University f Halabja

Directorate of Quality Assurance

SUBJECT OUTLINE

Academic Year: 2023-2024

1. Information on the Programme

Higher education institution	**University of Halabja**
College	**College of Science**
Department	**Computer**
Field of study	**Computer Science**
Cycle of study[1]	**First Cycle**
Specialization/ Study program	**N/A**
Form of education	**Full time**

2. Information on the Discipline

Discipline Name	**Data Mining**	**Discipline Code**	………….
ECTS	**6**	**Language**	English
Lecturer (Theory)	**Peshraw A. Abdalla**	**Home page**	https://tqa.uoh.edu.iq/uoh/profile/peshraw.abdalla@uoh.edu.iq/
Moodle Course link	https://moodle.uoh.edu.iq/course/view.php?id=339	**Google Scholar**	https://scholar.google.com/citations?user=hDSB67IAAAAJ&hl=en&oi=ao
E-mail	**peshraw.abdalla@uoh.edu.iq**	**Tel**	
Practical/Seminar / laboratory/ project Lecturer		**Home page**	
Moodle Course Link		**Google Scholar**	
E-mail		**Tel**	
Study Year	**4**	**Semester**	7th
Assessment type[2]	**Exam**	**Discipline status**	
Content[3]	**SD**	**Mandatory**[4]	MD

3. Prerequisites (if applicable)

Curriculum-related	Fundamentals of Programming with PYTHON, Mathematics, Statistics, Strutural Programming, OOP, Database.
Skills-related	Mathematics, Programming, Statistics

	Decipline:	Subject Name			ECTS:	6.00										
	Workload		164	Total Contact Hours:			56	Total Self Study Hours:			108					
	No. of Weeks	1st Week	2nd Week	3rd Week	4th Week	5th Week	6th Week	7th Week	8th Week	9th Week	10th Week	11th Week	12th Week	13th Week	14th and 15th Week (Final	Total
Contact Hours	Theoritical	2	2	2	2	2	2	2	2	2	2	2	2	2		26
	Practice	2	2	2	2	2	2	2	2	2	2	2	2	2		26
	Lab./Tutorial															0
	Fieldtrips/Visits															0
	Project					2					2					4
Self Study	Curriculum (articles+media+net)	5														5
	Curriculum (Books)		5	5		5	2		5		5	5				32
	Homework				2				5				5			12
	Quizzes						3			4		5				12
	Assignments			5		5			5			5				20
	Reports															0
	Presentation				3							3				6
	Midterm Exam (Thr. + Pr.)		7													7
	Final Exam (Thr. + Pr.)	2	2	2	2	2	2	2	2	2	2	2	2	2	14	14

5. Conditions (if applicable)	
For the Theoretical	**Policy Statement on Extensions** Extensions to the exams or project due dates will be given in the event of extenuating circumstances (such as illness, personal emergency, etc.) If you submit a brief written request to the lecturer as soon as possible after the circumstances arise. This request will be initialed (if approved) and will be returned to you. You must attach it to or cite it on the piece of work for which the extension was granted. **Academic Dishonesty** Academic dishonesty is regarded as a major violation of both the academic and professional principles of this community and may result in a failing grade or suspension. Academic dishonesty includes plagiarism, cheating (whether in or out of the classroom), and abuse or misuse of lab materials when such abuse or misuse can be related to course requirements. **Class Attendance and Participation Policy** Regular class attendance and participation is an essential component of this course and expected of all students. Class attendance and participation will be recorded. Please come to class having completed the assigned reading for the day and ready to discuss and unpack the material with your instructor and peers. Absences from class will be classified as "documented" or "undocumented." A documented absence is one where written documentation is submitted supporting the absence from class due to circumstances beyond the student's control. An undocumented absence is any other absence, including one that could qualify as documented if proper documentation were submitted. Multiple undocumented absences will impact your final course grade as follows: • Each student may take one (1) undocumented absence without penalty. • Each subsequent undocumented absence will cause the student's final course grade to be reduced by 2.5%. • Students with more than four (3) undocumented absences will automatically fail the course. • Students who arrive more than five (5) minutes late to class more than three (3) times during the semester will have each subsequent late arrival to class counted as a half undocumented absence for that class.
For the Practical/Lab. /Project	The same policy for the theory

6. Cumulated specific competences	
Professional competencies	Problem-solving, Numerical solution, Modelling, Programming (Python), Error analysis
Transversal competences	Data analyzing, Problem-solving, Programming (Python), teamwork, and critical thinking.

7. Discipline objectives (based on the cumulated specific competencies)	
General objective	This course is an introductory course on data mining. It introduces the basic concepts, principles, methods, implementation techniques, and applications of data mining, with a focus on two major data mining functions: (1) pattern discovery and (2) cluster analysis. In the first part of the course, which focuses on pattern discovery, you will learn why pattern discovery is important, what the major tricks are for efficient pattern mining, and how to apply pattern discovery in some interesting applications. The course provides you the opportunity to learn concepts, principles, and skills to practice and engage in scalable pattern discovery methods on massive data; discuss pattern evaluation measures; study methods for mining diverse kinds of frequent patterns, sequential patterns, and sub-graph patterns; and study constraint-based pattern mining, pattern-based classification, and explore their applications. In the second part of the course, which focuses on cluster analysis, you will learn concepts and methodologies for cluster analysis, which is also known as clustering, data segmentation, or unsupervised learning. We will introduce the basic concepts of cluster analysis and then study a set of typical clustering methodologies, algorithms, and applications. This includes partitioning methods, such as k-means, hierarchical methods, such as BIRCH, density-based methods, such as DBSCAN, and grid-based methods, such as CLIQUE. We will also discuss methods for clustering validation. The learning will be enhanced by clustering software and programming assignments. The technical contents of the course are based on the textbook Data Mining: Concepts and Techniques (3rd ed), as well as the on-campus course CS 412 – Introduction to Data Mining, which is

	offered in the Department of Computer Science at the University of Illinois. Please note several themes covered in the textbook are not covered in this online course, including (1) data preprocessing and preparation, (2) data warehouse and data cube technology, and (3) classification. This is because these themes have been covered or will be covered, with possible in-depth treatment, in several other courses offered in the Data Science Online Master program. Therefore, this course will focus on the in-depth study of the two major data mining functions illustrated above.
Specific objectives (Learning Outcomes)	Upon successful completion of this course, for pattern discovery, you will be able to: • Recall important pattern discovery concepts, methods, and applications, in particular, the basic concepts of pattern discovery, such as frequent pattern, closed pattern, max-pattern, and association rules. • Identify efficient pattern mining methods, such as Apriori, ECLAT, and FPgrowth. • Compare pattern evaluation issues, especially several popularly used measures, such as lift, chi square, cosine, Jaccard, and Kulczynski, and their comparative strengths. • Compare mining diverse patterns, including methods for mining multi-level, multi-dimensional patterns, qualitative patterns, negative correlations, compressed and redundancy-aware top-k patterns, and mining long (colossal) patterns. • Learn well-known sequential pattern mining methods, including methods for mining sequential patterns, such as GSP, SPADE, PrefixSpan, and CloSpan. • Learn graph pattern mining, including methods for subgraph pattern mining, such as gSpan, CloseGraph, graph indexing methods, mining top-k large structural patterns in a single large network, and graph mining applications, such as graph indexing and similarity search in graph databases. • Learn constraint-based pattern mining, including methods for pushing different kinds of constraints, such as data and pattern-based constraints, anti-monotone, monotone, succinct, convertible, and multiple constraints. • Learn pattern-based classifications, including CBA, CMAR, PatClass, and DPClass. • Enjoy various pattern mining applications, such as mining spatiotemporal and trajectory patterns and mining quality phrases. • Explore further topics on pattern analysis, such as pattern mining in data streams, software bug mining, pattern discovery for image analysis, and privacy-preserving data mining. For cluster analysis, you will be able to: • Recall basic concepts, methods, and applications of cluster analysis, including the concept of clustering, the requirements and challenges of cluster analysis, a multi-

	dimensional categorization of cluster analysis, and an overview of typical clustering methodologies. • Learn multiple distance or similarity measures for cluster analysis, including Euclidean and Minkowski distances; proximity measures for symmetric and asymmetric binary variables; distance measures between categorical attributes, ordinal attributes, and mixed types; proximity measures between two vectors – cosine similarity; and correlation measures between two variables covariance and correlation coefficient. • Learn popular distance-based partitioning algorithms for cluster analysis, including K-Means, K Medians, K-Medoids, and the Kernel K-Means algorithms. • Learn hierarchical clustering algorithms, including basic agglomerative and divisive clustering algorithms, BIRCH, a micro-clustering-based approach, CURE, which explores well-scattered representative points, CHAMELEON, which explores graph partitioning on the KNN Graph of the data, and a probabilistic hierarchical clustering approach. • Learn the density-based approach to cluster analysis, which can group dense regions of arbitrary shape, such as DBScan and OPTICS. • Learn the grid-based approach, which organizes individual regions of the data space into a grid-like structure, such as STING and CLIQUE. • Study concepts and methods for clustering evaluation and validation by introducing clustering validation using external measures and internal measures, and the measures for evaluating cluster stability and clustering tendency.

8. Content

Theoretical- Number of hours	**Teaching**	**Observation**
First week	**Registration**	2 hours
Second week	**Python Introduction**	2 hours
Third week	**Introduction to Data Mining:** • What is Data Mining? • Tasks and Applications • The Data Mining Process	2 hours
Fourth week	**Clustering 1:** • K-means Clustering	2 hours
Fifth week	**Clustering 2:** • Density-based Clustering, • Hierarchical Clustering, • Proximity Measures	2 hours

Sixth week	**Mid-Trem Exam**	
Seventh week	**Classification 1:** - Nearest Neighbor - Decision Trees and Forests	2 hours
Eighth week	**Classification 2:** - Rule Learning, - Naïve Bayes, - SVMs, - Neural Networks,	2 hours
Ninth week	**Classification 3:** - Model Evaluation	2 hours
Tens week	**Project Presentation**	2 hours
Eleventh week	**Seminar Papers**	2 hours
Twelfth week	Regression: Linear Regression, Nearest Neighbor Regression, Regression Trees, Time Series	2 hours
Thirteenth week	Text Mining: Preprocessing Text, Feature Generation, Feature Selection	2 hours

Practical Works– Number of hours	**Teaching**	**Observation**
First week	**Registration**	2 hours
Second week	**Python Introduction**	2 hours
Third week	**Introduction to Data Mining**	2 hours
Fourth week	Clustering 1	2 hours
Fifth week	Clustering 2	2 hours
Sixth week	**Mid-Trem Exam**	2 hours
Seventh week	Classification 1	2 hours
Eighth week	Classification 2	2 hours
Ninth week	Classification 3	2 hours
Tenth week	Project Presentation	2 hours

Eleventh week	Seminar Papers	2 hours
Twelfth week	Regression	2 hours
Thirteenth week	Text Mining	2 hours

9. Compulsory bibliography

1- **Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining. 2nd Edition. Pearson / Addison Wesley.**

2- **Aurélien Géron: Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow. 2nd or 3rd Edition, O'Reilly, 2019 or 2022**

Optional bibliography

Although the lectures are designed to be self-contained, it is recommended (but not required) to reference the textbook: Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Waltham: Morgan Kaufmann. You can download a PDF version of the chapters 1, 6, 7 and 2, 10, 11, 13 from Data mining: Concepts and techniques (3rd ed.) for free. Note that these are all the chapters related to the topics covered in this course, so the free PDF version of the chapters is sufficient for this course. If you would like to purchase the entire textbook, the publisher has an exclusive offer just for Coursera students. You can save 30% on either the print or eBook version of Data Mining: Concepts and Techniques, 3rd Edition and receive free shipping on all orders. Here is how it works: • Add the book to your cart. • Enter code COMP317 and click Apply. • The discount will be applied to the list price and cannot be combined with other promotions.

10. Corroborating the discipline content with the expectations of the epistemic community representatives, of the professional associations and of the relevant employers in the corresponding field

1.

2.

3.

4.

11. Assessment: **Students will be graded on their performance in the exams(theory and practice), assignments, presentations and one project, more precisely the grading is divided as follows:**			
Type of Activity	**Assessment criteria[2]**	**Assessment type**	**Final grade Percentage**
Final Exam	**Written Exam**	**writing examination**	**50%**
Practical/Laboratory	**Practical**	**Lab exam**	**%25**
Activity during semester	**Oral Exam**	**Assignment(10), Seminars Quiz(10) & Projects(5)**	**%25**
Minimum performance standards: Reading English well & Solving precalculus problems (Algebra) and having an introduction to Python basic commands			

Theoretical Lecturer	**Asst. Lec. Peshraw A. Abdalla**
Practice Lecturer	

Approved by the Curriculum Development Committee	
1	
2	
3	
Head of the Department/ Dean	

Notes:

1 Cycle of studies - choose one of the three options: Bachelor «1», Master «2», Ph.D. «3»

2 (Exam: oral examination, written exam), and (Continous Evaluation(CE), portfolio).

3 Discipline status (content) - for the Bachelor level, choose one of the options: FD (fundamental (General) discipline), PF (Preparatory Disciplines in the Field), SD (Specialty Disciplines), CD (Complementary Disciplines), DU (disciplines based on the university's options).

4 Discipline status (compulsoriness) - choose one of the options

– MD (Mandatory discipline),

- OD (optional discipline),
- ED (Elective (Facultative) discipline).

5 Note: 1 ECTS = 27 hours workload;   ECTS=WL/27, The first week is registration and introduction to the course.