Course syllabus «Data Mining»

Approved by
Programme Academic Council
Protocol Nr. 10 from 27.08.2019

Author	Dr. Janez Demsar
Number of credits	4
Contact hours	48
Self-study hours	104
Course	1,2
Educational format	Without use of online course

I. Goals and Results of Mastering the Discipline; Prerequisites

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.

As a result, students should:

Know:

- well-known sequential pattern mining methods, including methods for mining sequential patterns, such as GSP, SPADE, PrefixSpan, and CloSpan.
- various pattern mining applications, such as mining spatiotemporal and trajectory patterns and mining quality phrases.
- efficient pattern mining methods, such as Apriori, ECLAT, and FPgrowth.
- 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.

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.
- Compare pattern evaluation issues, especially several popularly used measures, such as lift, chisquare, 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.

Have:

- an ability to perform 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.

- an ability to select among the 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.
- an ability to use 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.

Basic knowledge of introductory statistics are required for this course.

The basics of this discipline should be used in all other program related courses.

The course is strongly related and complementary to other compulsory courses provided in the first year (e.g. Applied Linear Models II, Contemporary Data Analysis) and sets a crucial prerequisite for later courses and research projects as well as for the master thesis. The course gives students an important foundation to develop and conduct their own research as well as to evaluate research of others.

II. Content of the Course

Please note: some sessions, due to their complexity, will take more than one lecture to cover.

SESSION ONE: Introduction

Course Orientation; Course Pattern Discovery Overview; Pattern Discovery Basic Concepts; Efficient; Pattern Mining Methods; Pattern Discovery

SESSION TWO: Pattern evaluation

The session sets up the framework for pattern evaluation and mining diverse frequent patterns. It also addresses Sequential Pattern Mining; Pattern Mining Applications; Mining Spatiotemporal and Trajectory Patterns.

SESSION THREE: Pattern mining I

The session gives an overview into pattern-based mining, graph pattern mining, and pattern-based classification.

SESSION FOUR: Pattern mining II

This sessions builds the understanding of Pattern Mining Applications: Mining Quality Phrases-from Text Data; Advanced Topics on Pattern Discovery.

SESSION FIVE: Cluster analysis

Cluster Analysis Overview; Cluster Analysis Introduction; Similarity Measures for Cluster Analysis

SESSION SIX: Clustering Methods I

This session will continue the topic of clustering with Partitioning-Based Clustering Methods; Hierarchical Clustering Methods.

SESSION SEVEN: Clustering Methods II

Hierarchical Clustering Methods (continued); Density-Based and Grid-Based Clustering Methods

SESSION EIGHT – Clustering Methods III

This session will conclude clustering with methods for clustering validation.

III. Grading

Course grade will be completed as follows:

Course Element	**% Towards Final Grade**
Final Project	**50%**
Final take-home project	*50%*
Participation and responsibility grade	**50%**
Homework Assignments (5 x Varied points)	*20%*
In-Class Labs (9-10 x Varied points)	*20%*
Quizzes (Best 9 of 10, Varied points)	*10%*
Extra credit	As assigned
Total	100%

If the final grade is non-integer, it is rounded according to algebraic rules. If has a half (.5) at the end, we are rounding upward. Rounding of cumulative grades and other rounding issues are performed according to the HSE rules.

IV. Grading Tools

This class contains several assignments that test student knowledge and understanding throughout the course.

Quizzes

You cannot meaningfully participate in the seminar if you have missed my lecture and did not do any reading. Therefore, to encourage you to prepare for seminars, every seminar will have a quiz on the lecture material and all assigned readings for the week. This includes the very first seminar, which will focus on Lecture 1 material. You are allowed to miss any one quiz (skip a seminar, not prepare, etc.) – in other words, I will count the best 9 out of 10 quizzes that we will have. If you submit all ten, I will count best nine. All quizzes will be done online and submitted to me via SurveyMonkey (links will be given in class).

Important: I record IP addresses and only accept quizzes submitted from with the HSE IP address. Quizzes submitted from other locations are NOT counted towards your grade. In other words, to participate in a quiz, you have to be present in class.

In-class Labs
There will be a lab assignment in almost every seminar, depending on our progress. Since we will be learning SAS, and learning quickly, you will need to devote a substantial time to it. Seminar labs should help you with this task. At the end of the lab, you will submit your completed assignment for the day (or as much as you were able to complete) to me via LMS.

Homework assignments
There will be several homework assignments that will provide additional hands-on practice for the concepts we've learned in class and practiced during the seminar. Homeworks will be assigned as needed throughout the semester. All homework submissions must be done by the stated deadline via the LMS system.

Final project
For the final project you will be asked to analyze the patterns in a large set of data (TBD). Instructions will be provided during the course.

V. Resources

5.1 Main Literature

1. Aggarwal, Charu C. *Data mining: the textbook*. Springer, 2015. URL https://proxylibrary.hse.ru:2176/book/10.1007/978-3-319-14142-8. Springer Link.
2. Han, Jiawei, et al. Data Mining: Concepts and Techniques, Elsevier Science & Technology, 2011. ProQuest Ebook Central, URL https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=729031.
3. Larose, Daniel T., and Chantal D. Larose. Data Mining and Predictive Analytics, John Wiley & Sons, Incorporated, 2015. ProQuest Ebook Central, URL https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1895687.
4. Data Mining and Learning Analytics : Applications in Educational Research, edited by Samira ElAtia, et al., John Wiley & Sons, Incorporated, 2016. ProQuest Ebook Central, URL https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=4675499.
5. Mourya, S.K., and Shalu Gupta. Data Mining and Data Warehousing, Alpha Science International, 2012. ProQuest Ebook Central, URL https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=5218420.

5.2 Additional Literature

1. Brown, Meta S.. Data Mining for Dummies, John Wiley & Sons, Incorporated, 2014. ProQuest Ebook Central, URL https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=1780446.
2. Knobbe, A.J.. Multi-Relational Data Mining, IOS Press, 2006. ProQuest Ebook Central, URL https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=274726.
3. Active Mining : New Directions of Data Mining, edited by H. Motoda, IOS Press, 2002. ProQuest Ebook Central, URL https://ebookcentral.proquest.com/lib/hselibrary-ebooks/detail.action?docID=267470.

5.3 Software

№ п/п	Name	Access conditions
1.	MicrosoftWindows 7 Professional RUS MicrosoftWindows 10 MicrosoftWindows 8.1 Professional RUS	*From the university's internal network (contract)*
2.	Microsoft Office Professional Plus 2010	*From the university's internal network (contract)*
3.	R, R studio	*Open access. URL: https://www.r-project.org/*
4.	Orange	*Open access University Edition. URL: https://orange.biolab.si/*

5.3 Material and technical support

Classrooms for lectures on the discipline provide for the use and demonstration of thematic illustrations corresponding to the program of the discipline, consisting of:

- PC with Internet access (operating system, office software, antivirus software);
- multimedia projector with remote control.