City University of Hong Kong
Course Syllabus

offered by Department of Computer Science
with effect from Semester A 2015/16

Part I Course Overview

Course Title:	Fundamentals of Data Science
Course Code:	CS3481
Course Duration:	One semester
Credit Units:	3 credits
Level:	B3
Proposed Area: *(for GE courses only)*	☐ Arts and Humanities ☐ Study of Societies, Social and Business Organisations ☐ Science and Technology
Medium of Instruction:	English
Medium of Assessment:	English
Prerequisites: *(Course Code and Title)*	CS2204 Fundamentals of Internet Applications Development
Precursors: *(Course Code and Title)*	Nil
Equivalent Courses: *(Course Code and Title)*	Nil
Exclusive Courses: *(Course Code and Title)*	CS4483 Data Warehousing and Data Mining

Part II Course Details

1. Abstract

(A 150-word description about the course)

This course aims to explore the important field of data science. The syllabus covers the main techniques in statistical data modelling, and algorithms in data science, which include predictive modelling, cluster analysis, association rule mining and text mining. In addition, different applications of data science techniques in the real world such as web mining, business analytics and health informatics will be discussed.

2. Course Intended Learning Outcomes (CILOs)

(CILOs state what the student is expected to be able to do at the end of the course according to a given standard of performance.)

No.	CILOs[#]	Weighting* (if applicable)	Discovery-enriched curriculum related learning outcomes (please tick where appropriate)		
			A1	*A2*	*A3*
1.	Identify the main characteristics of different techniques in data science through observation of their operations.			✓	
2.	Perform a critical assessment of current techniques in data science.		✓	✓	
3.	Implement the main algorithms in data science in a computationally efficient way.			✓	
4.	Propose new solutions for real world information analytics problems by improving and combining current data science techniques.				
	** If weighting is assigned to CILOs, they should add up to 100%.*	100%			

[#] Please specify the alignment of CILOs to the Gateway Education Programme Intended Learning outcomes (PILOs) in Section A of Annex.

A1: Attitude
Develop an attitude of discovery/innovation/creativity, as demonstrated by students possessing a strong sense of curiosity, asking questions actively, challenging assumptions or engaging in inquiry together with teachers.

A2: Ability
Develop the ability/skill needed to discover/innovate/create, as demonstrated by students possessing critical thinking skills to assess ideas, acquiring research skills, synthesizing knowledge across disciplines or applying academic knowledge to self-life problems.

A3: *Accomplishments*
Demonstrate accomplishment of discovery/innovation/creativity through producing /constructing creative works/new artefacts, effective solutions to real-life problems or new processes.

3. Teaching and Learning Activities (TLAs)

(TLAs designed to facilitate students' achievement of the CILOs.)

Teaching pattern:

Suggested lecture/tutorial/laboratory mix: 2 hrs. lecture; 1 hr. tutorial.

TLA	Brief Description	CILO No.				Hours/week (if applicable)
		1	2	3	4	
Lecture	This course will focus on introducing the fundamental and state-of-the-art techniques in data science.	✓	✓	✓	✓	2 hrs/wk
Tutorial	Students will work on a set of problems on the principles and applications of data science, and present their solutions in the class.	✓	✓			1 hr/wk
Project	There will be two projects: The first project gives students an opportunity to implement existing algorithms in data science in a computationally efficient way. The second project allows students to create new designs for information analytics systems.			✓	✓	6 hrs/wk for 6 weeks

4. Assessment Tasks/Activities (ATs)

(ATs are designed to assess how well the students achieve the CILOs.)

Assessment Tasks/Activities	CILO No.				Weighting*	Remarks
	1	2	3	4		
Continuous Assessment: 50%						
Project 1 (Implementation of data science algorithms.)			✓		15%	
Project 2 (Application of data science algorithms to real world problems.)				✓	15%	
Quiz	✓	✓			20%	
Examination^: 50% (duration: 2 hours)						
** The weightings should add up to 100%.*					100%	

^ For a student to pass the course, at least 30% of the maximum mark for the examination must be obtained.

5. Assessment Rubrics

(Grading of student achievements is based on student performance in assessment tasks/activities with the following rubrics.)

Assessment Task	Criterion	Excellent (A+, A, A-)	Good (B+, B, B-)	Adequate (C+, C, C-)	Marginal (D)	Failure (F)
1. Project	1.1 Capacity for effectively implementing data science algorithms in a computationally efficient way.	High	Significant	Moderate	Basic	Not even reaching marginal levels
	1.2 Capability to create new solutions for real world information analytics problems by improving and combining different data science techniques.					
2. Quiz	2.1 Ability to explain in detail the principles of different data science techniques.	High	Significant	Moderate	Basic	Not even reaching marginal levels
	2.2 Capability to correctly apply a suitable data science technique to solve an information analytics problem.					
3. Examination	3.1 Capacity for understanding the main characteristics of different data science techniques in depth.	High	Significant	Moderate	Basic	Not even reaching marginal levels
	3.2 Capability to perform a critical assessment of current data science techniques.					
	3.3 Ability to integrate different data science techniques for addressing real world information analytics problems.					

Part III Other Information (more details can be provided separately in the teaching plan)

1. Keyword Syllabus

(An indication of the key topics of the course.)

Data pre-processing, statistical data modelling, predictive modelling, classifier evaluation, cluster analysis, association rule mining, data stream mining, text mining.

Syllabus

1. Knowledge discovery process

   Introduction of the knowledge discovery process in three stages: data pre-processing, data mining, and knowledge representation. Basic data pre-processing techniques including data cleaning, selection, integration, transformation and reduction will be discussed.

2. Statistical data modelling

   Introduction of fundamental concepts of statistical data modelling, which include random variables, probability distribution functions, probability density functions, covariance matrix, correlation coefficient, linear regression, sampling, statistical inference and multivariate statistical analysis.

3. Predictive modelling

   Introduction of the main predictive modelling techniques for data science, which include decision tree, nearest neighbour classifier and naïve Bayes classifier. In addition, the issues of classification performance evaluation and model selection will be discussed.

4. Cluster analysis

   Introduction of the main clustering techniques: partitional, hierarchical, and density-based clustering. Important algorithms such as k-means, agglomerative hierarchical clustering, and DBSCAN will be discussed. Related issues in outlier analysis and detection will be introduced.

5. Association rule mining

   Introduction of the Apriori algorithm for frequent pattern mining and association rule mining, and the comparison of different measures for evaluating the association patterns. Mining of frequent patterns in data streams will also be discussed.

6. Text mining

   Introduction of the vector space model for document representation, the term frequency-inverse document frequency (tf-idf) approach for term weighting, and proximity measures such as cosine similarity for document comparison. Different algorithms in text mining such as document clustering and text classification will also be discussed.

2. Reading List

2.1 Compulsory Readings

(Compulsory readings can include books, book chapters, or journal/magazine articles. There are also collections of e-books, e-journals available from the CityU Library.)

1.	Tan P. N., Steinbach M. and Kumar V. (2014) *Introduction to Data Mining.* Addison Wesley, 2nd edition.

2.2 Additional Readings

(Additional references for students to learn to expand their knowledge about the subject.)

1.	Bramer M. (2013) *Principles of Data Mining*. Springer, 2nd edition.
2.	Han J. and Kamber M. (2011) *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd edition.
3.	Witten I., Frank E. and Hall M. (2011) *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 3rd edition.