PS0002 Introduction to Data Science and Artificial Intelligence

Academic Year	AY2019/2020	**Semester**	1
Course Coordinator	Xiang Liming (lmxiang@ntu.edu.sg)		
Course Code	PS0002		
Course Title	Introduction to Data Science and Artificial Intelligence		
Pre-requisites	PS0001 Introduction to Computational Thinking		
No of AUs	3		
Contact Hours	Lecture: 26, Lab&tutorial: 23		
Proposal Date	2018		

Course Aims

In this era of information, vast amounts of new data are produced every day from various fields including scientific research, healthcare, industry and service processes. Using data effectively and extracting meaningful insights from data can significantly improve efficiencies, cut costs and add more value to organizations. This course aims to provide you with an understanding of basic techniques for data analysis, machine learning and dimension reduction for big data, and expose you to hands-on computational tools that are fundamental for data science. Besides supervised and unsupervised learning, another fundamental technology of Artificial Intelligence – reinforcement learning will also be introduced, including Markov decision process and Q- learning.

Suited for anyone from different backgrounds, this course will show you how you could apply various methods to data examples and case studies from both research and industrial sources in the Singapore context.

Intended Learning Outcomes (ILO)

By the end of the course, you should be able to:

1. Identify interesting data-driven questions in your respective field of study.
2. Formulate meaningful study problems that you want to explore.
3. Collect/extract relevant data, visualize and perform exploratory analysis on data.
4. Perform machine learning models to extract meaningful insights from data.
5. Implement above techniques with R.
6. Present your analysis results and problem solution via an engaging written communication.

Course Content

1. What is data science
    - Why it is essential?
    - Analytic thinking
2. Study problem formulation
    - Data acquisition
    - Data wrangling

- Coding essentials with R

3. Exploratory data analysis
   - Data visualization
   - Statistical insight
4. Machine learning models for prediction
   - Linear regression
5. Machine learning models for classification
   - Linear classifier
   - Logistic regression
   - Support vector machine
6. Machine learning for clustering
   - K-means method
   - Other clustering technics
7. Dimension reduction
   - Curse of dimensionality
   - Principal component analysis
8. Reinforcement learning and AI
   - Motivating examples
   - Markov decision process
   - Q-learning
9. The state of art
   - Neural networks
   - Deep learning

Assessment (includes both continuous and summative assessment)

Component	Course ILO Tested	Related Programme LO or Graduate Attributes	Weighting	Team/Individual	Assessment rubrics
1. Group Project	1-6	Competence, Communication, Civic-mindedness, Character, Creativity.	20%	Team	Appendix 2
2. Lab assignments	3-6	Competence	20%	Team	Appendix 1
3. Quiz (in Tutorial)	2-5	Competence, Creativity, Communication.	10%	Individual	Point-based marking (not rubrics based)
4. Written Examination	2-6	Competence, Creativity, Communication.	50%	Individual	Point-based marking

					(not rubrics based)
Total			*100%*		

Formative feedback

You will receive written and verbal feedback from the lecturer for Components 2 & 4.

You will receive summative group feedback on the group project in component 1.

Learning and Teaching approach

Approach	How does this approach support you in achieving the learning outcomes?
Lecture	Lectures will deliver the theoretical knowledge required to understand various components involved in data analysis and machine learning process.
Lab & tutorial	Labs sessions will: • Demonstrate practical applications of data science techniques in various application fields. • Enable you to code using R and address any coding issues.

Reading and References

The main references/textbooks relevant to the course materials are:

1. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2014). *An Introduction to Statistical Learning: with Applications in R*, Springer New York. Available at http://www-bcf.usc.edu/~gareth/ISL/ . ISBN 978-1-4614-7137-0
2. Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference and Prediction,* Springer. https://web.stanford.edu/~hastie/ElemStatLearn. ISBN 978-0-387-84857-0
3. Kabacoff, Robert (2011). *R in Action: Data analysis and graphics with R*, Manning. ISBN-13: 978-1617291388
4. Stuart Russell and Peter Norvig (2016) *Artificial Intelligence: A Modern Approach, 3rd edition.* Pearson. ISBN-13: 978-0136042594

Course Policies and Student Responsibilities

(1) General

You are expected to complete all assigned pre-class readings and activities, attend all seminar classes punctually and take all scheduled assignments and tests by due dates. You are expected to take responsibility to follow up with course notes, assignments and course related announcements for seminar sessions they have missed. You are expected to participate in all seminar discussions and activities.

(2) Absenteeism

Absence from class without a valid reason will affect your overall course grade. Valid reasons include falling sick supported by a medical certificate and participation in NTU's approved activities supported by an excuse letter from the relevant bodies.

If you miss a lecture, you must inform the course instructor via email prior to the start of the class.

(3) Absence Due to Medical or Other Reasons

If you are sick and not able to attend a quiz or midterm, you have to submit the original Medical Certificate (or another relevant document) to the administration to obtain official leave. In this case, the missed assessment component will not be counted towards the final grade. There are no make-up quizzes or make-up midterm.

Academic Integrity

Good academic work depends on honesty and ethical behaviour. The quality of your work as a student relies on adhering to the principles of academic integrity and to the NTU Honour Code, a set of values shared by the whole university community. Truth, Trust and Justice are at the core of NTU's shared values.

As a student, it is important that you recognize your responsibilities in understanding and applying the principles of academic integrity in all the work you do at NTU. Not knowing what is involved in maintaining academic integrity does not excuse academic dishonesty. You need to actively equip yourself with strategies to avoid all forms of academic dishonesty, including plagiarism, academic fraud, collusion and cheating. If you are uncertain of the definitions of any of these terms, you should go to the academic integrity website for more information. Consult your instructor(s) if you need any clarification about the requirements of academic integrity in the course.

Collaboration is encouraged for your homework because peer-to-peer learning helps you understand the subject better and working in a team trains you to better communicate with others. As part of academic integrity, crediting others for their contribution to your work promotes ethical practice.

You **must write up your solutions by yourself and understand anything that you hand in.**

If you do collaborate, **you must write on your solution sheet the names of the students you worked with. If you did not collaborate with anyone, please explicitly write, “No collaborators."** Failure to do so constitutes plagiarism.

Use of materials outside the course is strongly discouraged. If you use outside source, you must reference it in your solution.

Course Instructors

Instructor	Office Location	Phone	Email
Xiang Liming	SPMS MAS04-11	65137451	lmxiang@ntu.edu.sg

Planned Weekly Schedule

Week	Topic	Course ILO	Readings/ Activities
1	What is data science? Why it is essential? Analytic thinking	1	Readings include multiple case studies with data science applications in various fields of study. Activity: after reading, students will be asked to describe the application scenarios aligned with their major.
2	Study problem formulation, Data acquisition, Data wrangling	1,2,3	Readings: Lecture notes, reference book of Kabacoff (2011). Lab&tutorial starts. Show data manipulating process.
3	Coding essentials with R	1,2,3,5	Readings: Lecture notes, reference book of Kabacoff (2011). Demo of R programming essentials.
4	Data visualization, Exploratory data analysis	1,2,3,5	Readings: Lecture notes, reference book of Kabacoff (2011).

			Activity: Lab Assignment 1
5	Statistical insight	1,2,3,5	Readings: Lecture notes, reference books of James et al (2014) and Hasti et al (2017). Demo of standard statistical analysis tools.
6	Machine learning models for prediction, Linear regression	1,2,3,4,5	Readings: Lecture notes, reference books of James et al (2014) and Hasti et al (2017). Demo of linear regression analysis in R
7	Machine learning models for classification: Linear classifier, Logistic regression	1,2,3,4,5	Readings: Lecture notes, reference books of James et al (2014) and Hasti et al (2017). Show the use of R package MASS, LDA and GLM. Quiz will be conducted.
8	Support vector machine	1,2,3,4,5	Readings: Lecture notes, reference books of James et al (2014) and Hasti et al (2017). Demo of the SVR model with R package **e1071.**
9	Machine learning for clustering: K-means method and other clustering techniques	1,2,3,4,5	Readings: Lecture notes, reference books of James et al (2014) and Hasti et al (2017).

			Group project will be assigned. Demo of clustering tools in R packages “cluster” “mcluster”.
10	Dimension reduction: Curse of dimensionality, Principal component analysis	1,2,3,4,5	Readings: Lecture notes, reference books of James et al (2014) and Hasti et al (2017). Demo of PCA case studies
11	Reinforcement learning and AI	1,2,3,4,5	Readings: Lecture notes, reference books of
12	The state of art, Neural networks, Deep learning	1,2,3,4,5,6	Readings: Lecture notes, reference books of James et al (2014) and Hasti et al (2017). Show some real data examples
13	Written communication for data analysis results	1,2,3,4,5,6	Readings: Lecture notes; various case studies and corresponding sample papers

Appendix 1: Assessment Criteria for Lab Assignments
Standards Criteria

Levels of Performance	Criteria Description
A+ (Exceptional) A (Excellent)	Provides clear, efficient, working and well-documented code; evidence of programing understanding and concern for code efficiency beyond getting correct solution. Demonstrated ability to develop multiple approaches to programming task, and understanding of their respective advantages.
A- (Very good) B+ (Good)	Provides clear, efficient, working and well-documented code; evidence of programing understanding.
B (Average) B- (Satisfactory)	Working but limited documentation of code.

C+ (Marginally satisfactory)	
C (Bordering unsatisfactory) C- (Unsatisfactory)	Write the code with lots of help from TA and instructor. Limited code documentation or demonstration of conceptual understand.
D (Deeply unsatisfactory) F (0-44)	Lack of demonstrated conceptual understanding. Non-functional code.

Appendix 2: Assessment Criteria for Group Project
Standards Criteria

Levels of Performance	Criteria Description
A+ (Exceptional) A (Excellent)	Provides clear and meaningful study questions; appropriate methods for data presentation, manipulating and exploration; efficient, working and well-documented code; evidence of programing understanding and concern for code efficiency beyond getting correct solution. Takes an original approach to the questions; very well structured reports with good interpretations of results; evidence of excellent ability to apply knowledge taught in the course while thinking outside the box; provides clear, efficient, working and well-documented code Clearly identifies, illustrates and critically examines implications of the project in wider context of society. Provide source acknowledgement in standard citation format. All references and citations are present and correctly written.
A- (Very good) B+ (Good)	Takes a conventional approach to the question; good interpretation of results; evidence of ability to apply knowledge taught in the course; provides clear, efficient, working and well-documented code. Describes conventional links between project and wider context of society with clear illustrations, or identifies and examines implications of the project in the wider context of society. Provides source acknowledgement in standard citation format. One or two references or citations missing or incorrectly written.
B (Average) B- (Satisfactory) C+ (Marginally satisfactory)	Takes a conventional approach to the question; limited interpretation of results; evidence of some (but not significant) ability to apply knowledge taught in the course; working but limited documentation of code. States conventional links between project and wider context of society without clear illustrations, or acknowledges obvious implications of the project on the wider context of the society.

	Provides minimal source acknowledgement. Some information does not contain a citation.
C (Bordering unsatisfactory) C- (Unsatisfactory)	Limited understanding of process; incorrect or miss-interpreted results; limited evidence of ability to apply knowledge taught in the course. Non-functional or limited code documentation. Makes some weak connections or missed some obvious implications of the project and the wider context of society. Many references and citations are missing. Format has technical errors or is presented in inconsistent styles.
D, F (Deeply unsatisfactory)	Inadequate in addressing the question; incorrect and/or miss-interpretation of results; lacks structure and focus, and is mostly or wholly off topic; inadequate capacity to apply knowledge taught in the course; non-functional code. OR failure to submit the report. Makes little to no connection between the project and the wider context of society, or missed some obvious negative implications pf the project on the wider context of society. References and citation errors detract significantly from paper. Little or no acknowledgment of sources.