



COURSE SYLLABUS

I-Course Information :

Title of the course	Data Mining
Department	Information Management and Data Sciences
Specialization	Data Science
Academic Year	2020-2021
Semester Duration	
Credit Number	
Number of Hours	40 hours
Course Instructor	Dr.Maguy Medlej
Email	Maguy.medlejkhoury@hotmail.com
Office Hours	0

II- Course Description

This course focuses on both concepts and practice. We will introduce

- (a) the core data mining concepts and
- (b) practical skills for applying data mining techniques to solve real-world problems.

III- Course Prerequisites

- Knowledge of basic statistics and data structures;
- Basic knowledge of database design and database programming.
- R DataAnalysis
- R programming

IV-Course Objectives

- Study the major data mining problems as different types of computational tasks (prediction, classification, clustering, etc.) and the algorithms appropriate for addressing these tasks
- Learn how to analyze data through statistical and graphical summarization, supervised and unsupervised learning algorithms
- Systematically evaluate data mining algorithms and understand how to choose algorithms for different analysis tasks



Practice

- Learn how to gather and process raw data into suitable input for a range of data mining algorithms
- Critique the methods and results from a data mining practice
- Design and implement data mining applications using real-world datasets, and evaluate and select proper data mining algorithms to apply to practical scenarios

V- Course Outline and Calendar

Learning Objective	Learning Content	Pedagogical Strategy & Tools
Data Mining Overview Data Mining with R	• Introduction to Data Mining, Importance of Data Mining, Data Mining functionalities, Relationship between Data Warehousing, On-line Analytical Processing, and Data Mining. • Overview: Data mining tasks - Clustering, Classification, Regression Analysis, Rule learning, etc. • Data mining architecture, Major Issues in Data Mining, Data mining metrics, Applications of Data Mining • Data Mining Terminology	Lecture R environment preparation
Classification & Predictive Modeling Concepts, Algorithms, and Applications	• Basic Concepts in Classification & Prediction • Model Evaluation Techniques • Prediction using linear regression • Decision Tree Classification • Bayesian Classification	Lecture Lab Task Assignment
Classification & Predictive Modeling Concepts, Algorithms, and Applications (Cont.) Distance-Based Classification and Prediction Association rules	• Classification using Distances, Similarities and Predictive Modeling • Market Basket Analysis & Association Rule Mining	Lecture Lab Task Assignment



Clustering and Sampling Finding Groups and Similarities In Data	• Basic Clustering Concepts & Algorithms • Clustering Applications, User Profiling, and Personalization	Lecture Task Assignment
Use Case implementation in R	• Implementing practical use case	Project: Case study

P.S.: The course outline above may be modified during the semester.

VI- Teaching approach & methods

Slides, Practical exercises, assignments, documents reading, project implementation

VII- Assessment Methods

The success of the student in this course is based on:

- 1- His/her active participation in class.
- 2- His/her scientific contribution through research done outside the classroom.
- 3- His/her reading notes, research reports and practical applications as requested by course instructor.
- 4- Assignments delivery
- 5- Project implementation

The assessment of the student is carried out according to the following grid:

Assessment Type	Final Grade Weighting (%)	Dates/Deadlines
Participation (projects, research paper, oral presentation, etc.)	20%	
Midterm Exam – Practical Assignment	20%	
Final exam	60%	
Total	100%	

VIII- Specific Course Requirements

1- Regulations concerning deadlines for written and applied work submissions

The teacher has the right to:

- Refuse to receive written applied work outside the deadlines.
- Give the student additional time after omitting up to 25% of the mark awarded to the concerned activity.

2- Academic measures in case of plagiarism



The student shall be liable to the following penalties in case of any attempt of illicit scientific citation:

- Cancellation of the student's contribution, activity or scientific article and reduction of the final mark.
- If the same offense is reported in other courses, cancellation of the final exam and failure in the relevant course(s).

IX- Reference Book

- Jiawei Han and Micheline Kamber, [Data Mining: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems](#), Jim Gray, Series Editor. Morgan Kaufmann Publishers, August 2000. 550 pages. ISBN 1-55860-489-8.
- Ian H. Witten and Eibe Frank, [Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations](#), Morgan Kaufmann Publishers, October 1999. 371 pages. ISBN 1-55860-552-5.
- [Data Mining and Business Analytics with R](#), Johannes Ledolter, Wiley, 2013, ISBN: 978-1118447147 (online access via [Pitt network](#)) (primary book, hereafter referred as "DMR")
- [Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data](#) (2nd ed.), Bing Liu, Springer, 2011, ISBN: 978-3642194597 (available [online](#)) (secondary book, hereafter referred as "WDM")
- [Practical Data Science with R](#), Nina Zumel and John Mount, Manning Publications 2014, ISBN: 9781617291562 (online access via [Pitt network](#)) (third book, hereafter referred as "DSR")

X- Bibliography& Additional Sources

[simpleR](#)
[ggplot2](#)
[Cookbook for R](#)
[Quick-R](#)
[Guide to Getting Started in Machine Learning](#)
[Mining twitter with R](#)
[R client library for the Google Prediction API](#)
[R Tutorial ebook](#)
[R for Programmers](#)

XI- Data sets

[Public Data Sets from Amazon](#)
[dataset for social research](#)
[Public Datasets](#)
[Stack Overflow Dumps](#)



[Large Movie Review Dataset](#)

[GroupLens datasets](#)

[Jester](#)

[HetRec 2011](#)

[CAW 2.0](#)

[A Meta-index of Data Sets](#)

[UCI KDD Archive](#)

[Stanford Large Network Dataset Collection](#)

[SemEval-2013: Sentiment Analysis in Twitter](#)

[New York Times APIs](#)

[HealthData.gov](#)