COURSE OUTLINE

Approval status: cleared at TLPD level

Course Title	**Introduction to Data Science and Artificial Intelligence**		
Course Code	**PS0002**		
Offered	Study Year 1, Semester 2		
Course Coordinator	Xiang Liming (Assoc Prof)	lmxiang@ntu.edu.sg	6513 7451
Pre-requisites	PS0001 or CZ1003		
AU	3		
Contact hours	Lectures: 26, Tutorials: 13, Laboratories: 10		
Approved for delivery from			
Last revised	20 Feb 2020, 11:33		

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

Upon successfully completing this 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

What is data science? Why it is essential? Analytic thinking

Study problem formulation, Data acquisition, Data wrangling

Coding essentials with R

Data visualization, Exploratory data analysis

Statistical insight

Machine learning models for prediction, Linear regression

Machine learning models for classification: Linear classifier, Logistic regression

Support vector machine

Machine learning for clustering: K-means method

Other clustering technics

Dimension reduction: Curse of dimensionality, Principal component analysis

The state of art, Neural networks, Deep learning

Written communication for data analysis results

Assessment

Component	Course ILOs tested	SPMS-MAS Graduate Attributes tested	Weighting	Team / Individual	Assessment Rubrics
Continuous Assessment					
Laboratories					
Assignment	3, 4, 5, 6	**1.** a, b, c, d	20	team	See Appendix for rubric
Tutorials					
Project	1, 2, 3, 4, 5, 6	**1.** a, b, c, d **2.** a, b, c **3.** a, b **4.** a **5.** a	20	team	See Appendix for rubric
Quiz	2, 3, 4, 5	**1.** a, b, c, d **2.** a, b, c **3.** a, b	10	individual	See Appendix for rubric
Examination (2 hours)					
Short Answer Questions	2, 3, 4, 5, 6	**1.** a, b, c, d **2.** a, b, c **3.** a, b	50	individual	See Appendix for rubric
		Total	**100%**		

These are the relevant SPMS-MAS Graduate Attributes.

1. **Competence**
   a. Independently process and interpret mathematical theories and methodologies, and apply them to solve problems
   b. Formulate mathematical statements precisely using rigorous mathematical language
   c. Discover patterns by abstraction from examples
   d. Use computer technology to solve problems, and to communicate mathematical ideas
2. **Creativity**
   a. Critically assess the applicability of mathematical tools in the workplace
   b. Build on the connection between subfields of mathematics to tackle new problems
   c. Develop new applications of existing techniques
3. **Communication**
   a. Present mathematics ideas logically and coherently at the appropriate level for the intended audience
   b. Work in teams on complicated projects that require applications of mathematics, and communicate the results verbally and in written form

4. **Civic-mindedness**

   a. Develop and communicate mathematical ideas and concepts relevant in everyday life for the benefits of society

5. **Character**

   a. Act in socially responsible and ethical ways in line with the societal expectations of a mathematics professional, particularly in relation to analysis of data, computer security, numerical computations and algorithms

Formative Feedback

You will receive written and verbal feedback from the lecturer for lab assignments and written examinations.

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

Learning and Teaching Approach

Lectures (26 hours)	Lectures will deliver the theoretical knowledge required to understand various components involved in data analysis and machine learning process.
Tutorials (13 hours)	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.
Laboratories (10 hours)	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/ .
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
3. Kabacoff, Robert (2011). R in Action: Data analysis and graphics with R, Manning.
4. Stuart Russell and Peter Norvig (2016) Artificial Intelligence: A Modern Approach, 3rd edition. Pearson

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.

Course Instructors

Instructor	Office Location	Phone	Email
Xiang Liming (Assoc Prof)	SPMS MAS04-11	6513 7451	lmxiang@ntu.edu.sg

Planned Weekly Schedule

Week	Topic	Course ILO	Readings/ Activities
1	The state of art, Neural networks, Deep learning	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	Dimension reduction: Curse of dimensionality, Principal component analysis	1, 2, 3	Readings: Lecture notes, reference book of Kabacoff (2011). Lab&tutorial starts. Show data manipulating process.
3	Other clustering technics	1, 2, 3, 5	Readings: Lecture notes, reference book of Kabacoff (2011). Demo of R programming essentials.
4	Support vector machine	1, 2, 3, 5	Readings: Lecture notes, reference book of Kabacoff (2011). Activity: Lab Assignment 1
5	Machine learning for clustering: K-means method	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 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). Demo of linear regression analysis in R
7	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). Show the use of R package MASS, LDA and GLM. Quiz will be conducted.
8	Coding essentials with R	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	Data visualization, Exploratory data analysis	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	Statistical insight	1, 2, 3, 4, 5	Readings: Lecture notes, reference books of James et al (2014) and Hasti et al (2017).
11	What is data science? Why it is essential? Analytic thinking	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
12	Study problem formulation, Data acquisition, Data wrangling	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 Rubrics

Rubric for Laboratories: Assignment (20%)

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) C+ (Marginally satisfactory)	Working but limited documentation of code.
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.

Rubric for Tutorials: Project (20%)

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
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.
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.
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.
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.

Rubric for Tutorials: Quiz (10%)

Point-based marking (not rubrics based)

Rubric for Examination: Short Answer Questions (50%)

Point-based marking (not rubrics based)