SDSC6009: MACHINE LEARNING AT SCALE

Effective Term
Semester A 2025/26

Part I Course Overview

Course Title
Machine Learning at Scale

Subject Code
SDSC - Data Science
Course Number
6009

Academic Unit
Data Science (DS)

College/School
College of Computing (CC)

Course Duration
One Semester

Credit Units
3

Level
P5, P6 - Postgraduate Degree

Medium of Instruction
English

Medium of Assessment
English

Prerequisites
Nil

Precursors
Nil

Equivalent Courses
Nil

Exclusive Courses
Nil

Part II Course Details

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.

Course Intended Learning Outcomes (CILOs)

	CILOs	Weighting (if app.)	DEC-A1	DEC-A2	DEC-A3
1	Describe principles of scalable machine learning and parallel computing	20	x		
2	Discuss big data management tools and ecosystem	20	x		
3	Design and develop parallel computing and scalable machine learning algorithms	20	x	x	
4	Conduct assessment, comparison, and selection for scalable learning models	20	x	x	
5	Implement parallel compute frameworks for industrial applications	20		x	x

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.

Learning and Teaching Activities (LTAs)

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

Assessment Tasks / Activities (ATs)

	ATs	CILO No.	Weighting (%)	Remarks ("-" for nil entry)	Allow Use of GenAI?
1	Group Project	2, 3, 4, 5	40	-	Yes
2	Individual Coursework	1, 2, 3, 4	25	-	Yes

Continuous Assessment (%)

65

Examination (%)

35

Examination Duration (Hours)

2

Minimum Examination Passing Requirement (%)

30

Additional Information for ATs

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

Assessment Rubrics (AR)

Assessment Task

Group Project (for students admitted before Semester A 2022/23 and in Semester A 2024/25 & thereafter)

Criterion

40%

Excellent

(A+, A, A-) High

Good

(B+, B, B-) Significant

Fair

(C+, C, C-) Moderate

Marginal

(D) Basic

Failure

(F) Not even reaching marginal levels

Assessment Task

Individual Coursework (for students admitted before Semester A 2022/23 and in Semester A 2024/25 & thereafter)

Criterion

25%

Excellent

(A+, A, A-) High

Good

(B+, B, B-) Significant

Fair
(C+, C, C-) Moderate

Marginal
(D) Basic

Failure
(F) Not even reaching marginal levels

Assessment Task
Examination (for students admitted before Semester A 2022/23 and in Semester A 2024/25 & thereafter)

Criterion
35%

Excellent
(A+, A, A-) High

Good
(B+, B, B-) Significant

Fair
(C+, C, C-) Moderate

Marginal
(D) Basic

Failure
(F) Not even reaching marginal levels

Assessment Task
Group Project (for students admitted from Semester A 2022/23 to Summer Term 2024)

Criterion
40%

Excellent
(A+, A, A-) High

Good
(B+, B) Moderate

Marginal
(B-, C+, C) Basic

Failure
(F) Not even reaching marginal levels

Assessment Task
Individual Coursework (for students admitted from Semester A 2022/23 to Summer Term 2024)

Criterion
25%

Excellent
(A+, A, A-) High

Good
(B+, B) Moderate

Marginal
(B-, C+, C) Basic

Failure
(F) Not even reaching marginal levels

Assessment Task
Examination (for students admitted from Semester A 2022/23 to Summer Term 2024)

Criterion
35%

Excellent
(A+, A, A-) High

Good
(B+, B) Moderate

Marginal
(B-, C+, C) Basic

Failure
(F) Not even reaching marginal levels

Part III Other Information

Keyword Syllabus

- 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.

Reading List

Compulsory Readings

	Title
1	Lecture Notes

Additional Readings

	Title
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