COURSE DESCRIPTION CARD - SYLLABUS

Course name
Mining of Massive Datasets

Course

Field of study
Computer Science

Year/Semester
1/2

Area of study (specialization)
Intelligent Information Technologies

Profile of study
general academic

Level of study
Second-cycle studies

Course offered in
polish

Form of study
full-time

Requirements
compulsory

Number of hours

Lecture
30

Laboratory classes
30

Other (e.g. online)

Tutorials

Projects/seminars

Number of credit points

5

Lecturers

Responsible for the course/lecturer:
Professor Tadeusz Morzy, PhD
email: Tadeusz.Morzy@put.poznan.pl
tel: 61 665 2906
faculty: Faculty of Computing and Telecommunications
address: ul. Piotrowo 2, 60-965 Poznań

Responsible for the course/lecturer:

Prerequisites

The student should have basic knowledge regarding data mining algorithms and techniques. Additionally, basic knowledge of statistics, probability theory, and graph theory are expected.

The student should be able solve basic data mining tasks and know how to seek knowledge (research a given topic) on his/her own.

The student should also be willing to expand his/her competencies and be able to work as part of a team. Moreover, the student should showcase such characteristics as honesty, responsibility, perseverance, curiosity, creativity, respect for other people.

Course objective

1. Passing on knowledge about advanced data mining algorithms and working with complex data representations at various stages of the knowledge discovery process.

2. Developing problem solving skills related to the above-mentioned topics (through case studies related to complex data preprocessing and machine learning).

3. Developing practical skills through laboratories involving supervised classification tasks, unsupervised learning, time series analysis, or exploring social network data.

4. Acquiring knowledge about techniques and algorithms of knowledge discovery and pattern recognition, with special focus on relational and text data.

5. Promoting reproducible research related to the above-mentioned topics by using the R and python programming languages.

Course-related learning outcomes

Knowledge

The students have advanced knowledge regarding data mining, especially with respect to complex data representations.

They have organized, theoretically-grounded knowledge of data mining.

The knowledge concerns topics such as data stream mining, visual data analysis, natural language processing, distributed data mining, classification methods for evolving and streaming data.

They have detailed knowledge concerning data preprocessing techniques, continuous target predictions (regression methods, neural networks), classifier selection, parameter tuning, evaluation methods and metrics for classification and clustering.

They know the basic methods, techniques, and tools that can be used to solve complex data mining problems, they know the basic concepts of natural language processing and text mining, and they know how deploy the developed data mining models as part of an IT system.

Skills

The student is able to — while formulating and solving engineering tasks — integrate knowledge from different fields of computer science connected to data acquisition from different sources, its preprocessing and mining, as well as to the evaluation and practical application of the discovered patterns.

Knows how to assess the usefulness of new data mining algorithms, by reviewing scientific literature.

Is capable of formulating and solving simple research problems related to classification, regression, clustering and natural language processing.

Knows how to research information on data mining methods from literature, knowledge bases and other sources, knows how to integrate that information, interpret it, and critically assess, while formulating conclusions and opinions.

Knows how to critically assess existing data mining methods and propose improvements.

Knows how to perform simple research experiments, using the R and Python programming languages, and communicate experimental results through reproducible reports (using, e.g., knitr and jupyter).

Knows how to evaluate the pros and cons of selected data mining algorithms and their implementations, depending on the problem at hand.

Social competences

The students know the impact that data mining can have on solving practical tasks in companies, and its potential effect on entire societies.

The students understand that data mining is an ongoing field of study, and that one must keep learning to be up to date with the state-of-the-art.

Methods for verifying learning outcomes and assessment criteria

Learning outcomes presented above are verified as follows:

Lecture:

- questions connected to topics from previous lectures and whiteboard problem solving tasks

- a written exam involving multiple choice questions, short answer questions, problem-solving questions (students can bring their own learning materials). The exam consists of 5-6 questions worth 10 points each. The exams total point worth is 50-60 points. To pass the exam requires at least 50% of the maximum amount of available points.

Laboratories:

- tests with multiple choice and short-answer questions

- assessment of reports from two case study projects

- assessment of a web application (ML proof-of-concept system) programming task

Students can gain additional points by:

- presenting advanced topics during the lecture or laboratory class

- helping the lecturers improve study materials

- taking part in machine learning competitions

Programme content

Lecture:

Characteristics of the process of knowledge discovery from databases. Main methods of data preprocessing (in particular detecting conflicts when combining data from multiple sources, data cleaning, dealing with missing values), data reduction techniques (feature selection, feature extraction, data reduction for visualization, SVD), data transformations, and discretization methods. Web mining: link analysis. Web ranking algorithms (PageRank, HITS). Text mining: models and algorithms. Recommender systems.

Laboratories:

Data mining in R and Python. Tw projects (case studies), focusing on practical aspects of various steps of the knowledge discovery process. The classes include a basic R programming course, tutorials involving popular R packages for solving regression, classification, clustering, visualization, and data preprocessing problems. Students are also required to create a web application in R as a proof-of-concept machine learning system. Students get to know machine learning libraries in Python, with special focus on natural language processing (nltk, gensim, spaCy). Selected classes are devoted to visualization techniques and the methodology of reproducible research by using libraries such as caret, knitr (R) and scikit-learn, jupyter (Python). Students are also introduced to methods of tackling larger datasets by using parallel and distributed computation.

Students may be required to realize some of the topics mentioned above as part of their own work.

Teaching methods

Lecture: multimedia presentations, whiteboard examples, simple algorithmic tasks, software demos

Lab: multimedia presentations, problem solving, computational experiments, discussion, case studies, programming competitions, tests

Bibliography

Basic

A. Rajaraman, J. Lescovec, J.D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2014

Han J., Kamber M., Data Mining: Concepts and techniques, San Francisco, Morgan Kaufmann, 2000.

B. Liu, Web Data Mining: Exporing Hyperlinks, Contents, and Usage Data, Springer, 2015

Additional

Tufte, Edward R. The visual display of quantitative information. Vol. 2. Cheshire, CT: Graphics press, 2001.

Wickham, Hadley. Tidy data. Journal of Statistical Software 59.10 (2014): 1-23.

Ng, Andrew. Machine Learning Yearning, 2019.

Breakdown of average student's workload

	Hours	ECTS
Total workload	125	5,0
Classes requiring direct contact with the teacher	60	2,4
Student's own work (preparation for laboratory classes, preparation for exam, preparation of two projects and one web application) [1]	65	2,6

[1] delete or add other activities as appropriate