SDSC3002: DATA MINING

Effective Term
Semester A 2024/25

Part I Course Overview

Course Title
Data Mining

Subject Code
SDSC - School of Data Science

Course Number
3002

Academic Unit
School of Data Science (DS)

College/School
School of Data Science (DS)

Course Duration
One Semester

Credit Units
3

Level
B1, B2, B3, B4 - Bachelor's Degree

Medium of Instruction
English

Medium of Assessment
English

Prerequisites
MA2506 Probability and Statistics or MA2510 Probability and Statistics

Precursors
Nil

Equivalent Courses
Nil

Exclusive Courses
Nil

Part II Course Details

Abstract

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.

Course Intended Learning Outcomes (CILOs)

	CILOs	Weighting (if app.)	DEC-A1	DEC-A2	DEC-A3
1	Desribe the abstract representation of data, such as vectors, matrices, sets and graphs, with modelling considerations, for use in downstream applications	10	x		
2	Discuss classical data mining methods such as pattern mining, classification, dimensionality reduction and clustering	30	x	x	
3	Implement scalable algorithms to conduct data mining tasks	30	x	x	x
4	Demonstrate the ability of working with other students on projects addressing challenging problems from real-world data mining applications	30	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 lectures and class projects	1, 2, 3, 4	3 hours/week
2	Tutorial	Students will engage in tutorials teaching the software packages and coding	3, 4	6 hours/semester, included in lecture time

Assessment Tasks / Activities (ATs)

	ATs	CILO No.	Weighting (%)	Remarks (e.g. Parameter for GenAI use)
1	Assignments	1, 2, 3, 4	40	
2	Project	2, 3, 4	30	

Continuous Assessment (%)

70

Examination (%)

30

Examination Duration (Hours)

2

Additional Information for ATs

Note: To pass the course, apart from obtaining a minimum of 40% in the overall mark, a student must also obtain a minimum mark of 30% in both continuous assessment and examination components.

Assessment Rubrics (AR)

Assessment Task

Coursework

Criterion

Assignment, Participation, Project presentation and report

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

Criterion

Open-book and notes exam

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

Additional Information for AR

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

Part III Other Information

Keyword Syllabus

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; principal 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.

Reading List

Compulsory Readings

	Title
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

Additional Readings

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