COURSE OUTLINE: MH4510

Course Title	**Statistical Learning and Data Mining**		
Course Code	**MH4510**		
Offered	Study Year 4, Semester 1		
Course Coordinator	Fedor Duzhin (Dr)	fduzhin@ntu.edu.sg	6513 7469
Pre-requisites	MH2500, MH3500, MH3510, MH3511		
AU	4		
Contact hours	Technology-enhanced Learning: 32, Laboratories: 24, Tutorials: 12		
Approved for delivery from	AY 2021/22 semester 1		
Last revised	21 Jul 2021, 13:04		

Course Aims

This course gives an overall view of the modern statistical/machine learning techniques for mining massive datasets, ranging from generalized linear models, over model selection, to the state-of-the-art techniques like LASSO, neural networks, etc. This course will not only discuss individual algorithms and methods, but also tie principles and approaches together from a theoretical perspective. Moreover, students can gain a hands-on experience through team project. This course equips students with the necessary skills for being a data analyst.

Intended Learning Outcomes

Upon successfully completing this course, you should be able to:

1. Resolve data mining problems with various modern statistical techniques;
2. Summarise the strengths and shortcomings of different techniques;
3. Evaluate learning methods statistically and recommend the optimal one for applications;
4. Implement the modern statistical techniques with statistical software such as R.
5. Work in a team, i.e., effectively communicate on distribution of roles and work tasks and help other students with their tasks.
6. Present project findings in the form of a report written at the appropriate level for the intended audience.

Course Content

Main concepts (supervised vs unsupervised learning, regression vs classification, overfitting, training, validation and test sets etc.). Overview of linear regression. KNN.

Linear regression: detailed view. Gradient descent. Logistic regression.

Resampling methods (cross-validation, bootstrap etc.)

Regularization (lasso and ridge regression). Best subset selection.

Tree-based methods

Support vector machines

Artificial Neural Networks

Principal Component Analysis. Clustering algorithms (K-means & hierarchical clustering).

Word embeddings for natural language processing. GloVe

Association Analysis

Assessment

Component	Course ILOs tested	SPMS-MAS Graduate Attributes tested	Weighting	Team / Individual	Assessment Rubrics
Continuous Assessment					
Laboratories					
Assignment	1, 6	**1.** d	10	individual	See Appendix for rubric
Tutorials					
Project	1, 2, 3, 4, 5, 6	**1.** a, b, d **2.** a, b, c **3.** a, b **4.** a **5.** a	40	both	See Appendix for rubric
Test	4, 5	**1.** a, c **2.** b	10	individual	See Appendix for rubric
Mid-semester Quiz					
Short Answer Questions	3, 4, 5	**1.** a, b, c **2.** b	20	individual	See Appendix for rubric
Short Answer Questions 1	3, 4, 5	**1.** a, b, c **2.** b	20	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

Course instructors will provide constant feedback to students on group work in tutorials and on lab handouts. Besides, students are welcome to contact the instructors at any time to get advice on their group project. There will also be two structured feedback sessions where the

instructors grade the course project as if it were final, but the grade does not count towards the course mark - after submitting the proposal and after the project presentation.

In tutorials, students will also get feedback from their peers.

Learning and Teaching Approach

Technology-enhanced Learning (32 hours)	Video lectures with theoretical material and interactive quizzes to help in understanding course material.
Laboratories (24 hours)	Computer labs where students learn practical skills or programming in R, data manipulation, and implementing machine learning models by going through lab handouts.
Tutorials (12 hours)	Tutorials consist of practice questions aimed at helping students to understand theoretical material.

Reading and References

Textbook:
1. G. James, D. Witten, T. Hastie, and R. Tibshirani. (2013) An Introduction to Statistical Learning - with Applications in R. Springer. ISBN 978-1-4614-7138-7

Reference:
2. T. Hastie, R. Tibshirani, and J. Friedman. (2009) The Elements of Statistical Learning - Data Mining, Inference, and Prediction (2nd Ed.). Springer. ISBN 978-0-387-84858-7
3. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Course Policies and Student Responsibilities

(1) General

You are expected to attend all tutorials punctually and submit all quizzes. You are also expected to complete all weekly lab reports by due dates and lab reports should be your own work (but you are not required to attend lab sessions). You are expected to complete and present a team project. You are expected to attend both midterm tests and do them individually. You are expected to take responsibility to follow up with course notes, assignments and course-related announcements.

(2) Lab assignments

All lab assignments equally contribute to the 10% of total score. Late submissions will be subject to mark deduction.

(3) Team Project

The team project accounts for 40% of total score. Each team needs to submit a project proposal, a deck of slides for presentation, and final project report by the due dates. Moreover, each team is required to present its project in the last teaching week.

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
Fedor Duzhin (Dr)	SPMS-MAS-05-23	6513 7469	fduzhin@ntu.edu.sg

Planned Weekly Schedule

Week	Topic	Course ILO	Readings/ Activities
1	Introduction. Main concepts (supervised vs unsupervised learning, regression vs classification, overfitting, training, validation and test sets etc.). Overview of linear regression. KNN.	1	Reading #1 Chapters 1
2	Linear regression again. Gradient descent. Logistic regression.	1, 2	Reading #1 Chapters 2, 3 Labs & tutorials
3	Resampling methods (cross-validation, bootstrap etc.)	1, 2, 4	Reading #1 Chapters 4 Labs & tutorials
4	Regularization (lasso and ridge regression). Best subset selection.	1, 2, 4	Reading #1 Chapters 5, 6 Labs & tutorials
5	Tree-based methods	1, 3	Reading #1 Chapters 8 Labs & tutorials
6	Support vector machines	1, 2	Reading #1 Chapters 9 Labs & tutorials
7	Artificial Neural Networks	1, 2, 4	Reading #2 Chapters 11.1-11.8, Labs & tutorials MIDTERM TEST 1
8	Principal component analysis. Clustering algorithms (K-means & hierarchical clustering)	1, 2, 4	Reading #1 Chapter 10 Labs & tutorials
9	Word embeddings. GloVe	1, 2, 4	Reading #3 Labs & tutorials
10	Association Rules	1, 2, 4	Reading #2 Chapter 14.2 Labs & tutorials
11	Project work	1, 2, 4	Labs & tutorials
12	Project work	1, 2	MIDTERM TEST 2
13	Group Project Presentation	1, 2, 3, 4	

Appendix 1: Assessment Rubrics

Rubric for Laboratories: Assignment (10%)

Completely incorrect or not submitted - 0%

Barely started - 25%

Half-correct - 50%

Correct with errors - 75%

Completely correct - 100%

Rubric for Tutorials: Project (40%)

		Unsatisfactory	Poor	Average	Good	Excellent
	MAX	0	0.25	0.5	0.75	1
Problem choice	10	The problem was suggested to the students by the course instructor.	The students seem to have chosen a random problem from an online source (i.e., Kaggle) without researching its background.	The students made a minimal effort to research the problem's background.	There is a comprehensive explanation of the problem background (literature review).	There is a comprehensive explanation of the problem background (literature review). The problem is either particularly challenging (e.g., requires merging several datasets) or particularly important for society (e.g., climate change or cure for cancer).
Exploratory data analysis	5	Data exploration is not in the report at all.	There is an attempt to do data exploration, but it does not help the reader to understand the data at all.	Either the summary or the visualization is missing.	Data summary and visualizations are there but are not completely perfect. For example, an ordinal variable may be called categorical or there may not be a histogram of some important predictor. Alternatively, there may be too much visualization to the extent of distracting the reader from the problem. Process of data cleaning may not be explained.	The data are properly visualized and summarized in a way that helps to get meaningful insights of the data. Process of data cleaning is fully explained.
Feature engineering	10	The is no variable selection	The response variable is identified incorrectly. Feature selection is not done at all.	The response variable is clearly identified but feature selection is not done at all.	The response variable and the features are clearly identified but the justification is incomplete.	The response variable and the features are clearly identified and variable selection is properly explained
Modelling	15	Fewer models than required are included in the report.	3 models (for 3-student teams) or 4 models (for 4-student teams) have been chosen. Model	3 models (for 3-student teams) or 4 models (for 4-student teams) have been chosen.	3 models (for 3-student teams) or 4 models (for 4-student teams) have been chosen.	3 models (for 3-student teams) or 4 models (for 4-student teams) have been chosen.

			parameters are selected randomly and overfitting has not been addressed. Models may not be chosen correctly (i.e., doing a linear regression to predict a binary outcome).	There are serious inconsistencies in either parameter selection or dealing with overfitting.	Either the explanation of parameter choice is incomplete or overfitting has not been addressed properly.	The report carefully explains reasons for the choice of model parameters and addresses overfitting.
Coding	10	There is no R code at all, it contains syntax errors, or it does not do what it is supposed to.	The code does what it is supposed to do but is either very inefficient (e.g., because of an out-of-place recursion) or incomprehensible. This includes breaking loops, creating matrices by assigning values entry by entry etc.	The code does what it is supposed to do, but may be inefficient because of lack of vectorization, unnecessary loops or control statements, nested loops or failure to use appropriate native functions.	The code does what it is supposed to do, but may be a bit inefficient. Sometimes the code may be hard to understand because of lack of commenting or obscure variable names.	The code is efficient, easy to understand and it does what it is supposed to do. It makes use of existing R libraries and avoids loops whenever possible. Important variables have meaningful names and there are comments explaining the code whenever it is needed.
Material beyond the course syllabus	25	The project is completely within the course syllabus.	The team has learned some material that is not in the course syllabus, but the progress was minimal.	The team has gone way beyond the course syllabus by reading research literature on data mining and applying results to their project. Extra research affected one aspect of the project (i.e., one of the models they used is not covered in class). The team showed good understanding of the theory.	The team has gone way beyond the course syllabus by reading research literature on data mining and applying results to their project. Extra research affected more than one aspect of the project. The team showed good understanding of the theory.	The team has gone way beyond the course syllabus by reading research literature on data mining and applying results to their project. Extra research affected several aspects of the project. The team showed deep understanding of the theory.
Presentation	10	The team was not able to prepare the presentation on class of week 13.	The number of slides exceeds 4. The presentation is very messy and the speaker doesn't seem to be familiar with the scope of the project.	Not more than 4 slides that may be prepared on Powerpoint / Google drive. The presentation may be very messy and by far exceed 10	Not more than 4 slides professionally typed in R Notebook. The presentation is very clear but may last slightly more than 10 minutes. Students may	Not more than 4 slides professionally typed in R Notebook. The presentation is very clear and lasts not more than 10 minutes. All team members

				minutes. Some team members may not be familiar with the structure of the project.	not be familiar with their teammates' contribution.	are able to answer questions about the project
Final report	15	There is no report at all, i.e., it is either not submitted or submitted as commented R code rather than a structured R Notebook published knitted into a PDF.	The report is not written in full English sentences and is hard to follow. There may be serious mistakes in style, grammar, and structure that make understanding the report a challenging task. There may not be an abstract or the reference list and the length of the report may by far exceed 5000 words.	The report is mostly written in full English sentences, but is not very easy to follow due to mistakes in style or grammar. The structure of the report is not completely as required, i.e., the abstract may be in the end or the number of words exceeds 5000. The report may be produced in MS Word. Sources may not be properly cited.	The report is written in full English sentences, is properly structured and is easy to follow. The report has been prepared in R Notebook and knitted into PDF. The report does not exceed 5000 words. There are mistakes in style or grammar that don't make the report too hard to understand, such as small errors in LaTeX equations or occasional grammar errors. The abstract may exceed 200 words or may not be informative.	The report is written in full English sentences, is properly structured and is easy to follow. The report has been prepared in R Notebook. The length of the report does not exceed 5000 words (including R codes). The abstract is not longer than 200 words and summarizes all important findings. All the sources are appropriately cited. There are no mistakes in style or grammar.
Total	100					
		The final project mark (40% of the course total) is calculated as follows: 0.35 * team score + 0.55 * individual contribution * team score + 0.05 * mark for consistency of scores one submits with the team's opinion + 0.05 * mark for writing good peer reviews				
		WARNING: If your individual score for the project exceeds 100, then it will be changed to exactly 100 (in other words, it is impossible to get more than 40% of the total course mark from the project). If individual contribution to the team project is 0, then your score for the project will be changed to 0.				

Remark: 35% of the total score for the project is due to team work and 65% is due to individual work.

Rubric for Tutorials: Test (10%)

5 multiple choice questions

Rubric for Mid-semester Quiz: Short Answer Questions (20%)

Completely incorrect - 0%

Some right ideas - 25%

Half-correct - 50%

Correct with errors - 75%

Completely correct - 100%

Rubric for Mid-semester Quiz: Short Answer Questions 1 (20%)

Completely incorrect - 0%

Some right ideas - 25%

Half-correct - 50%

Correct with errors - 75%

Completely correct - 100%