NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

COURSE CONTENT

Academic Year	2022/2023	**Semester**	2
Course Coordinator	Asst Prof. Xunyuan YIN		
Course Code	CB4247		
Course Title	Statistics & Computational inference to Big Data		
Pre-requisites	CH2010 Engineering Statistics		
No of AUs	3		
Contact Hours	26 Lecture hours and 12 tutorial hours (from TW2)		
Proposal Date	29 July 2022		

Course Aims

The advent of the big data era has highlighted great new opportunities and challenges for statistical inference in manufacturing and daily life. To embrace big data (from an industrial manufacturing perspective), there is an urgent need to truly understand the core concepts and become capable of leveraging key algorithms/techniques/methodologies pertaining to data (big-data) statistics and computational inference, which is essential for extracting useful and valuable information for informed decision-making. This course will start with the core principles of data analytics and will equip you with the statistics and computational inference (including regression, dimensionality reduction, modeling) suitable for coping with big data case scenarios. This course is expected to help students develop interpretation of easy-to-use techniques/algorithms/methods and equip the students with essential skills in addressing big data inference problems in the chemical and biomedical industries.

Intended Learning Outcomes (ILO)

Student will:

1. Understand the concept of big data, and apply concepts of probability and probability distributions. Identify the different type of statistical distribution (including normal distribution, Chi-square and *F* distribution) and describe the key characteristics of these distributions.
2. Master big data pre-processing techniques, including how to deal with the missing data, detection and processing methods of outliers, and resampling methods.
3. Revisit and apply ordinary least square (OLS) and nonlinear least-squares.
4. Learn and apply weighted least-square (WLS) methods to estimate the parameters in a regression model.
5. Learn and apply machine learning methods (e.g., principal components analysis (PCA), PCA based least squares methods, partial least squares methods
6. Learn and apply Lasso regression methods and Ridge regression methods) for big data regression.
7. Use commonly used programming-based computing platforms (e.g., MATLAB or Python) to process data, conduct regression, build regression model, and visualize and analyze the obtained results.

Course Content

Key topics taught:

1. Review of probability and probability distributions
2. Data Pre-processing for big data analytics, regression/data-driven predictive modeling
3. Fundamentals of regression (ordinary least-squares, weighted least-squares)
4. Nonlinear regression
5. Principal component analysis (PCA) for dimensionality reduction
6. PCA-based regression and Partial least-squares
7. Other variants of least-squares in the context of big data (e.g., Lasso regression and Ridge regression)
8. Applications of the methods/techniques to real-world problems for analysis and regression/modeling

Assessment (includes both continuous and summative assessment)

Component	Course LO Tested	Related Programme LO or Graduate Attributes	Weighting	Team/ Individual	Assessment rubrics
1. Final Examination (2hrs, Closed-book exam)	1, 2, 3, 4, 5, 6	EAB SLO* a, b, c, d	50%	Individual	
2. CA1: Quiz	1, 2, 3, 4	EAB SLO* a, b, c, d	20%	Individual	
3. CA2: Project	1, 2, 3, 4, 5, 6, 7	EAB SLO* a, b, c, d, e, h, j, l	30%	Individual	See Appendix 1
Total			100%		

Formative feedback

During tutorials, the instructor will articulate expected learning outcomes in detail and use examples/case studies to better illustrate the methods introduced in the previous lectures.

After each CA, the instructor will go through the problems during tutorials. Common mistakes and misunderstanding in concepts will also be addressed.

Specific feedback to the progress of project work may be returned to students via email. General feedback to project work will be published online.

Learning and Teaching approach

Approach	How does this approach support students in achieving the learning outcomes?
LECTURE	Course materials covering all the topics
TUTORIAL	12 classroom discussion sessions on tutorial questions and related topics

Reading and References

The following textbooks may be used as references.

1. D. C. Montgomery, G. C. Runger, Applied Statistics and Probability for Engineers.
2. G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning.
3. J. Warren, N. Marz. Big Data: Principles and Best Practices of Scalable Realtime Data Systems.

Course Policies and Student Responsibilities

Students are responsible for meeting all course requirements, observing all deadlines, examination times, and other course procedures.

You will be awarded ZERO mark for being absence from quizzes unless it is due to the following reasons:

- Illness (valid medical certificate is required, not from Chinese doctor)
- Passing away of immediate family member (parents, siblings or grandparents)
- Participate in an activity representing NTU (support letter from participating organization)
-

There will be no makeup given for missed quizzes. Final grade will be determined based on the participated quiz and final examination.

You are responsible for following the university regulations for final examination.

You are responsible for being on time for all lectures and tutorials. Sufficient efforts should be put into solving or attempting the tutorial problems prior to attending the respective tutorial classes.

You might be awarded an “F” for a component or expelled from the university if you are caught cheating.

You are responsible for seeking academic help in a timely fashion.

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 may 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
Yin Xunyuan	N1.2 B2-22	63168746	xunyuan.yin@ntu.edu.sg

Planned Weekly Schedule

Week	Topic	Course LO	Readings/ Activities
1	Introduction to big data analytics and probability distributions	1,	
2	Data pre-processing for big datasets	2	
3-4	Revisiting ordinary least-squares	3	
4-5	Revisiting nonlinear least-squares regression	3	
5-6	Weighted least-squares regression	4	
7	CA1	1, 2, 3, 4	
8	Introduction to principal component analysis (PCA) and PCA-based regression	5	
9	Partial least-squares regression	5	
10	Lasso regression	6	
11	Ridge regression	6	
12	CA2	1, 2, 3, 4, 5, 6, 7	
13	Applications in Engineering Problems	1, 2, 3, 4, 5, 6, 7	

Appendix 1: Assessment Criteria for the Project

Criteria	Exceed Expectations 71%-100%	Meet Expectations 41% - 70%	Meet Baseline Expectations 26% – 40%	Below Expectations 0 – 25%
Identify and formulate a good regression/predictive modeling problem, pre-process the data; explore and choose relevant and appropriate algorithm(s) to conduct regression or build predictive models (LO 1-6)	Systematically and most appropriately pre-process raw data to retain usable variables and extract the most valuable information from the raw data to facilitate further analysis and results. Truly understand the nature of the considered problem. Through exploration, choose the most appropriate regression/modeling methodology (may also explore some algorithms/concepts that are relevant to this course but not discussed in detail in lectures) for the specific problems and well justify the selection and adoption of the method using clear and concrete language. Showcase the obtained results in effective ways (visualization,	Appropriately pre-process raw data to mitigate the negative effect of missing data/poor-quality samples in the raw dataset on further analysis and results. Select appropriate regression/modeling methodology for the specific regression/modeling problems and present explanations about the selection and adoption of the method. Visualize the obtained results, discuss and interpret the obtained results, discuss the advantages and limitations of the proposed solution, and justify the conclusions based on the results. In the report, present clear figures with necessary legends and labels, and	Apply necessary algorithms/techniques to pre-process raw data so that the processed dataset may be acceptable for regression and modeling. Apply regression/modeling methodology covered in this course to the processed data to generate some regression/modeling results Present reasonable and acceptable interpretation of the results but without drawing solid contributions and presenting convincing justifications of the conclusions. Submit a complete project report, but without appropriately presenting visualization results. Not well written with ambiguous statements and typos and grammatical errors seen	Unable to formulate a well-defined regression/modeling problem. Unclear ultimate goals. Unable to apply appropriate data pre-processing and regression/modeling algorithms to pursue the objectives. No result and/or interpretation to showcase.

Criteria	Exceed Expectations 71%-100%	Meet Expectations 41% - 70%	Meet Baseline Expectations 26% – 40%	Below Expectations 0 – 25%
	tables, etc.), discuss in detail and interpret the obtained results, discuss the advantages and limitations of the adopted algorithms and the developed solution, discuss current perspectives and future directions, and justify the conclusions based on the results. In the report, present very nicely-plotted, easy to interpret figures with labels and legends, draw a schematic of the proposed solution, conduct extensive literature review, and present very well-structured project report with (almost) no grammatical errors/typos.	prepare well-written project report with minimal grammatical errors/typos.	throughout the report.	

Appendix 2: The EAB (Engineering Accreditation Board) Accreditation SLOs (Student Learning Outcomes)

a) **Engineering knowledge:** Apply the knowledge of mathematics, natural science, engineering fundamentals, and an engineering specialisation to the solution of complex engineering problems
b) **Problem Analysis:** Identify, formulate, research literature, and analyse complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
c) **Design/development of Solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for public health and safety, cultural, societal, and environmental considerations.
d) **Investigation:** Conduct investigations of complex problems using research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
e) **Modern Tool Usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations
f) **The engineer and Society:** Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
g) **Environment and Sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for the sustainable development.
h) **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
i) **Individual and Team Work:** Function effectively as an individual, and as a member or leader in diverse teams and in multidisciplinary settings.
j) **Communication:** Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
k) **Project Management and Finance:** Demonstrate knowledge and understanding of the engineering and management principles and economic decision-making, and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
l) **Life-long Learning:** Recognise the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change