **other attendance requirements**, as well as **personal study hours,** including completion of online learning activities and assessment requirements. Note that "attendance at specified classes will be a mandatory requirement for satisfactory completion of some subjects" (Learning, Teaching and Assessment Policy, 5.10); and that additional hours may be required per week for those students in need of **English language, numeracy** or **other learning support.**

Other mandatory attendance requirements:

Students are expected to participate in the Learn Ultra discussion boards. Discussion boards give you a place to interact with staff and other students about subject content and topics, and help students to clarify and extend their understanding of key content. These are a forum for students to present their thoughts/ideas in an online version of an in-person classroom discussion and therefore the same courtesy rules apply.

While attendance in the Collaborate sessions is not mandatory, it is highly recommended. These sessions will provide you with the opportunity to have synchronous (at the same time) conversations with your Subject Coordinator (or your tutor) and with your fellow students from across the subject, to have your questions answered and to receive further clarification about any concerns or questions you may have.

1.2 Key dates

Key dates	**Date**
O week	Starts on Monday 27 April, 2020
Census date	See2020 Study Period and Census Dates
Last date to withdraw without academic penalty	See2020 Study Period and Census Dates
Assessment 1 (20%)	Open on Monday (27/04/2020). Due on Sunday, Week 2 (17/5/2020, AEST 11:59pm).
Assessment 2 (35%)	Open on Monday (4/05/2020). Due on Sunday, Week 4 (31/05/2020, AEST 11:59pm).
Assessment 3 (45%)	Open on Monday (18/05/2020). Due on Wednesday, Week 7 (17/06/2020, AEST 11:59pm).

2 Subject details

2.1 Subject description

This subject will provide students with a range of algorithms based on machine learning techniques for advanced data analysis and mining. These algorithms and techniques fall within the most common machine learning paradigms. In particular, students will learn sophisticated supervised learning methods.

2.2 Subject learning outcomes

Students who successfully complete this subject will be able to:

1. understand roles of machine learning in the realm of data mining to diverse of audiences;
2. compare and different machine learning methods;
3. analyse real world tasks using machine learning techniques learnt in this subject, in particular
   a. describe, choose, and apply appropriate supervised machine learning methods for descriptive data mining tasks
4. synthesise and communicate the method and findings to diverse audiences.

These outcomes will contribute to your overall achievement of **course learning outcomes.** Your course learning outcomes can be located in the entry for your course in the electronic Course and Subject Handbook 2019 (click on 'Course Information' bar/ select 'Undergraduate Courses' or 'Postgraduate Courses'/ select relevant course/ scroll down to 'Academic Requirements for Course Completion', 'Course learning outcomes').

2.3 Learning and teaching in this subject

Week 1: The introductory week begins with the revision of some content on Data mining that have been covered in the pre-requisite subject. This is followed by problem based motivation and heuristic introduction to the four supervised learning algorithms tree-based method, support vector machines, neural networks and deep learning.

Week 2 is devoted to foundational mathematics, probability theory and optimisation methods commonly used in supervised machines learning. This is to provide students theories underpinning the supervised learning techniques covered in this subject. We will revisit some of topics covered in MA5801 and MA5820, as well as learning method theories.

In Week 3, students will be introduced to the well-known tree-based methods. The topics include classification and regression trees (CARTS), PRIM-Bump hunting, multivariate adaptive regression splines (MARS), boosted trees and random forest (RF).

Week 4: Another successful classifier support vector machines (SVM) is introduced. We make a distinction between linear and non-linear SVMs. Students learn the implementation of hard and soft margin SVM using R.

Week 5: Students are introduced to the basic neural networks as an alternative to deterministic, hyperplane-based classifiers. Students would implement these using R.

Week 6 Deep learning: Students are introduced to the fundamental concept of deep learning, understand key factors behind the popularity of deep learning, methods to train deep learning and use the Keras package in R to solved basic classification and regression problem.

Week 7 is kept free to work on your capstone project.

Week 1: Introduction

Topic 1: Learning algorithms -the story so far and some principles.
Topic 2: Supervised learning
Topic 3: Supervised learning for tabular data

Topic 4: Supervised learning for Artificial Neural Networks

Learning Outcomes:

- Understand the roles of supervised learning within the broader ambit of data science.
- Link real problems that can be solved by application of supervised learning algorithms.
- Appreciate the utility, distinction and challenges around the two alternative supervised learning paradigms: deterministic decision rules vs recursive probabilistic algorithms.

Week 2: Essential mathematics for data-mining

2.1 Linear Algebra
2.2 Probability and Distributions
2.3 Optimisation

Learning Outcomes:

- understand the roles of linear algebra, probability theory and optimisation in the realm of machines learning
- conceptually understand the fundamental probability axioms and rules of discrete and continuous random variable
- understand and apply appropriate probability distributions to descriptive problems
- understand algorithm underpinning various optimisation methods
- apply and implement concepts in linear algebra, probability theory and optimisation using R

Week 3: Tree-based methods

3.1. Classification and Regression Trees (CARTs)
3.2. PRIM-Bump
3.3. Multivariate adaptive regression splines (MARS)
3.4. Bagging
3.5. Random forest
3.6. Boosted trees

Learning Outcomes:

- Comprehend the use of tree-based methods in the realm of data science
- Conceptually understand algorithms underpinning variety of tree-based methods
- Differentiate and appraise a variety of tree-based methods
- Identify appropriate tree-based method for descriptive problems
- Apply the tree-based methods covered in Week 3 to real datasets using the computer language R and the software environment RStudio

Week 4: Support Vector Machines

4.1. Classification: a method of drawing hyperplanes.
4.2. Linear support vector machines: Maximum margin classifiers.
4.3. Soft-margin classifiers.
4.4. Limitations of SVM

Learning Outcomes:

- Identify and translate a data science problem into a classification problem.
- Develop a conceptual and mathematical understanding of conventional support vector classifiers.
- Implement two class and multiclass support vector classification algorithm in R.
- Understand the benefits and limitations of support vector machines against comparators.

Week 5: Neural Network

5.1 Introduction – emulation of nature
5.2 Vanilla Neural Network.
5.3 Fitting Model Parameters.
5.4 Issues surrounding ANN.
5.5 Exercises.

Learning Outcomes:

- Conceptually describe and apply the vanilla neural network.
- Use R to fit neural networks.
- Understand the limitations surrounding training neural networks
- Translate a data science problem into an application of neural networks.

Week 6: Deep Learning

6.1. The deep learning paradigm.
6.2. Focus on DNN training.
6.3. Introduction to Keras deep-learning library.
6.4. Classification neural network.
6.5. Regression neural network.

Learning Outcomes:

- Learn high-level definitions of fundamental concepts of deep learning.
- Understand the key factors behind deep learning’s rising popularity and future potential.
- Learn how to train and test deep learning networks.
- Understand how deep learning networks are trained via backpropagation and gradient descent.
- Learn how to use Keras to solve basic classification and regression problems.

2.4 Student feedback on subject and teaching

As part of our commitment at JCU to improving the quality of our courses and teaching, we regularly seek feedback on your learning experiences. Student feedback informs evaluation of subject and teaching strengths and areas that may need refinement or change. ***YourJCU Subject and Teaching Surveys*** provide a formal and confidential method for you to provide feedback about your subjects and the staff members teaching within them. You will receive an email invitation when the survey opens. We value your feedback and ask that you provide constructive feedback about your learning experiences for each of your subjects, in accordance with responsibilities outlined in the Student Charter. Refrain from providing personal feedback on topics that do not affect your learning experiences. Malicious comments about staff are deemed unacceptable by the University.

2.5 Subject resources and special requirements

All subject readings and resources, including journal articles, book chapters, websites, videos, print and e-Textbooks, are available to view online from your *Readings list* via your LearnJCU subject site.

The following textbooks are required for this subject and are either free online or available electronically through the library.

- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York, NY: Springer. (freely available online at)
- Trevor, H., Robert, T., & JH, F. (2009). *The elements of statistical learning: data mining, inference, and prediction.* New York, NY: Springer. (freely available online at).
- Chollet, F., & Allaire, J. J. (2018). Deep Learning with R. Greenwich, CT, USA:Manning Publication Co.

The mandatory software for the subject is R. Although R alone is sufficient, students may want to install R-studio for the user friendly interface. Both software are open source and R-studio is available via https://www.rstudio.com/products/rstudio/download/#download

3 Assessment details

3.1 Requirements for successful completion of subject

In order to pass this subject, you must:

- Achieve an overall percentage of 50% or more;
- Achieve a percentage of 50% or more in the capstone project

Assessment items and final grades will be reviewed through moderation processes (Learning, Teaching and Assessment Policy, 5.13-5.18). It is important to be aware that assessment "is always subject to final ratification following the examination period and that no single result represents a final grade in a subject" (Learning, Teaching and Assessment Policy, 5.22.). The final grades will comply with the "Student Results Policy" (Learning, Teaching and Assessment Policy, 5.23).

3.1.1 Inherent requirements

Inherent requirements are the fundamental abilities, attributes, skills and behaviours needed to achieve the learning outcomes of a course while preserving the academic integrity of the university's learning, assessment and accreditation processes. Students and prospective students must be able to demonstrate that they have acquired or have the ability to acquire the inherent requirements for their degree.

Reasonable adjustments may be made to assist students manage additional circumstances impacting on their studies provided these do not change the academic integrity of a degree. Reasonable adjustments do not alter the need to be able to demonstrate the inherent requirements of the course. Students who believe they will experience challenges completing their degree or course because of their disability, health condition or other reason should discuss their concerns with an Access Ability Services team member or a member of College staff, such as the Course Coordinator. In the case where it is determined that inherent requirements cannot be met with reasonable adjustments, the University staff can provide guidance regarding other study options.

3.2 Feedback on student learning

Feedback for students will be provided on all assessment items.

3.3 Assessment tasks

ASSESSMENT 1

Aligned subject learning outcomes	It addresses the following learning outcome(s): • understanding the roles of linear algebra, probability theory and optimisation in the realm of machines learning; • understanding algorithms underpinning various optimisation methods; • applying and implementing concepts in linear algebra, probability theory and optimization using R.
Group or individual	Individual
Weighting	20%
Due date	Due on Sunday, Week 2 (17/5/2020, AEST 11:59pm).

ASSESSMENT TASK 1: DESCRIPTION

This assessment aims to assess a student's understanding on the topics covered in week 2.

ASSESSMENT TASK 1: CRITERIA SHEET

Refer to the rubric on the subject’s website.

ASSESSMENT 2

Aligned subject learning outcomes	The assessment addresses the following learning outcome(s): • developing a conceptual and mathematical understanding of conventional support vector classifiers; • identifying and translating a data science problem into a supervised learning problem; • identifying appropriate tree-based methods, and support vector classification for descriptive problems; • application of support vector classifier and tree-based methods covered in Week 3 and 4 to a dataset.

Group or individual	Individual
Weighting	35%
Due date	Due on Sunday, Week 4 (31/5/2020, AEST 11:59pm).

ASSESSMENT TASK 2: DESCRIPTION

In this assessment, you will implement and compare two machine learning algorithms learnt from Week 3 and Week 4 on a real data. In addition, you will solve some analytical questions to develop a conceptual and mathematical understanding of conventional support vector classifiers.

ASSESSMENT TASK 2: CRITERIA SHEET

Refer to the rubric on the subject's website.

ASSESSMENT 3

Aligned subject learning outcomes	The purpose of the assignment is to enable you to: • apply machine learning methodologies; • undertake independent research investigating machine learning parameters; • compare and contrast machine learning methodologies; • construct a written communication and interpretation of findings resulting from machine learning methodologies.
Group or individual	Individual assessment task
Weighting	45%
Due date	Due on Wednesday, Week 7 (17/06/2020, AEST 11:59pm).

ASSESSMENT TASK 3: DESCRIPTION

During this assessment you will produce a written report on analyses of a real-world problem using a neural network and an alternative machine learning method that have been presented in this course. For both techniques, you will investigate the properties of the machine learning methodologies by altering the machine learning parameters.

Additionally, you will compare and contrast the two machine learning methodologies using the same data source.

ASSESSMENT TASK 3: CRITERIA SHEET

See assessment document for detail and rubrics.

4 Submission and return of assessment

4.1 Submission of assessment

All assessments are submitted through Learn Ultra.

Note that the Learning, Teaching and Assessment Policy (5.22.3) outlines a uniform formula of penalties that will be imposed for submission of an assessment task after the due date. This formula is 5% of the total possible marks for the assessment item per day including part-days, weekends and public holidays. Due to the dynamic nature and pace of the online delivery mode, **no submission will be accepted after 1 week**. In other words, after 1 week, any assessment item would be awarded 0 marks (100% penalty).

4.2 Return of assessment

Feedback on marked assessments will be available in the Gradebook in Learn Ultra.

Please see for other important student information pertaining to plagiarism and referencing, examinations advice and student support services.

5 Subject calendar

Week/Date/Module		Topics Covered
O	Orientation	• Getting Started
1	Introduction	• Learning algorithms – the story so far • Supervised learning • Supervised learning for tabular data • Supervised learning for Artificial Neural Networks
2	Essential mathematics for data mining	• Linear Algebra • Probability and Distributions • Optimisation
3	Classification and Regression Trees	• Classification and Regression Trees (CARTs) • PRIM-Bump • Multivariate adaptive regression splines (MARS) • Bagging • Random forest • Boosted trees
4	Support Vector Machines	• Classification: a method of drawing hyperplanes. • Linear support vector machines: Maximum margin classifiers. • Soft-margin classifiers. • Limitations of SVM
5	Artificial Neural Networks	• Introduction – emulation of nature

		• Vanilla Neural Network. • Fitting Model Parameters. • Issues surrounding ANN. • Exercises.
6	Deep learning	• High-level concepts of deep learning. • Mathematics of deep learning. • Introduction to Keras deep-learning library.