THE UNIVERSITY OF SYDNEY

School of Computer Science

COMP5318 Machine Learning and Data Mining

semester 1, 2021

Irena Koprinska and Chang Xu

Course web site on Canvas: https://canvas.sydney.edu.au/login/canvas (login with your unikey)

Welcome to COMP5318 Machine Learning and Data Mining!

Machine learning is the process of automatically building mathematical models that explain and generalise datasets. It integrates elements of statistics and algorithm development into the same discipline. Data mining is a discipline within knowledge discovery that seeks to facilitate the exploration and analysis of large quantities for data, by automatic and semiautomatic means. This subject provides a practical and technical introduction to machine learning and data mining. Topics to be covered include problems of discovering patterns in the data, classification, regression, feature extraction and data visualisation. Also covered are analysis, comparison and usage of various types of machine learning techniques and statistical techniques.

The course requires that students have basic linear algebra and probability theory background and are competent at programming in a high level language. We will use Python for the practical part.

Learning outcomes

At the completion of this unit, a student should:

- be able to understand the basic principles, strengths, weaknesses and applicability of machine learning algorithms for solving classification, regression, clustering and reinforcement learning tasks
- have obtained practical experience in designing, implementing and evaluating machine learning algorithms
- have gained practical experience in using machine learning software and libraries
- be able to present and interpret data and information in verbal and written form

1. Teaching team

Unit of study coordinator and lecturers

A/Prof Irena Koprinska, irena.koprinska@sydney.edu.au, unit coordinator and lecturer (weeks 1-6 and 13)
Office: Computer Science Building (J12), level 4, room 450

Dr Chang Xu, c.xu@sydney.edu.au, lecturer (weeks 7-12 and 13)
Office: Computer Science Building (J12), level 3, room 316

Teaching assistants

Henry Weld, hwel4188@uni.sydney.edu.au

Givanna Putri, ghar1821@uni.sydney.edu.au (PASTA)

Tutors

Henry Weld, hwel4188@uni.sydney.edu.au	Claire Hardgrove, claire.hardgrove@sydney.edu.au
Stephen McCloskey, smcc7913@uni.sydney.edu.au	Mashud Rana, mashud.rana@sydney.edu.au
Chen Chen, cche4088@uni.sydney.edu.au	Gio Picones, gpic4558@uni.sydney.edu.au
Nicholas Rhodes, nrho8098@uni.sydney.edu.au	Thomas Selvaraj, tsel7884@uni.sydney.edu.au

How to contact us

If you have questions about the course content, post them on the discussion board PIAZZA, assessable via Canvas. This is the fastest way to get a response from the teaching team or your classmates, almost in real time! You can post your questions anonymously or not anonymously.

2. Timetable

The lectures will be online, the tutorials are either face-to-face or online based on your enrolment.

Lectures (start in week 1)

The lectures are on Monday 6-8pm. They will be **pre-recorded** and available on Canvas to watch during the lecture time. You can access them from "Recorded Lectures".

Tutorials (start in week 2)

The tutorials are 1-hour on Monday 8-9pm, Tuesday 5-6pm and Wednesday 5-6pm; at each of these timeslots there are six tutorials running in parallel. You need to attend only one tutorial as per your timetable.

All tutorials on Monday are online, half of the tutorials on Tuesday and Wednesday are online, the other half are face-to-face in the School of Computer Science labs (level 1, School of Computer Science building, J12).

All online tutorials will be **live streamed via Zoom**. One tutorial will be recorded every week and made available on Canvas in "Recorded Lectures".

Please attend your allocated tutorial; if there are issues with Zoom during the online streaming, you can attend another one.

For the face-to-face tutorials we need to take attendance – this will only be used for COVID tracing (required by the University), not for any other purpose. We don't want to impose attendance as a class requirement –

we believe in academic freedom – you are mature and responsible individuals and should attend your classes because they are beneficial for you, not because this is mandated.

Tutorial	Time	Tutor
RE (remote = online)		
1 M20A Zoom	Monday 20-21	Henry
2 M20B Zoom	Monday 20-21	Stephen
3 M20C Zoom	Monday 20-21	Nicholas
4 M20D Zoom	Monday 20-21	Mashud
5 M20E Zoom	Monday 20-21	Chen
6 M20F Zoom	Monday 20-21	Gio
7 T17A Zoom	Tuesday 17-18	Claire
8 T17B Zoom	Tuesday 17-18	Nicholas
9 T17C Zoom	Tuesday 17-18	Stephen
10 W17A Zoom	Wednesday 17-18	Gio
11 W17B Zoom	Wednesday 17-18	Claire
12 W17C Zoom	Wednesday 17-18	Mashud
CC (on campus = face-to-face)		
1 face-to-face	Tuesday 17-18, CSC Lab 115	Mashud
2 face-to-face	Tuesday 17-18, CSC Lab 114	Chen
3 face-to-face	Tuesday 17-18, CSC Lab 130A	Thomas
4 face-to-face	Wednesday 17-18, SCS Lab 130A	Stephen
5 face-to-face	Wednesday 17-18, SCS Lab 118	Thomas

3. Course website

The main place for this course is the Canvas COMP5318 website, accessible from:
https://canvas.sydney.edu.au/login/canvas

We will use it for all teaching materials (lecture slides, tutorial notes and tutorial solutions), assignment specifications and posting of your marks.

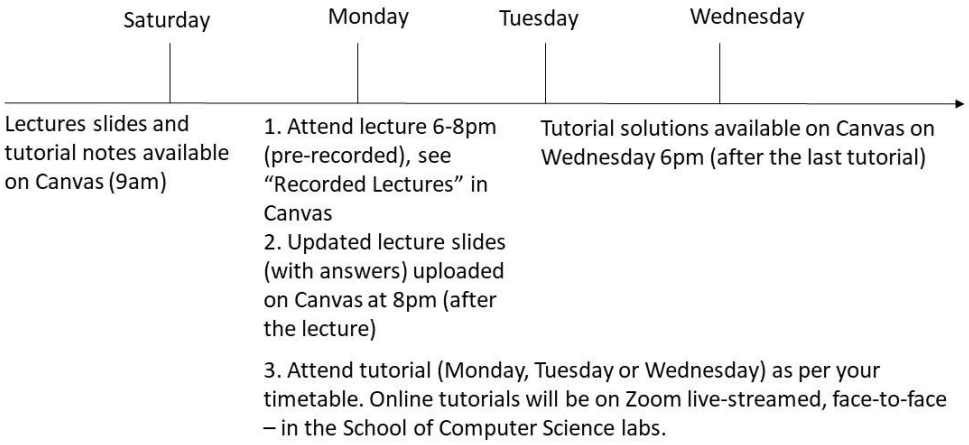
In addition to Canvas, we will also use the discussion board Piazza and the autograding system PASTA. They will be linked to Canvas.

4. Availability of teaching materials

The course materials (lecture slides, tutorial notes and homework submission box) will be available every week in advance on Saturday morning on Canvas. E.g. the materials for week 2 will be available on Saturday 9am in week 1.

Sometimes the lecture slides may not include the answers to all questions and exercises that we will do during the lecture; in this case, the complete version with the answers will be available after the lecture (at 8pm on Monday).

The tutorial solutions will be available on Canvas after the last tutorial on Wednesday which finishes at 6pm, i.e. the tutorial solutions for week 2 will be available on Wednesday 6pm in week 2.



5. Weekly schedule

Week	Date	Topic	Lecturer
1	1 March	Administrative matters and course overview. Introduction to machine learning and data mining. Data: cleaning, pre-processing and similarity measures.	Irena
2	8 March	Nearest neighbour. Rule-based algorithms.	Irena
3	15 March	Linear regression. Logistic regression. Overfitting and regularization.	Irena
4	22 March	Naïve Bayes. Evaluating machine learning methods. Assignment 1 out (Monday)	Irena
5	29 March	Decision trees. Ensembles.	Irena
Mid-semester break			
6	12 April	Support vector machines. Kernels. Dimensionality reduction.	Irena
7	19 April	Neural networks - perceptrons and backpropagation algorithm. Assignment 1 due (Friday)	Chang
8	26 April	Deep neural networks: convolutional and recurrent.	Chang
9	3 May	Clustering I:Partitional, model-based and hierarchical. Assignment 2 out (Monday)	Chang
10	10 May	Clustering II: Density-based and grid-based. Evaluating clustering results.	Chang
11	17 May	Markov models – HMM, MEMM, CRF.	Chang
12	24 May	Reinforcement learning. Assignment 2 due (Friday)	Chang
13	31 May	Guest lecture. Revision.	Irena

6. Assessment overview

Summary

1. Assignment 1: 15%; due on Friday week 7 (submission in PASTA)
2. Assignment 2: 25%; due on Friday week 12 (submission in PASTA and Canvas)
3. Exam: 60%; online (take-home, 2 hours)

Component	Due date and submission	Notes
Assignment 1 Weight: 15%	Friday week 7, 11.59pm Individual Submitted in PASTA Late submissions: Late submissions are allowed up to 3 days late. A penalty of 5% per day late will apply. Assignments more than 3 days late will not be accepted (i.e. will get 0).	Given a problem, you will be required to apply one or more machine learning algorithms to solve it. This will include writing a computer program in Python to solve the problem.
Assignment 2 Weight: 25%	Friday week 12, 11.59pm Individual or in pairs (no more than 2 people). We encourage working in pairs. You can pair with a student from your or another tutorial. Late submissions: late submissions are allowed up to 3 days late. A penalty of 5% per day late will apply. Assignments more than 3 days late will not be accepted (i.e. will get 0).	As in Assignment 1, you need apply machine learning algorithms to a given problem. In addition to the computer program in Python, you need also to submit a report discussing the results. Assignments 1 and 2 will be posted on Canvas with information how to submit them.
Exam Weight:60% Online exam, take-home	During the exam period Individual Exam duration: 2 hours	During exam period. At least 40% on the exam is required to pass the course. Information about the exam and a sample exam paper will be available in week 13.

Special considerations: If you experience short-term circumstances beyond your control, such as illness, injury or misadventure, which affect your preparation or performance in an assessment, you may apply for special consideration. There is a centralised University system; all applications are submitted online after login to "myUni" and are processed by the Student Administration Services unit.

Important: You are required to submit your special consideration application form within 3 working days from the date when the assessment was due. For more information see: http://sydney.edu.au/special-consideration

Passing this unit of study: The School of Computer Science has the following policy: To pass a unit of study, a student must achieve at least 40% in the written exam. A student must also achieve an overall final mark of 50 or more in order to pass a unit of study.

7. Academic honesty

All cases of plagiarism and academic dishonesty will be investigated. There is a new process and a centralized University system and database. Please read the University Policy on Academic Honesty carefully: http://sydney.edu.au/elearning/student/EI/academic_honesty.shtml

Please note that:

- If you copy from another student, website or other source, you have committed an act of **plagiarism.** This includes copying the whole assignment/exam answer or only a part of it.
- If you make your work available to another student to copy, you have committed an act of **academic dishonesty**
- If you engage another person to complete your assignment/exam (or a part of it), for payment or not, you have committed an act of **academic misconduct.** Your case will be forwarded to the University Registrar for investigation which is very serious.

The penalties for academic dishonesty, plagiarism and misconduct are severe and include: 1) a permanent record of academic dishonesty, plagiarism and misconduct in the University database and your student file, 2) mark deduction, e.g. from 0 for the assignment to fail for the course, 3) expulsion from the University and cancelling of your student visa.

To detect plagiarism in reports (Assignment 2 report part), we will use TurnItIn, which is linked to Canvas.

To detect plagiarism in programming code we will use the similarity detection system MOSS. MOSS is designed especially for detecting plagiarism in programming code. We will compare your programming assessments with these of other students (current and previous) and the Internet. MOSS is an extremely good system – it cannot be fooled by changing the variable names or the order of the conditions in `if` statements.

Please do not confuse legitimate cooperation with cheating. In individual assignments, you can discuss the assignment with another student, this is a legitimate collaboration, but you cannot complete the assignment together – everyone must write their own code and report.

Important: If someone asks you to see or copy your assignment, or to complete the assignment instead of them, just say: *I can't do this. This is against the University policy. I will not risk my reputation and future by doing this.* **Be smart and don't risk your future by engaging in plagiarism and academic dishonesty!**

8. Textbooks

Textbooks

Ian H. Witten, Eibe Frank, Mark Hall and Christopher J. Pal (2017). *Data Mining - Practical Machine Learning Tools and Techniques*, 4th edition, Morgan Kaufmann. (You can also use the 3nd edition)

Pang-Ning Tan, Michael Steinbach, Anuj Karpathe and Vipin Kumar (2019). *Introduction to Data Mining*, 2nd edition. Pearson. (you can also use the previous edition)

Books for the practical part using Python:

Andreas C. Mueller and Sarah Guido (2016*). Introduction to Machine Learning with Python: A Guide for Data Scientists*, O'Reilly.

Aurelien Geron (2019). *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow*, O'Reilly.

All books are available from the library as hard copies. All books except Tan are also available online from the library: https://library.sydney.edu.au/ - there is no electronic version for Tan from the publisher, that's why it is not available online.