Annexe A: New/Revised Course Content in OBTL+ Format

Course Overview

Expected Implementation in Academic Year (New format)	AY2026-2027
Semester/Trimester/Others (specify approx. Start/End date)	Semester 1
Course Author * Faculty proposing/revising the course	Fedor Duzhin
Course Author Email	fduzhin@ntu.edu.sg
Course Title	Statistical Learning and Data Mining
Course Code	MH4510
Academic Units	4
Contact Hours	68
Research Experience Components	Not Applicable

Course Requisites (if applicable)

Pre-requisites	MH2500 and MH3510 and MH3500 OR MH2500 and MH3510 and MH2510
Co-requisites	
Pre-requisite to	
Mutually exclusive to	
Replacement course to	
Remarks (if any)	

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, you can gain a hands-on experience through team project. This course equips you with the necessary skills for being a data analyst.

Course's Intended Learning Outcomes (ILOs)

Upon the successful completion of this course, you (student) would be able to:

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

Reading and References (if applicable)

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).

NOTE: The above listing comprises the foundational readings for the course and more up-to-date relevant readings will be provided when they are available.

Planned Schedule

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

Week or Session	Topics or Themes	ILO	Delivery Mode	Activities	Readings
12	Project work	1,2	In-person	MIDTERM TEST 2	Some readings
13	Group Project Presentation	1,2,3,4	In-person	Group presentations	Some readings

Learning and Teaching Approach

Approach	How does this approach support you in achieving the learning outcomes?
Technology-enhanced Learning	Video lectures with theoretical material and interactive quizzes to help in understanding course material.
Laboratories	Computer labs where you learn practical skills or programming in R, data manipulation, and implementing machine learning models by going through lab handouts.
Tutorials	Tutorials consist of practice questions aimed at helping you to understand theoretical material.

Assessment Structure

Assessment Components (includes both continuous and summative assessment)

No.	Component	ILO	Related PLO or Accreditation	Weightage	Description of Assessment Component	Team/Individual	Rubrics	Level of Understanding
1	Continuous Assessment (CA): Others(Laboratories - Assignment)	1, 6		10	Weekly lab reports to be completed and submitted as an R notebook knitted to PDF. Every one of them is very simple, only 1% of the total course mark, and is graded out of "Not done / Poor / Average / Good / Excellent" scale.	Individual	Holistic	Multistructural

No.	Component	ILO	Related PLO or Accreditation	Weightage	Description of Assessment Component	Team/Individual	Rubrics	Level of Understanding
2	Continuous Assessment (CA): Project(Tutorials - Projects)	1, 2, 3, 4, 5, 6		40	This is the main assessment component - group project. Note that even though it is a group project, your score will be the product of the team score and your individual contribution to the project.	Team	Analytic	Extended Abstract
3	Continuous Assessment (CA): Test/Quiz(Tutorials - test)	4, 5		10	Weekly MCQ quizzes.	Individual	Analytic	Multistructural
4	Continuous Assessment (CA): Test/Quiz(Mid-semester Quiz - Short Answer Questions 2)	3,4,5		20	Theoretical midterm exam based on material of Lectures 6-10. Mixture of short answer questions and MCQ.	Individual	Analytic	Relational

No.	Component	ILO	Related PLO or Accreditation	Weightage	Description of Assessment Component	Team/Individual	Rubrics	Level of Understanding
5	Continuous Assessment (CA): Test/Quiz(Mid-semester quiz - Short Answer Questions 1)	3,4,5		20	Theoretical midterm exam based on material of Lectures 1-5. Mixture of short answer questions and MCQ.	Individual	Analytic	Relational

Description of Assessment Components (if applicable)

Formative Feedback

Course instructors will provide constant feedback to you on group work in tutorials and on lab handouts. Besides, you are welcome to contact the instructors at any time to get advice on your 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.

NTU Graduate Attributes/Competency Mapping

This course intends to develop the following graduate attributes and competencies (maximum 5 most relevant)

Attributes/Competency	Level
Collaboration	Advanced
Communication	Advanced
Curiosity	Intermediate
Digital Fluency	Advanced
People Orientation	Basic

Course Policy

Policy (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. On the use of technological tools (such as Generative AI tools), different courses / assignments have different intended learning outcomes. Students should refer to the specific assignment instructions on their use and requirements and/or consult your instructors on how you can use these tools to help your learning. Consult your instructor(s) if you need any clarification about the requirements of academic integrity in the course.

Policy (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.

Policy (Absenteeism)

You are allowed to be absent from up to 2 tutorials without a valid reason. For more absences, you are required to produce a valid leave of absence, failing which you will get a 0 score for that tutorial. If you are absent from a midterm test, you need to produce a valid leave of absence - in that case, a make-up test will be arranged for you in the next few days.

If you don't submit a lab report on time (every Monday), you need to contact the course instructor immediately, explain the reason and ask for an extension. A short extension (1-2 days) will be usually given to you.

If you don't contribute in the course project, you will fail the course automatically.

Policy (Others, if applicable)

Diversity and inclusion policy

Integrating a diverse set of experiences is important for a more comprehensive understanding of science.

It is our goal to create an inclusive and collaborative learning environment that supports a diversity of perspectives and learning experiences, and that honours your identities; including ethnicity, gender, socioeconomic status, sexual orientation, religion or ability.

To help accomplish this:

- If you are neuroatypical or neurodiverse, have dyslexia or ADHD (for example), or have a social anxiety disorder or

social phobia;

- If you feel like your performance in the class is being impacted by your experiences outside of class;
- If something was said in class (by anyone, including the instructor) that made you feel uncomfortable;

Please speak to your teaching team, our school pastoral officer or a peer or senior (either in-person or via email) about how we can help facilitate your learning experience.

As a participant in course discussions, you should also strive to honour the diversity of your classmates. You can do this by: using preferred pronouns and names; being respectful of others opinions and actively making sure all voices are being heard; and refraining from the use of derogatory or demeaning speech or actions.

All members of the class are expected to adhere to the NTU anti-harassment policy. if you witness something that goes against this or have any other concerns, please speak to your instructors or a faculty member.

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

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.

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	There 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. Mode parameters are selected randomly and	3 models (for 3 student teams) or 4 models (for 4-student teams) have been chosen. There are serious inconsistencies in either	3 models (for 3 student teams) or 4 models (for 4-student teams) have been chosen. Either the explanation of parameter choice is incomplete or overfitting has not	3 models (for 3 student teams) or 4 models (for 4-student teams) have been chosen. The report carefully explains reasons for the choice of model parameters

			overfitting has not been addressed. Models may not be chosen correctly (i.e., doing a linear regression to predict a binary outcome).	parameter selection or dealing with overfitting.	been addressed properly.	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 incomprehensi ble. 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 minutes. Some team members may not be familiar with the structure of the project.	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 be familiar with their teammates' contribution.	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 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 Rcode rather than a structured R Notebook published knitted into a PDF.	The report is notwritten in full English sentencesand is hard to follow. There maybe serious mistakes in style, grammar, and structure that	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	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	The report is written in full English sentences, is properly structured andis easy to follow. The report has been preparedin R Notebook.The length of the report doesnot

			make understanding thereport 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.	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.	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.	exceed 5000 words (including R codes). The abstract is not longer than 200 words and summarizes all important findings. All thesources 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 scorefor 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%