School of Mathematics and Statistics, Faculty of Science
The Red Centre, Centre Wing - NSW 2052 Australia
Kensington campus

COURSE OUTLINE

DATA3001
Data Science and Decisions in Practice
6 Units of Credit (6UOC)
(Term 3, 2020)

Course authority and lecturer: Dr Pierre Lafaye De Micheaux

Contact: Office: RC-2050, School of Mathematics and Statistics
E-mail: `lafaye@unsw.edu.au`
Phone: (00.[+612]) 9385 7029
Web: `https://web.maths.unsw.edu.au/~lafaye`

Consultation hours: Please use e-mail to arrange an appointment.

Other lecturers: A/Prof Gustavo Batista, Computer Science & Engineering (g.batista@unsw.edu.au); Dr Michele de Nadai, School of Economics (m.denadai@unsw.edu.au); Dr Zdravko Botev, School of Mathematics and Statistics (botev@unsw.edu.au).

1 Course Overview

This is the capstone course for the Data Science and Decisions program. The course will bring students in the three streams together to share their knowledge, expertise and training in a way that is typical of industry. Students will attend seminars by industry representatives from Data Science industries, and students will work on group projects related to real world industry problems. Typical groups will be composed of students across the three different streams of the Data Science and Decisions program. The course will expose students to Data Science as it is practiced in industry.

2 Pre-requisites, Exclusions

Pre-requisites: students are assumed to have completed all level I and level II courses in the 3959 program before enrolling in this course.

3 Schedule and Format of the Course

The class timetable is available at `http://timetable.unsw.edu.au/2020/DATA3001.html`.

Week 1 will be dedicated to seminars given by industry representatives from Data Science industries or research organisations. These representatives will discuss about their day-to-day practice of data science. They will present some problems they have in their workplace. They might provide data sets that could be used by the students to start addressing these problems. Moreover, at the end of Week 1, groups of around 5 students will be formed and each group will have to pick up a project. (Note that due to potential dropouts a group might end up with less than 5 students.) During Weeks 2–5,7–10, students will have to meet either online (e.g., in their BlackBoard Collaborate breakout group dedicated space) or in Active Learning Spaces (`http://learningenvironments.unsw.edu.au`) during the two pre-booked weekly 2-hour slots. All instructors will be present in order to help, in turn, each group of students with their projects.

4 Assessments

Assessments are designed to help you build a portfolio. This will be very useful during job interviews to show your real-world experience: you will be able to explain to an employer the entire *data science iterative workflow* of your project, which can be potentially briefly described as follows:

- **Step 1:** choosing and understanding of a problem to solve;
- **Step 2:** identifying one or several appropriate raw data sets;
- **Step 3:** wrangling, exploring and cleaning the data at hand;
- **Step 4:** planning the project, the roles in the team and the tasks to perform;
- **Step 5:** assessing the state of the art in the domain by conducting a literature review;
- **Step 6:** selecting the appropriate statistical and machine learning techniques and computing tools;
- **Step 7:** modelling, explaining, inferring and predicting in order to add value;
- **Step 8:** creating visualisations that will give better insight;
- **Step 9:** reporting, implementing and deploying your findings.

Students will work on a *group project* of relevance to the broad Data Science industry (private companies, startups, governmental bodies of research, etc.) involving all steps of the data science workflow. Problems and data sets will either come from the Data Science industries involved in the seminar series of Week 1, or from publicly available data sets (see, e.g., https://www.springboard.com/blog/free-public-data-sets-data-science-project/ or https://www.kaggle.com/datasets or https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research). Several groups of students can work independently on the same project.

Groups of approximately 5 students coming from the three different streams of the Data Science and Decisions program will be created during the seminar period of Week 1. If a student cannot find by him/herself other team mates, he/she will be randomly allocated to a group by an instructor. Each student must read the information contained at https://student.unsw.edu.au/groupwork and confirm having read it. This page will inform you about the nature of group work, about what you should expect and the expectations teachers have of you in group learning situations.

All the results of a project should be entirely accessible and reproducible (e.g., think of using a random seed in the case your analysis leads you to randomly generate data). To this end, students are requested to use Markdown and Github (with Git). For your quantitative analyses, you will use either R (e.g., through RStudio and Rmarkdown) or Python (e.g., via an ipython notebook such as Jupyter). Authorisation should be requested from the lecturers and properly justified if you plan to use another software. Note that Linux is more efficient than Microsoft Windows to deal with large data sets. All mathematical formulas will need to be written using the LaTeX language.

A first assessed activity will be the **Choice of a Problem** (10%). Each group will have to choose a complex open-ended real-world problem to solve (and related data sets, potentially big, to use) among those presented during the seminar in Week 1. The first step is to choose a topic and delimit the problem which will be studied. You should understand and define the overall problem and propose a solution. Having only a short time to complete the project, it is crucial that the problem that will be studied is well defined. The approach to solve the problem should be original, so it will be necessary to carry out a preliminary "literature review". This will prevent any plagiarism. This will also enable you to situate the project in a more global context. At this stage, one can identify potential approaches and software that will be used to solve the problem. It is necessary to carefully plan any simulation and to decide what statistical analyses will have to be carried out. All the sub-steps of the project should be planned precisely (a detailed schedule will be created). A two-page proposal will be submitted and discussed with one of the instructors in Week 4 for approval (which

implies that this task should be finished before Thursday 4pm in Week 4). Assessment criteria involve the following: a clear definition of the problem; a clear description of the data format and their storage; a clear description of the data (variables, missing and corrupt values, etc.); the level of difficulty of the chosen data sets (size, complexity, messiness) and its relevance for the chosen problem; the appropriate choice of software and statistical methods to solve your research questions; the precision of your schedule activities.

In order to assess **Group Project Dynamics** (10%+5%), each student will have to write an individual $\leq$ 500 words report on the effectiveness of their team work process and submit this report by the end of Week 10. To ensure this task is done smoothly to its proper completion, each student will have to:

- fill in a weekly *Teamwork Checklist*; (A teamwork checklist template can be found at this page `https://student.unsw.edu.au/groupwork`. Such a checklist will be made available each week to the instructors so that they can monitor your work. They will include a peer assessment, where each student will be given the opportunity to comment on the contribution made by other team members.)
- maintain a *Reflective ePortfolio*. (The reflective ePortfolio task requires you, among other things, to write reflectively about your role as a team member. Reflection has been demonstrated to be a useful tool to support professional learning in authentic contexts where there are many uncontrolled variables. Often the process of writing these things down brings aspects or relationships to light that you might have been previously unaware of.)

The individual *Group Dynamics Report* will consist in an assessed written report of less than 500 words that is a summary of the above two tasks. Each student's report will be given a mark M (over 10%) by the instructor and a mark M' which is the average of the marks given by the other students within the same team. The global mark (over 15%) will be $M + M'$. There are several reasons behind this work. You should reflect about what you have learned in terms of working as a team during this course. (Note that Data Science is meant to be a highly interdisciplinary field of research.) Here are a few suggestions: explain how you split the various tasks between the members of the team, what (computing) tools you have used to organize your work efficiently, what went wrong, what worked well, what you would change in the future to improve the interaction between the members of a team, what are the main ingredients to make a team achieve its goals. You can comment about your contributions and the contributions of the others (attendance is recorded, so the lecturers can also forge their own impression of the team interaction during the time they spent with you). We also hope to give some good advice to students next year based on your reports.

A third assessed activity will be a **Group Report** (45%). This group report will consist of a PDF file of no more than 30 pages (appendices excluded). A LaTeX template will be provided and it is mandatory to use it for your group report. Note that we strongly advise that you create first either an RMarkdown or a Python+Markdown source file that could be then converted to a LaTeX document. Also, all the files of your project (data, codes, group report, etc.) will have to be stored (and updated regularly) on one of the team's member private Github Pro account (see `https://education.github.com/pack` to get free access as a student). You must name your GitHub repository as follows: `Team xx - Name of the project`. It is your responsibility to provide an access to this repository to all the lecturers so that they are able to verify what is done each week. Computing codes for all numerical and graphical results will be made available and can be part of the assessment for this task. Each group report will be assessed by one of the instructors. You will then have to distribute the number of marks given to your report among you. (For instance, say the group work is given 7/10 and there are 5 group members. The group is then given $7 \times 5 = 35$ marks to share amongst yourselves, with a cap of 10 for each one of you. So one person might get 7.5, but then someone else has to get 6.5 with the other three staying at 7, i.e. it still has to add up to 35.) Criteria for a good report include: a table of contents; page numbers; an appendix; appropriate tables and graphs; a nice presentation of the data sets; a goal clearly explained; the quality of written expression, English grammar; respect of due time, maximum page numbers, etc.; the ability to clean and wrangle your data efficiently; the speed of execution of your codes; the quality of the exploratory data analysis (EDA) performed; a deep understanding of the statistical/machine learning/computational/optimisation methods and algorithms used, and their appropriate choice given the problem at hand; the level of difficulty of the statistical analyses conducted; a high prediction rate if applicable; the proper interpretation of the data and the results of statistical analyses; the communication of the results via a nice data storytelling.

The last assessed activity will be a **Group Presentation** (30%). Oral presentations (3 mn per student to present orally their contribution for a total of $\leq$15 mn per group) will consist in two tasks:

- creation of a team video;
- peer review of the videos of 4 other groups.

The group video will have to be uploaded before the end of November. Instructions on how to submit this video will be provided in due time. (Confidentiality issues should be taken into account here.) Every student will have to start by very briefly presenting him/herself and show their face and their student ID card with your zID clearly visible. The format of the video could be made of PDF or Powerpoint slides with a voice over. Note that RShiny web apps can also be used for visualisation purposes and are a nice add-on to a portfolio. They can be freely hosted at http://www.shinyapps.io/ for example. On a final note, a video is a nice enhancement of a portfolio.
Moreover, each student will have until the 10th of December to peer review the videos of four other teams. They will provide an evaluation mark (over 100) to each student in the video, as well as constructive comments/criticisms on both the presentation itself and the way the problem was tackled. A list of criteria for this assessment, as well as general advice for giving a good presentation, will be provided to the students beforehand. Each student s is expected to receive an individual mark (over 100) computed according to the following formula:

$$0.5R_s + 0.5G_s$$

where R_s is the average (over 100) of the marks received by s from all students belonging to four other groups and by the instructors (and eventually the representatives from the Data Science industries), and where

$$G_s = 100 - (N_s - 1)^{-1} \sum_{s'} |g_{s\to s'} - \mathrm{Mg}_{s'}|$$

where

- N_s is the total number of students s' (belonging to four other groups) that the student s is assigned to mark,
- $g_{s\to s'}$ is the mark given by student s to student s' (if missing, $g_{s\to s'}$ will be replaced by $\overline{IM}_{s'}$),
- $\mathrm{Mg}_{s'}$ is the median of the marks given to student s'.

Task	Weighting	Duration	Date Due
Choice of a Problem	10%	4 weeks	week 4
Group Project Dynamics	15%	9 weeks	end of week 10
Group Report	45%	9 weeks	end of week 10
Group Presentation	30%	3 mn/student, 15 mn/group	30/11 and 10/12

Knowledge and Abilities Assessed

- The Choice of a Problem will assess your ability to understand and appreciate Data Sciences (and Data Analytics) in the modern world.
- The Group Project Dynamics will assess your ability to collaborate within teams, and to undertake significant project work in a self disciplined, organised, and professional manner from conception to documentation.
- The Group Report will assess your ability to apply statistical and computational techniques, and business sensibilities, to real-world problems, as well as to write technical reports in a professional manner.
- The Group Presentation will assess your ability to express your ideas orally and to assess the work of others.

5 Course Aims

The aim of DATA3001 is that at the end of the term, you should be able to identify and place the data-driven problem into an analytical framework, solve the problem through computational means, interpret the results and communicate your findings to a diverse audience. All such skills are highly valued by employers. This course will foster the ability to work in an interdisciplinary team, to translate problems between two or more disciplines and this is essential for both professional and research pathways in the future.

The activities and assessment for the course will contribute to the core science graduate attributes of 'Research, inquiry and analytical thinking abilities', 'Capability and motivation for intellectual development' and 'Communication'.

6 Course Learning Outcomes

The successful student will

- ✓ gain an understanding and appreciation of Data Sciences (and Data Analytics) in the modern world.
- ✓ apply mathematical and computational techniques, and business sensibilities, to real-world problems.
- ✓ gain skills in writing technical reports.
- ✓ gain skills in assessing reports.
- ✓ gain skills in oral presentations.
- ✓ be experienced and empowered in undertaking significant project work in a self disciplined, organised, and professional manner from conception to documentation.
- ✓ be able to build statistical models and understand their power and limitations
- ✓ be able to acquire, clean and manage data
- ✓ be able to visualise data for exploration, analysis, and communication
- ✓ be able to collaborate within teams
- ✓ be able to deliver reproducible data analysis
- ✓ be able to manage and analyze massive data sets
- ✓ be able to assemble computational pipelines to support data science from widely available tools
- ✓ be able to apply problem solving strategies to open ended questions

The above outcomes are related to the development of the Science Faculty Graduate Attributes, in particular: 1. **Research, inquiry and analytical thinking abilities**, 4. **Communication**, 6. **Information literacy**.

7 Resources

No lecture notes will be provided for this course. Some useful books are listed in the References section at the end of this document.

You should check regularly UNSW Sydney's Moodle web page (https://moodle.telt.unsw.edu.au) for new materials, as well as for announcements about assessment tasks.

8 Course Evaluation and Development

The School of Mathematics and Statistics evaluates each course each time it is run. We carefully consider the student responses and their implications for course development. It is common practice to discuss informally with students how the course and their mastery of it are progressing. Thank you in advance for your contribution!

9 Administrative Matters

It is the student's responsibility to be familiar with UNSW and School of Mathematics and Statistics policies.

Assessment Policies

The School of Mathematics and Statistics has a set of assessment policies that you can consult at
http://www.maths.unsw.edu.au/currentstudents/assessment-policies

If you are absent (e.g., ill) from the final examination, you must apply for special consideration using the UNSW Special Consideration online service. Information regarding additional assessments are available at
http://www.maths.unsw.edu.au/currentstudents/additional-assessment

School Rules and Regulations

Fuller details of the general rules regarding attendance, release of marks, special consideration, etc., are available at
http://www.maths.unsw.edu.au/currentstudents/student-services

Plagiarism and Academic Honesty

Plagiarism is presenting another person's work or ideas as your own. Plagiarism is a serious breach of ethics and is not taken lightly. It undermines academic integrity and it will not be tolerated.

UNSW Sydney has a set of rules and regulations for academic conduct, honesty and plagiarism. See
http://www.lc.unsw.edu.au/academic-integrity-plagiarism
and
https://www.gs.unsw.edu.au/policy/documents/studentcodepolicy.pdf

10 References

1. P. Lafaye de Micheaux, R. Drouilhet and B. Liquet (2013). *The* R *Software : Fundamentals of Programming and Statistical Analysis*. Statistics and Computing. New York: Springer.
   http://link.springer.com/book/10.1007/978-1-4614-9020-3
   Translations exist in French, Chinese and Indonesian.
   http://biostatisticien.eu/springeR

2. D. P. Kroese, Z. Botev, T. Taimre, R. Vaisman (2020). *Data Science and Machine Learning: Mathematical and Statistical Methods*. Chapman & Hall/Crc Machine Learning & Pattern Recognition.
   https://www.amazon.com/Data-Science-Machine-Learning-Mathematical/dp/1138492531

3. T. Hey, S. Tansley and K. Tolle (2009) . *The Fourth Paradigm, Data-Intensive Scientific Discovery*. Microsoft Research.

4. *A Beginner's Guide to Getting Your First Data Science Job*. Springboard.

5. S. Gutierrez. *Data Scientists at Work*. friendsoft apress.

6. G. James, D. Witten, T. Hastie and R. Tibshirani (2015). *An Introduction to Statistical Learning with Applications in R*. Springer.

7. M. Kirk (2017). *Thoughtful Machine Learning with Python*. O'Reilly Media, Inc.

Some of the above books might be available at your local library: https://primoa.library.unsw.edu.au/primo-explore/search?vid=UNSWS

Last update: August 8, 2020.