City University of Hong Kong
Course Syllabus

offered by School of Data Science
with effect from Semester A 2021/22

Part I Course Overview

Course Title:	**Data Mining**
Course Code:	**SDSC3002**
Course Duration:	**One Semester**
Credit Units:	**3**
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)*	**MA2506 Probability and Statistics**
Precursors: *(Course Code and Title)*	**Nil**
Equivalent Courses: *(Course Code and Title)*	**Nil**
Exclusive Courses: *(Course Code and Title)*	**Nil**

Part II Course Details

1. Abstract

(A 150-word description about the course)

Data mining is about the extraction of non-trivial, implicit, previously unknown and potentially useful principles, patterns or knowledge from massive amount of data. This course introduces the foundation of data mining techniques, including basic concepts of data representation, new software stack for processing massive data such as MapReduce and Spark, and popular data mining tasks like mining frequent itemsets, nearest neighbor search, clustering analysis and graph mining. Students will also learn how data mining techniques are used in real-world applications such as online advertising and recommender systems.

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.	**Understand** the abstract representation of data, such as vectors, matrices, sets and graphs, with modelling considerations, for use in downstream applications	10%	√		
2.	**Familiarize** classical data mining methods such as pattern mining, classification, dimensionality reduction and clustering	30%	√	√	
3.	**Implement** scalable algorithms to conduct data mining tasks	30%	√	√	√
4.	**Demonstrate** the ability of working with other students on projects addressing challenging problems from real-world data mining applications	30%	√	√	
** 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.)

TLA	Brief Description	CILO No.				Hours/week (if applicable)
		1	2	3	4	
Lecture	Lectures and class projects	√	√	√	√	3 hours/week
Tutorial	Teach the software packages and coding			√	√	6 hours/semester, included in lecture time

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: 70%						
Assignments	√	√	√	√	40%	
Project		√	√	√	30%	
Examination: 30% (duration: 2 hours)						
Examination	√	√	√	√	30%	
**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 should 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-)	Fair (C+, C, C-)	Marginal (D)	Failure (F)
1. Coursework	Assignment, Participation, Project presentation and report	High	Significant	Moderate	Basic	Not even reaching marginal levels
2. Examination	Open-book and notes exam	High	Significant	Moderate	Basic	Not even reaching marginal levels

Examination, test, continuous assessment and laboratory reports will be numerically-marked.

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.)

Introduction to Data Mining: data representation; data mining tasks; overlaps with machine learning, database systems and theoretical computer science; new computing software like MapReduce and Spark.

Itemset Mining: market-basket model; frequent itemsets; A-priori algorithms; sampling-based frequent itemset mining algorithms; association rules.

Similarity/Distance between data points: nearest neighbor search; Minhashing algorithm; locality sensitive hashing; dimensionality reduction; principle component analysis; random projections.

Clustering: k-means algorithm; hierarchical clustering; density-based clustering; spectral clustering; graph Laplacian matrix.

Graph Analysis: graph centrality measures; PageRank; hubs and authorities in networks; stochastic diffusion models; Markov chains and random walks; graph representation learning; link prediction.

Applications: online advertising; the matching problem; recommender systems; matrix factorization; collaborative filtering; social network mining; community detection and graph partition; network sampling.

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.	Jure Leskovec, Anand Rajaraman, Jeff Ullman, *Mining of Massive Datasets*. 3rd edition, Cambridge University Press
2.	Lecture notes and reading materials selected by the instructor

2.2. Additional Readings

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

1.	Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques, 3rd ed. The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, July 2011. ISBN 978-0123814791