158.755

Data Science – Making Sense of Data

Course and Assessment Guide

Dr Teo Susnjak

School of Natural and Computational Sciences

2020

The Course

Prescription

This paper studies the science of drawing knowledge and insights from data. The essence of the course is learning how to reason about data-oriented problems – how to formulate a problem from data, how to ask questions from data, and how to devise, analyse and present data-driven solutions. Some of the key skills taught are data acquisition and wrangling as well as data visualization. Several machine learning/data mining algorithms are covered in a context of real-world problems which also includes natural language processing. Readings and discussions on the emerging topics in this field form an important component of this paper, as well as programming and the usage of industry-based tools. Course work includes a combination of both individual and group work, together with presentations.

Learning outcomes

Students who successfully complete this paper should be able to:

1. Perform data wrangling and data visualization skills using an open-source programming library.
2. Demonstrate the ability to formulate a problem from data, ask questions of data, devise a solution and present findings from real-world problems
3. Apply several machine learning and data mining algorithms both programmatically and from software toolkits to solve classification problems
4. Reason about which machine learning/data mining algorithms to apply on given problems
5. Compare the generalizability of different machine learning/data mining algorithms for a given problem

Teaching Team

Course Coordinator:	Teo Susnjak (t.susnjak@massey.ac.nz)
Lecturing:	Teo Susnjak
Tutoring:	Rahila Umer (rahiumer@gmail.com) Stan Wang (yulinzxc@gmail.com) Tyrel Glass (glaty851@gmail.com)

Course Times and Location

Course day and time: Mondays 8:00 am to 11:00 am

Location: CLQB1 Computing Lab

Topics

Weeks	Topic
1	Course intro, Data Science, Python intro, Jupyter IDE intro
2	Data wrangling - Python Pandas: Series, DataFrame
3	Plotting, DataFrame apply functions, integration, group by, aggregation
4	Introduction to Regression
5	Data acquisition, Web Scraping, Web APIs
6	kNN, Normalization, Distance Metrics
7	Presentations (Classification, Generalization, Overfitting, Evaluation)
8	Clustering with k-Means
9	Naïve Bayes Classification
10	Machine Learning Theory and Catch up session
11	Working with Time Series Data
12	Project Presentations and Course Review

Study Resources

There are a range of study resources for this course. During the course, the primary learning resources will be the Jupyter Notebooks which will facilitate learning the Python and machine learning components of the course. The Jupyter Notebooks will be provided each week as well as readings materials for selected topics.

Given the nature of programming, you will also be required to carry out a lot of self-directed learning and troubleshooting which will necessitate you learning how to search the internet for help around specific programming challenges that you will inevitably come across. Learning how to do this well, is an important part of learning how to program in this domain space.

This content in this course has been multiple times already. In previous years we have made screencast recordings of some course content. We have decided to make this archived content available to this course's cohort as it could be valuable to some students. This archived course material is slightly different to some topics delivered this year because we are always updating the content. However, in essence, all the same key topics are covered. This material is available on Stream under the "Archive Course Material (optional)" section. The relevant screencasts from this catalogue are Python recordings from 1 through to 30.

Coronavirus Disruption Mitigation Strategy

Disruption due to the outbreak of the coronavirus is likely to affect some international students. This is a developing situation and so we will keep you updated with latest information.

However, for those students who expect to have their travel to NZ delayed due to this, you are expected to make a start to your studies using the "Archive Course Material (optional)" mentioned above. By working through this online material, you will ensure that you cover all the key topics covered in the first 4-5 weeks of the course, which will enable you to complete your assignments and meet the learning outcomes associated with the first third of the course.

If you expect to be prevented from attending the courses, then you are urged to immediately begin your studies using the online material provided. Begin with screencasts in installation of Anaconda and getting started with Jupyter, then work through all the

Python notebook screencasts. The relevant screencasts from this catalogue are Python recordings from 1 through to 30.

More information regarding Coronavirus mitigation plans fir affected students will be announced on Stream.

Recommended Textbooks

There are no set texts for this course. The following textbooks are not compulsory but may be of assistance to students wishing to explore the topics being covered in more depth.

McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media.

For interested students wishing to pursue data mining further:
Witten, I. H., Frank, E. & Hall M. A (2011). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Compulsory reading

Some weeks may include compulsory reading material on the Stream site.

Recommended reading

Some weeks may include supplementary reading articles and links on the Stream site.

Software and Hardware

You are encouraged to bring your own laptops, loaded with the Python development environment to the classes. This is not compulsory though.

You are however, strongly encouraged to install the software we will be using in this course on your home machines, which will enable you to devote more time to practicing programming and working on your assignments. Otherwise, you will need to use the Massey Computing Labs for this.

We will be using Python 3.7 for this course. A suitable distribution of Python will be installed in the teaching rooms. We will be using the free Anaconda 2018.12 distribution for this course. You can download the software from here: https://repo.continuum.io/archive/ Please install the Anaconda3-2018.12 distribution that is suitable for your platform.

Stream: Your online learning environment

Accessing Stream helps you do well in the course in three ways:

1. **Lecturer-to-Student Communication:** I will post any important notices, instructions and additional readings that arise on Stream. By checking Stream often you will always know 'what's going on'.

2. **Student-to-Lecturer Communication:** I encourage you to communicate with me via Stream if you have any questions. Often these questions alert me to potential areas of confusion that the whole class can benefit from. I aim to respond to your inquiries in a timely manner.

3. **Student-to-Student Communication:** Stream allows you to communicate with other students via a forum. Post a message introducing yourself to the class.

Stream discussion forum etiquette and expectations:

If you have a general question about the concepts covered in class, you are encouraged to post the query on the forum first, so that other students can also benefit. This gives other students the chance to provide answers. The teaching staff do their best to keep track of the questions and, if no suitable answers have been provided within a couple days, a member of the teaching team will offer a response. Sometimes we may miss them, so email us if this happens. Sending an email query to each member of the teaching staff (about the same query) separately is not appreciated. Send it to one staff member only please.

When engaging in the forum discussions:

1. All communication is expected to conform to the Code of Student Conduct found here: http://www.massey.ac.nz/massey/about-massey/calendar/studying-at-massey-university/code-of-student-conduct.cfm
2. The forum is designed to facilitate student learning; therefore, make sure that discussion topics are confined to relevant course topics.
3. The forum is not the appropriate place for expressing complaints, criticisms or concerns you might have about the course, technologies being used, teaching staff or the university. If you want to express any of the above, raise them directly with the teaching staff, the class advocate or other formal avenues.

Failure to comply with the above expectations may lead to disciplinary action. We are all responsible for creating and maintaining a professional and safe forum to enhance learning, so let's all work to that end.

How to approach your study

This course is entirely taught in computing labs and will be delivered in a practical, hands-on workshop-mode. This means that teaching and practical work will be fully integrated in the 3-hour teaching blocks each week.

You will be learning by doing and therefore you will be expected to spend a significant amount of time outside of class working through the programming notebooks provided with this course, assignments, as well as doing your own exploratory exercises.

There are no shortcuts to learning how to program. Reading about it will not advance your skills. You simply have to dive in and do lots of it.

Each week I suggest you read the online material before the lecture. I would also like you to make a start on the Jupyter Notebooks before class, solving some exercises and familiarising yourselves with the content.

Assessment

Assessment	Due date	Learning outcomes assessed	Weighting
Executable Jupyter Notebook Report	Wk 3/4	1, 2	15%
Executable Jupyter Notebook Report	Wk 6	1, 2, 3	25%
Executable Jupyter Notebook Report	Wk 8	1, 2, 3, 4	25%
Executable Jupyter Notebook Report, Software Artefact and Presentation	Wk 12	1, 2, 3, 4, 5	35%

Final Examination

There is no final examination for this course.

Requirements for completing the course

50% or above in total across all assessments.

Extensions and late assignments

Assignment deadlines are strict. However, each student will be given an allowance of **5 days (in total, not for each assignment)** worth of late submissions. You may use your late submission 'grace-days' however you like, with no questions asked. You can ‘spread’ the allowance across the 4 assignments. **Use this allowance wisely as no further extensions will be granted once it is used up.** Late submissions are rounded up to a day, so if you submit 30 minutes late, this counts as a full day late submission. Late submissions, or ‘grace-days’ do not apply for scheduled presentations.

Please do not ask for any further assignment extension (unless circumstances are catastrophic and verifiable), instead, refer to official university guidelines for an Aegrotat Pass or Impaired Performance.

Assignment submission

Please submit electronic versions of your assignments through Stream. The expected turnaround time is 21 working days. Feedback will be provided with assignment marks. A marking guide will accompany each assignment brief.

Academic integrity

It is mandatory that any assessment items that you submit during your University study are your own work. Massey University takes a firm stance on academic misconduct, such as plagiarism and any form of cheating.

Plagiarism is the copying or paraphrasing of another person's work, whether published or unpublished, without clearly acknowledging it. It includes copying the work of other students and reusing work previously submitted by yourself for another course.

It also includes the copying of code from unacknowledged sources.

Academic integrity breaches impact on students as it disadvantages honest students and undermines the credibility of your qualification. Massey has purchased a licence to utilise Turnitin®, a text matching web application to assist with the detection of copying. Assignments submitted to Turnitin will be compared with material available on the world wide web including electronic books, journals, newspapers, cheat sites (or paper mills), web pages and previously submitted assignments. Your lecturer will let you know if your assignments are going to be submitted to Turnitin.

Plagiarism, and cheating in tests and exams will be penalised; it is likely to lead to loss of marks for that item of assessment and may lead to an automatic failing grade for the course and/or exclusion from reenrolment at the University.

Please see the Academic Integrity Guide for Students on the University website for more information. The Guide steps you through the University Academic Integrity Policy and Procedures. For example you will find definitions of academic integrity misconduct, such as plagiarism; how misconduct is determined and managed; and where to find resources and assistance to help develop the skills of academic writing, exam preparation and time management. These skills will help you approach university study with academic integrity.

Conditions for aegrotat pass and impaired performance

If you are prevented by illness, injury or serious crisis from attending a compulsory learning experience, an examination or completing an element of assessment (worth 10% or more) by the due date, or if you consider that your performance has been seriously impaired by such circumstances, you may apply for aegrotat or impaired performance consideration. You must apply on the Aegrotat & Impaired Performance Application form available from the Massey University website. The completed form must be accompanied by a certificate signed by a health professional, and/or corroborating evidence.

Grievance procedures

A student who claims that he/she has sustained academic disadvantage as a result of the actions of a University staff member should use the University Grievance Procedures. Students, whenever practicable, should in the first instance approach the University staff member concerned. If the grievance is unresolved with the staff member concerned, the student should then contact the relevant Head of Institute/School/Department or College office for further information on the procedures. The procedures can be found on the University website in the University Calendar.