POZNAN UNIVERSITY OF TECHNOLOGY

EUROPEAN CREDIT TRANSFER AND ACCUMULATION SYSTEM (ECTS)

pl. M. Skłodowskiej-Curie 5, 60-965 Poznań

COURSE DESCRIPTION CARD - SYLLABUS

Course name

Data Mining

Course

Field of study	Year/Semester
Computing	1/1
Area of study (specialization)	Profile of study
Data Processing Technologies	general academic
Level of study	Course offered in
Second-cycle studies	Polish
Form of study	Requirements
full-time	elective

Number of hours

Lecture	Laboratory classes	Other (e.g. online)
15	15	
Tutorials	Projects/seminars	
15	15	

Number of credit points

5

Lecturers

Responsible for the course/lecturer:	Responsible for the course/lecturer:
prof. dr hab. inż. Tadeusz Morzy	dr hab. inż. Mikołaj Morzy, prof. PP
email: Tadeusz.Morzy@put.poznan.pl	email: Mikolaj.Morzy@put.poznan.pl
tel: +48 61 665 2906	tel: +48 61 665 2961
Faculty of Computing and Telecommunications	Faculty of Computing and Telecommunications
address: Piotrowo 2, 60-965 Poznań	address: Piotrowo 2, 60-965 Poznań

Prerequisites

Learning outcomes from first cycle studies: K1st_W1-8, K1st_U2-14, verified in the process of enrollment in second cycle studies - these effects are presented on the department's website. A student starting this course should have basic knowledge of database systems, statistics, probability, and combinatorial optimization.

Basic knowledge of Java and Python programming languages is required for the laboratory classes. The student should have the ability to solve basic problems in data processing and analysis, and the ability to obtain information from the indicated sources. He or she should also understand the need to expand his or her competence / have a willingness to cooperate as part of a team.

In terms of social competence, the student must present such attitudes as honesty, responsibility, perseverance, cognitive curiosity, creativity, personal culture, respect for other people.

Course objective

1. To provide students with basic knowledge of data mining, in terms of:

- association mining methods,

- sequential pattern mining,

- data classification,

- data clustering.

2. To develop in students the ability to solve data mining problems and discover knowledge from large data repositories.

3.To develop in students the skills of teamwork and integration of knowledge from different areas of computer science.

4. To develop in students the ability to formulate and test hypotheses related to engineering problems and simple research problems in data analysis and data mining.

Course-related learning outcomes

Knowledge

has advanced and in-depth knowledge in the field of information systems based on machine learning, theoretical foundations of their construction and methods, tools and programming environments used for their implementation (K2st_W1)

has structured and theoretically underpinned general knowledge related to key issues in statistics and computer science (K2st_W2)

has advanced detailed knowledge of data mining, machine learning, statistics and data processing (K2st_W3)

has knowledge of development trends and the most significant new developments in machine learning and data mining (K2st_W4)

knows advanced methods, techniques and tools used in solving complex engineering tasks and conducting research work in the area of data mining (K2st_W6)

Skills

is able to acquire information from literature, databases and other sources (in Polish and English), integrate them, interpret and critically evaluate, draw conclusions and formulate and fully justify opinions (K2st_U1)

is able to plan and carry out experiments, and interpret the obtained results and draw conclusions, as well as formulate and verify hypotheses related to complex personal and technical problems (K2st_U3)

is able to use analytical, simulation and experimental methods to formulate and solve engineering tasks and simple research problems (K2st_U4) is able - when formulating and solving engineering tasks - to integrate knowledge from different areas of computer science and statistics and apply a system approach, also taking into account non-technical aspects (K2st_U5)

is able to assess the usefulness and possibility of using new libraries for machine learning (K2st_U6)

is able to critically analyze existing machine learning processes and propose their improvements (K2st_U8)

is able - using, among others, machine learning methods - to solve complex computer tasks, including atypical tasks and tasks with a research component (K2st_U10)

Social competences

understands that in machine learning, knowledge, skills and tools become obsolete very quickly (K2st_K1)

understands the importance of using the latest knowledge in machine learning in solving research and practical problems (K2st_K2)

Methods for verifying learning outcomes and assessment criteria

Learning outcomes presented above are verified as follows:

Formative assessment:

(a) for lectures/exercises:

- on the basis of evaluations of the implemented exercises/tasks at the blackboard

b) in terms of laboratories / project classes:

- on the basis of the evaluation of the current progress of the tasks,

Summative evaluation:

a) in terms of lectures, verification of the established learning outcomes is realized by:

- evaluation of knowledge and skills demonstrated in an open problem-based written exam (the student can use any teaching materials), The exam consists of 6-8 problem-based tasks, for which 10 points can be obtained. A total of 60-80 points can be obtained. A passing grade of 3.0 requires 50% of the maximum number of points.

- Discussion of the results of the exam,

b) in terms of exercises, verification of the established learning outcomes is realized by:

a written colloquium of an open nature. Within the framework of the colloquium, 4-5 problem tasks should be solved. To pass the colloquium requires obtaining 50% of the maximum number of points.

c) in the field of laboratories, verification of the established learning outcomes is realized by:

- evaluation of the degree of assimilation of knowledge presented during the laboratory through a short quiz containing questions on the issues covered during the week of classes.

- realization of individual independent tasks of project or problem nature after each class,

d) in terms of project activities, verification of the assumed learning outcomes is realized by:

- realization of a larger task of a project or problem nature.

Obtaining additional points for activity during classes, especially for:

- correctly solving puzzles thematically related to statistics, machine learning and data mining,

- participation in international programming competitions, with special emphasis on teamwork.

Programme content

The lecture program covers the following topics:

Introduction to data mining: methods and applications. Association discovery: problem formulation and definition of association rules. A table of observations. Binary association discovery: association rule, rule evaluation measures. Apriori binary association rule discovery algorithm. FP-Growth binary association rule discovery algorithm. Closed and maximal association rules. Discovery of multilevel association rules. Discovery of multidimensional association rules. Binarization and discretization of data. Measures of attractiveness of association rules. Sequence data types. Discovery of sequential patterns: problem formulation. Basic sequential pattern discovery algorithm. Prefix sequential pattern discovery algorithm. Discovery of sequential patterns with time constraints - problem formulation. Algorithm for discovering sequential patterns with time constraints. Introduction to data classification. Methods of data classification. Data classification by decision tree induction. Decision tree induction algorithms using entropy measures and Gini index. The phenomenon of classifier overfitting. Decision tree pruning methods. Rule classifiers: definitions of basic concepts. Derivation of rule classifiers from decision trees. Sequential coverage algorithm and general algorithm for classifier rule extraction. Associative classification: problem definition. Association classification algorithms. Bayesian classifiers. Bayesian networks. Nearest neighbor classifier. Combination of classifiers. Quality assessment of classifiers: evaluation measures, ROC space and curve. Components of the clustering process. Definitions of measures of dissimilarity of objects. Classification of clustering methods. Hierarchical clustering: agglomerative and divisive. Algorithms of hierarchical grouping. Iterative-optimization grouping. Density grouping methods. Model-based methods. Grouping of objects described by categorical attributes. Detection of singular points.

Laboratory and project classes are conducted in the form of fifteen 2-hour meetings held in the laboratory. The laboratory program includes the following topics:

Data preprocessing for data mining processes: discretization, normalization, replacement of missing values, determination and elimination of outliers using RapidMiner, Orange Data Mining and KNIME environments as examples. Attribute preprocessing from the Python language level. Attribute validity assessment, attribute weighting methods, chi-square test, description length minimization (MDL) rule, attribute weighting using entropy. Discovery of association rules and Apriori and FP-Growth algorithms. Introduction to classification problems, partitioning a data set into a learning and testing set. Rule-based classifiers, simple tree classifiers, decision tree induction methods, measures for evaluating the quality of set partitioning: Gini index, entropy, Information Gain. Naive Bayes classifier, optimal Bayes classifier, Bayesian networks. Methods for evaluating and testing classifiers, multi-criteria evaluation of learned models. Lift, ROC, Precision-Recall measures in evaluating the quality of models. Learning classifiers using cost matrix. Family of SVM algorithms. Advanced classification methods: methods for aggregation of multiple models by voting, family of ensemble methods, multilayer classifiers. Basic cluster analysis algorithms, practical limitations of k-means and k-medoids algorithms, density-based cluster analysis algorithms, EM family of cluster analysis algorithms. A low-level programming interface for machine learning in Python Sci-Kit language. Feature extraction methods: family of PCA, SVD and NNMF algorithms.

Teaching methods

Lecture: multimedia presentation illustrated by examples given on the blackboard.

Exercises: multimedia presentation illustrated by examples given on the blackboard and performance of tasks given by the instructor.

Laboratory: independent work on the basis of examples provided by the instructor, tutorials, quizzes, tasks to be carried out independently

Project classes: independent work in project groups.

Bibliography

Basic

1. Eksploracja danych: metody i algorytmy, T. Morzy, PWN, 2013.

2. Introduction to Data Mining, Tan, P-N., Steinbach, M., Kumar, V., Pearson Education, 2006.

3. Data Mining: Concepts and Techniques, Han, J., Kamber, M., Pei, J., Morgan Kaufmann, 2012.

4. Systemy uczące się, Cichosz, P., WNT, 2000.

5. Data Mining: Practical Machine Learning Tools and Techniques, Witten, I., Frank, E., Morgan Kaufmann, 2005.

Additional

1. Statystyczne systemy uczące się, Koronacki, J., Ćwik, J., WNT, 2005.

2. Uczenie maszynowe i sieci neuronowe, Krawiec, K., Stefanowski, J., Wydawnictwo PP, 2003.

3. Programmer’s Guide to Data Mining, Zacharski, R. http://guidetodatamining.com/

4. Machine Learning, Ng, A., https://www.coursera.org/course/ml

Breakdown of average student's workload

	Hours	ECTS
Total workload	125	5,0
Classes requiring direct contact with the teacher	60	3,0
Student's own work (literature studies, preparation for exercises, laboratory and project classes, writing and testing programs, exam preparation) [1]	65	2,0

[1] delete or add other activities as appropriate