City University of Hong Kong
Course Syllabus

offered by School of Data Science
with effect from Semester A 2024/25

Part I Course Overview

Course Title:	Machine Learning at Scale
Course Code:	SDSC6009
Course Duration:	One Semester
Credit Units:	3
Level:	P6
Medium of Instruction:	English
Medium of Assessment:	English
Prerequisites: *(Course Code and Title)*	SDSC5001 Statistical Machine Learning I
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

This course teaches the underlying principles required to develop scalable machine learning pipelines for structured and unstructured data at the petabyte scale. The course covers principles of scaling machine learning process under big data via deploying the MapReduce parallel computing. In addition, the hands-on algorithmic design and development of machine learning algorithms in parallel computing environments (Spark) will be discussed. Students will use MapReduce parallel computing frameworks for machine learning in industrial applications and deployments for various fields, including advertising, finance, healthcare, and search engines.

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.	Describe principles of scalable machine learning and parallel computing	20%	✓		
2.	Discuss big data management tools and ecosystem	20%	✓		
3.	Design and develop parallel computing and scalable machine learning algorithms	20%	✓	✓	
4.	Conduct assessment, comparison, and selection for scalable learning models	20%	✓	✓	
5.	Implement parallel compute frameworks for industrial applications	20%		✓	✓
		100%			

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 real-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. Learning and Teaching Activities (LTAs)

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

LTA	Brief Description	CILO No.						Hours/week (if applicable)
		1	2	3	4	5		
Lecture	Students will engage in formal lectures to gain knowledge about principles of scalable machine learning pipelines covered in this course	✓	✓	✓	✓			26 hours/semester
Laboratory work	Students will participate in lab activities to develop the ability of implementing scalable machine learning pipelines		✓	✓	✓	✓		13 hours/semester

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	5			
Continuous Assessment: 65 %								
Group Project		✓	✓	✓	✓		40%	
Individual Coursework	✓	✓	✓	✓			25%	
Examination: 35 % (duration: 2 hours , if applicable)								
Examination	✓	✓	✓	✓	✓		35%	
							100%	

5. Assessment Rubrics

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

Applicable to students admitted before Semester A 2022/23 and in Semester A 2024/25 & thereafter

Assessment Task	Criterion	Excellent (A+, A, A-)	Good (B+, B, B-)	Fair (C+, C, C-)	Marginal (D)	Failure (F)
1. Group Project	40%	High	Significant	Moderate	Basic	Not even reaching marginal levels
2. Individual Coursework	25%	High	Significant	Moderate	Basic	Not even reaching marginal levels
3. Examination	35%	High	Significant	Moderate	Basic	Not even reaching marginal levels

Applicable to students admitted from Semester A 2022/23 to Summer Term 2024

Assessment Task	Criterion	Excellent (A+, A, A-)	Good (B+, B)	Marginal (B-, C+, C)	Failure (F)
1. Group Project	40%	High	Moderate	Basic	Not even reaching marginal levels
2. Individual Coursework	25%	High	Moderate	Basic	Not even reaching marginal levels
3. Examination	35%	High	Moderate	Basic	Not even reaching marginal levels

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.)

A review of big databases: Distributed file storage, Hadoop, Spark, MLLib
Machine learning under big data environment: Implement machine learning methods via Spark to analyse big data, Principles in decomposing large-scale learning tasks into distributed individual sub-learning tasks, Optimization contents in distributed learning.
Transfer learning: Domain source and target source learning; transfer learning methods, residual function transfer, discrepancies between domain source model and target source model, industrial case studies using transfer learning.
Recommendation Systems at Scale: Graph-networks, Link Analysis, collaborative filtering, Sparsity and Scalability in recommendation systems.
Introductory Real-time Computer Vision: Organization of training image samples, Transfer learning in CNN, You Only Look Once method.

Programming in Spark will be covered in the lab sessions.

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.	Lecture Notes

2.2 Additional Readings

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

1.	Jure Leskovec, Anand Rajaraman, Jeff Ullman, Mining of Massive Datasets
2.	Sandy Ryza, Uri Laserson, Sean Owen & Josh Wills. Advanced Analytics with Spark
3	Ron Bekkerman, Mikhail Bilenko, John Langford. Scaling up Machine Learning: Parallel and Distributed Approaches