
Faculty of Medicine
School of Medical Sciences

# HDAT9500
# Machine Learning and Data Mining

COURSE OUTLINE

Term 3 2019

16 September to 29 November

CONTENTS PAGE

**Machine Learning and Data Mining: Course Information**

Healthcare organisations have a vast amount of data: electronic medical records, claims, registries, medical databases, and other bulk digital health data. Data mining analytics and machine learning are powerful tools used to find high-level order, structure and meaning within this huge quantity of information.

Data mining can be defined as the discipline that takes methods from statistics, computer science and database analytics and applies them to large data sets in order to extract useful information. Machine learning is an approach to data mining which itself comprises many powerful techniques and algorithms. These algorithms learn from previous experience to discover patterns and relationships in data, and have been found to perform extremely well in large datasets.

## OBJECTIVES OF THE COURSE

This course provides an introduction to data mining and machine learning techniques through a series of health applications.

Algorithms for supervised and unsupervised learning are covered, including linear regression and classification, tree-based methods, kernel methods, clustering, dimensionality reduction, ensemble methods and neural networks.

Students will learn about the underlying supporting theory of these techniques, as well as gain the practical know-how required to effectively apply these techniques to new health data problems.

## COURSE CO-ORDINATOR and LECTURERS

Course Coordinator:
**Dr Oscar Perez-Concha**
**E: o.perezconcha@unsw.edu.au**
Centre for Big Data Research in Health
Level 4, Lowy Building (C25)
UNSW Sydney
T: +61 (2) 9385 1403

Lecturers in this course:

| Name | Chapters | Email |
|---|---|---|
| **Dr Oscar Perez-Concha** | 1 to 6, 8 & 10 | o.perezconcha@unsw.edu.au |
| **Dr Sebastiano Barbieri** | 7 | s.barbieri@unsw.edu.au |
| **Professor Massimo Piccardi** | 9 | massimo.piccardi@uts.edu.au |
| **Mr Inigo Jauregi** | 9 | inigo.jauregiunanue@student.uts.edu.au |

## COURSE STRUCTURE and TEACHING STRATEGIES

This is a blended learning course comprising of 10 chapters.
Each chapter consists of approximately 10 hours of learning activities distributed as follows:

1. Readings prior to Face to Face - 3.5 hours
2. Face to Face sessions/Workshop: 3 hours
    a. Workshop key concepts and ideas, further reading and slides.
    b. Guided Exercises
3. Web-based online assessment: Weekly exercise - 3.5 hours

Students are expected to attend all face to face sessions for their full duration (3 hours of per chapter): Tuesday 1pm-4pm in Mathews Building Room 105 (F23 Kensington campus).

The course will be hosted online on OpenLearning: www.openlearning.com

## COURSE CONTENT AND TIMETABLE

| Week | Chapter topic | Chapter Dates | Assignment Deadline |
|---|---|---|---|
| 1 | Chapter 1: Introduction | 14 Sep – 20 Sep | 20/Sep/2019 17:00 |
| 2 | Chapter 2: Linear Prediction | 21 Sep – 27 Sep | 27/Sep/2019 17:00 |
| 3 | Chapter 3: Model Evaluation & Improvement | 28 Sep – 4 Oct | 4/Oct/2019 17:00 |
| 4 | Chapter 4: Tree-Based Methods | 5 Oct – 11 Oct | 11/Oct/2019 17:00 |
| 5 | Chapter 5: Support Vector Machines | 12 Oct – 18 Oct | 18/Oct/2019 17:00 |
| 6 | Chapter 6: Artificial Neural Networks | 19 Oct – 25 Oct | 25/Oct/2019 17:00 |
| 7 | Chapter 7: Unsupervised Learning | 26 Oct – 1 Nov | 1/Nov/2019 17:00 |
| 8 | Chapter 8: Sequential Data Models | 2 Nov – 8 Nov | 8/Nov/2019 17:00 |
| 9 | Chapter 9: Natural Language Processing | 9 Nov – 15 Nov | 15/Nov/2019 17:00 |
| 10 | Chapter 10: Recap Session | 16 Nov – 24 Nov | 29/Nov/2019 17:00 |

## COURSE MATERIALS AND TEXTBOOKS

1. Lecture notes provided. Notes will be posted periodically.

2. The following core books are recommended:

   **Book 1**: Introduction to machine learning with Python : a guide for data scientists by Andreas C. Müller and Sarah Guido. Sebastopol, CA, O'Reilly Media, Inc, 2017, 9781449369903, ISBN

   - E-book: http://shop.oreilly.com/product/0636920030515.do
   - Slides: http://www.cs.columbia.edu/~amueller/comsw4995s18/schedule/
   - Example Python code: https://github.com/amueller/introduction_to_ml_with_python

   **Book 2**: An Introduction to Statistical Learning with Applications in R by by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.

   - Free to download: http://www-bcf.usc.edu/~gareth/ISL/
   - Slides and videos: https://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/

   **Book 3:** Deep Learning with Python by François Chollet.

   - Three first chapters are free: https://www.manning.com/books/deep-learning-with-python
   - Very good introduction to Machine Learning and Deep Learning

3. Other books:

   **Book 4:** Pattern Recogntion and Machine Learning by Christopher M. Bishop. Springer. 2007.

   - Free online: https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/

   **Book 5:** The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Trevor Hastie, Robert Tibshirani, Jerome Friedman.

   - Free to download: https://web.stanford.edu/~hastie/ElemStatLearn/

## PROGRAMMING LANGUAGE AND SOFTWARE TOOLS

1. Python: Students are expected to have an intermediate level of Python.
2. Python Jupyter Notebooks (Preferably installed via Anaconda: https://anaconda.org/)
3. www.openlearning.com: OpenLearning will be the primary resource for obtaining course materials: lecture notes, Jupyter Notebooks, videos and links to relevant sites. In addition, OpenLearning will be the online platform to discuss ideas, ask questions and participate in forums.
4. Git and GitHub (https://github.com/). **Students need to have a GitHub account, which has free to sign up,** in order to download the guided Jupyter notebooks and assignments.

## COURSE LEARNING OUTCOMES

1. Distinguish a range of task specific machine learning techniques appropriate for Health Data Science.
2. Design machine learning tasks for Health Data Science scenarios.
3. Construct appropriate training and test sets for health research data.
4. Generate knowledge via the application of machine learning techniques to health data.
5. Appraise methods of training error rate optimisation.

## COURSE EVALUATION AND DEVELOPMENT

For course evaluation, feedback will be gathered at the completion of the course using an online student survey. Student feedback is taken seriously, and continual improvements will be made to the course based, in part, on such feedback.

**ASSESSMENT PROCEDURES: 40% ON HOMEWORKS + 60% FINAL PROJECT**

1. **Chapter assignments** (chapter 2 to chapter 9): 5% each (8 assignments x 5% per assignment): 40% of the final score.
    a. The assignments will assess the chapter content. The assignments will consist of Jupyter Notebooks with several programming tasks.
    b. These assignments will be based on the corresponding chapter readings and previous practice assignment(s).
    c. Each assignment will contribute 5% of their final mark. A rubric is provided (see rubric in the next page).
    d. A penalty will apply for late submissions: the mark will be 0 if not submitted on time.

2. **Project assignment:** Machine learning and data mining algorithms applied to a health question: 60% of the final score.
    a. This assignment targets the pedagogy of 'assessment for learning'.
    b. This project assignment will consist of a health-based scenario, a given health dataset and a health question. In order to find a solution to the question, students will have to apply machine learning and data mining techniques.
    c. Students will provide a written report that will include a description of the methods applied, program developed to solve the health question, the output results and an interpretation of the findings. The written report will be assessed by a rubric (see rubric in the next page).
    d. 20% of the total value of that assignment will be deducted for every day late. The assignment will not be marked if submitted more than 5 days after the due date and will receive a value of 0. For example, if you submit your assignment 2 days late, then 40% (20% x 2 days) will be deducted from the mark.

**A note on the Student Code of Conduct and Honour**: We encourage students to help each other, form study groups and have open discussions.

However, each student must provide the solutions for the chapter assignments and project assignment **independently.** In other words, **each student must understand the solution provided and be able to reproduce it independently**.

In addition, students must write the group of people with whom they collaborated and include an anti-plagiarism declaration that the work submitted is original and the student's own work.

## ASSESSMENT TASKS

| Week | Task (% of total course mark) | Due Date and Time |
|---|---|---|
| **1** | Assignment 1: Introduction (0%) | **Formative Assessment** 20/Sept/2019 17:00 |
| **2** | Assignment 2: Linear Prediction (5%) | 27/Sept/2019 17:00 |
| **3** | Assignment 3: Model Evaluation & Improvement (5%) | 4/Oct/2019 17:00 |
| **4** | Assignment 4: Tree-Based Methods (5%) | 11/Oct/2019 17:00 |
| **5** | Assignment 5: Support Vector Machines (5%) | 18/Oct /2019 17:00 |
| **6** | Assignment 6: Artificial Neural Networks (5%) | 25/Oct/2019 17:00 |
| **7** | Assignment 7: Unsupervised Learning (5%) | 1/Nov/2019 17:00 |
| **8** | Assignment 8: Sequential Data Models (5%) | 8/Nov/2019 17:00 |
| **9** | Assignment 9: Natural Language Processing (5%) | 15/Nov/2019 17:00 |
| **10** | Project Assignment (60%) | 29/Nov/2019 17:00 |

## GRADING RUBRIC FOR ASSIGNMENTS

| **Criteria:** | **High Distinction 85-100 Marks** | **Distinction 75-84 Marks** | **Credit 65-74 Marks** | **Pass 50-64 Marks** | **Fail 0-50 Marks** |
|---|---|---|---|---|---|
| Program Specifications / Correctness | No errors, program always works correctly and meets the specification(s) | Minor details of the program specification are violated, program functions incorrectly for some inputs | Significant details of the specification are violated, program often exhibits incorrect behaviour | Program only functions correctly in very limited cases or not at all | Program does not function correctly at all. |
| Program Documentation / Comments / | Program is well documented / commented | Minor details of the comments are missing | Significant details of the comments are missing | Comments are not clear and/or significant details are missing | No comments or minimum comments |
| Knowledge, Understanding and Analysis | Correct answers, with thorough information and thoughtful reasoning | Correct answers, with complete information | Mainly correct answers, with some misinterpretation | Limited or incomplete answers | Very limited or incomplete answers |

## APPROACH TO LEARNING AND TEACHING

The learning and teaching philosophy underpinning this course is centred on student learning and aims to create an environment which interests and challenges students. The teaching is designed to be engaging and relevant in order to prepare students for future careers.

## COMMUNICATION WITH TEACHING STAFF

We strongly encourage students to communicate with lecturers and peers using the forum in Open Learning (www.openlearning.com). Every section in Open Learning has a forum or comments section at the bottom of the page. By having questions sent to the public forum and using the forum for discussion, all your colleagues can benefit.

For questions regarding assessment marks, please send Oscar an email. If you would like your assignment to be reassessed, please provide an explanation justifying your reason.

All communications will be done through Open Learning. It is each student's responsibility to check Open Learning in regular basis. Major changes (e.g. bugs in the homework) will be also posted in Open Learning.

## GENERAL INFORMATION

### Special Consideration

Please see UNSW-Special Consideration and Student Advice-Special Consideration

If you unavoidably miss assessment deadlines, you must lodge an application with UNSW Student Central for special consideration. If your request for consideration is granted an alternative assessment deadline will be organised.
See: Student-Advice-Reviews and Appeals

Tell the course convenor as soon as possible, via email: o.perezconcha@unsw.edu.au

Applications for special consideration will not normally be received more than 3 days after the assessment due date.

### Student Support Services

See: Student Advice-Student support services.

### Academic Integrity and Plagiarism

The UNSW Student Code outlines the standard of conduct expected of students with respect to their academic integrity and plagiarism.

More details of what constitutes plagiarism can be found here