NATIONAL UNIVERSITY OF SINGAPORE

Department of Statistics and Data Science

2023/2024 IND 5003 Data Analytics for Sense-making Semester I

Lecturer

Assoc. Prof. Lim Tiong Wee
Department of Statistics and Data Science
tiong_wee_lim@nus.edu.sg

Course Resources

1. Canvas > IND5003
   - Contains all teaching materials and datasets.
   - Submission of all assignments, take-home exam, and all project-related material (except project videos).
2. Microsoft Teams > Teams > 2310_IND5003_L1
   - Our weekly lessons are **on-site**: Tuesdays, 19:00–22:00, LT34.
   - Contains the recordings of our weekly **on-site** lessons.
   - Can be used as a chatroom for project discussions.
   - Contact instructors here.
3. Microsoft Stream > My Content > Groups > 2310_IND5003_L1 > Videos
   - Project videos should be uploaded here.

All lessons this semester are **on-site**. You are highly encouraged to send feedback on anything and everything about the course.

Course Description

Technological advancements such as cyber-physical systems and the Internet of Things are enabling connected machines which collect a tremendous volume of structured and unstructured data. This course covers essential analytics tools and techniques for performing supervised and unsupervised learning on that data. It focuses on applications in such domains as consumer, human resource, manufacturing, medical and retail to identify patterns and insights for process improvements and decision-making. These tools include the Python programming language; the techniques include frequently used time series models and predictive models such as regression, random forests, neural networks and deep learning.

Learning Outcomes

The learning outcomes for this course are

- to obtain familiarity with data science tools such as Python;
- to obtain a general understanding of data science techniques; and
- to gain an idea of domain specific data science techniques and tools.

Schedule

This is a tentative schedule:

Wk	Date	Lecture	Assessment Due
1	15 Aug	Introduction	
2	22 Aug	Data Manipulation	
3	29 Aug	Statistical Inference	DataCamp: Introduction to Python
4	5 Sep	Simulation	
5	12 Sep	Customer Analytics	Python Assignment 1 due
6	19 Sep	(Project proposal review)	DataCamp: Intermediate Python, Data Manipulation with pandas
		Recess Week	
7	3 Oct	Regression	
8	10 Oct	Supervised Learning	Take-home Exam
9	17 Oct	Vision-based Techniques	
10	24 Oct	Natural Language Processing	DataCamp: Machine Learning with scikit-learn
11	31 Oct	Time Series Analysis	Python Assignment 2 due
12	7 Nov	(Buffer)	Project due
13	14 Nov	Projects Review	

Each lecture will be conducted as a workshop with an emphasis on hands-on methodology. We will demonstrate lots of Python code and you will have ample opportunity to try things out for yourselves.

The intention is not to delve deeply into any method or domain. Instead, we aim to expose you to various methods and tools, and to point you to the appropriate courses to follow up with.

The lessons will be recorded and made available for reference by the class only on Microsoft Teams.

Assessment Components

This is a breakdown of the assessment components and their weights:

DataCamp(4 courses)	20%	Due at 23:59 on the Sunday of Weeks, 3, 6, and 10.
Assignments (2 Python worksheets)	10%	Due at 23:59 on the Sunday of Weeks 5 and 11.
Take-home Exam (on Python)	20%	Due at 23:59 on the Sunday of Week 8 (open-book).
Project (code, report, and video)	50%	Due at 23:59 on the Tuesday of Week 12.

DataCamp Courses

As part of our course, you have been given access to all DataCamp courses. As part of our assessment, you will be assigned four DataCamp courses to go through. Each course usually takes 3–4 hours. You are also free to do any other courses on DataCamp. Your DataCamp classroom access will expire at the end of January 2024.

Python Assignments

The two Python assignments will be short coding questions or short qualitative questions on the techniques we covered in class. You are welcome to discuss these with your classmates but your final submission must be individual work.

Take-home Exam

The take-home exam will consist of similar questions but these must be tacked individually without discussion.

Group Project

The group project will be the final portion of the assessment. Please form groups of four people and let me know the group compositions as soon as you can. The project can be on any one of the following topics.

- A work-related data analysis project. In this case, there should be a few components to the proposal:
  - A problem statement and the business decision you hope to achieve a summary of the data.
  - A proposed data collection technique.
  - A proposed method of analysis.
  - Some thoughts on how the project might fail and how to fix it.
- A data analysis (not just machine learning) of a publicly available data set.
  - Original python code.
  - Original data exploration, methods of analysis.
  - New methods that we did not cover in class.
- A review of a particular data science technique or domain, e.g.,
  - Support Vector Machines
  - Computer Vision Techniques
  - Robot Operating System (ROS), etc.

As part of the project, you will need to submit a 6-page report (single-spaced, 12-pt font) to Canvas and a 15-minute video presentation to Microsoft Stream. Your presentation will be peer-reviewed and discussed in the Week 13 lecture period.

Project Assessment

For the project submission, you will have to do the following tasks:

1. Create a 15-minute video presentation and upload it to Microsoft Stream.
2. Upload a zip file consisting of your analysis and output to Canvas. The zip file should contain:
   - A Jupyter notebook with your code.
   - A pdf report (6-page limit, 12-pt font).
   - Data set you used (if possible and not confidential).
3. Grade your peers' projects according to three criteria, each out of 10. Upload this excel sheet to Canvas.
4. Watch your peers' videos and post questions about them on the Microsoft Stream site.
5. Answer questions that others posted about your project during our Week 13 lecture.

Marks Allocation

The project contributes 50 marks to your final grade. Here is a breakdown of those marks:

1. Report quality (10 marks, group grade): A good report will

- be well-organized, with clear and necessary sectioning, including a references section,
- use figures and tables judiciously to convey the results, and
- be free of spelling/grammar mistakes.

2. Presentation quality (10 marks, group grade): A good presentation would
   - be well-organized,
   - use visualizations and tables, and
   - be well-paced.
3. Code quality (10 marks, group grade): Good code would capitalize on Python's strengths.
   - There would be good use of iterators (instead of for loops), lambda functions and self-defined functions.
   - There would be little duplication of code throughout the notebook.
   - There would also be comments to help the reader along.
4. Analysis quality (10 marks, group grade):
   - Good analysis is iterative. We will look for evidence of that, even just verbal. If you need to leave this part out of the video and report, that is understandable. However, please keep it in the notebook.
   - New techniques (that you read up on) used, and used correctly.
   - Good understanding of techniques used.
5. Grading & Questions (5 marks, individual grade):
   - We will take a look at the questions you posted. Good, insightful questions will get you good marks in this section.
6. Clarifications (5 marks, group grade):
   - On the final call, you will have a chance to answer/clarify the questions from your classmates.

Project Titles/Groupings

Ideally, I would prefer you to work on a work-related problem. However, I understand if confidentiality issues prevent this. Please form groups as soon as you can and begin working on your projects. If you need help forming a group, or need ideas on titles or data sets, please let me know and we can discuss with you.

Our Week 6 lecture is reserved for project proposal discussions. This is when I will meet with your groups individually to talk about your project. If it does not have sufficient scope, we will discuss ways to increase it.