COURSE DESCRIPTION CARD - SYLLABUS

Course name
Data mining [S1SI1E>EDAN]

Course

Field of study
Artificial Intelligence

Year/Semester
2/4

Area of study (specialization)
–

Profile of study
general academic

Level of study
first-cycle

Course offered in
English

Form of study
full-time

Requirements
compulsory

Number of hours

Lecture
30

Laboratory classes
30

Other
0

Tutorials
0

Projects/seminars
0

Number of credit points

5,00

Coordinators

dr hab. inż. Mikołaj Morzy prof. PP
mikolaj.morzy@put.poznan.pl

Lecturers

Prerequisites

A student starting this course should have basic knowledge of database systems, statistics, probability, and combinatorial optimization. Basic knowledge of 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: - data types, similarity and distance measures, - association mining methods, - sequential pattern mining, - 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:
- on the basis of evaluations of the implemented exercises/tasks at the blackboard
b) in terms of laboratories:
- 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.
- additional points for presence during lectures
- additional 20% of points obtained for the laboratories
- discussion of the results of the exam,

b) 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,
- 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

This course provides a comprehensive introduction to the principles and techniques of data mining. Students will learn about various data types and preprocessing methods, dimensionality reduction, association rule mining, sequence and graph pattern mining, and clustering algorithms. The course covers both foundational concepts and advanced topics, equipping students with the skills to discover valuable knowledge from large datasets. Topics include data types, graph and high-dimensional data analysis, kernel methods, dimensionality reduction techniques, association rule mining, sequence mining, graph pattern mining, distance and similarity measures, various clustering algorithms, and cluster evaluation.

List of lectures

* Introduction to data mining: This lecture introduces the core concepts of data mining, its importance in extracting knowledge from large datasets, and the data mining process.
* Data types: This lecture explores the different types of data, including quantitative, qualitative, and specific types like numerical, nominal, and text data, and their relevance in data mining
* Graph data & high-dimensional data: This lecture covers the representation and analysis of graph data, the challenges of high-dimensional data, and dimensionality reduction techniques
* Kernel methods
* Dimensionality reduction (PCA, SVD, TSNE and UMAP): This lecture focuses on techniques for reducing the dimensionality of data while preserving essential information, including PCA, SVD, TSNE, and UMAP.
* Association rules I (frequent itemsets, association rules, apriori): This lecture introduces association rule mining, focusing on frequent itemsets, association rules, and the Apriori algorithm.
* Association rules II (prefix-span, evaluation of rules, interestingness measures): This lecture builds upon association rules, covering the PrefixSpan algorithm, rule evaluation, and interestingness measures.
* Sequence mining
* Graph pattern mining
* Distance and similarity: This lecture covers methods for measuring distance and similarity between data points, which are fundamental to many data mining algorithms.
* Representative-based clustering (k-means/medoids, kernel-k-means, EM clustering): This lecture explores clustering algorithms that use representative points, such as k-means, k-medoids, kernel k-means, and EM clustering.
* Hierarchical & density-based clustering: This lecture introduces hierarchical and density-based clustering methods for grouping data points based on their relationships.
* Spectral & graph clustering
* Clustering evaluation: This lecture focuses on methods for evaluating the quality and effectiveness of clustering results.

Course topics

This course provides a comprehensive introduction to the principles and techniques of data mining. Students will learn about various data types and preprocessing methods, dimensionality reduction, association rule mining, sequence and graph pattern mining, and clustering algorithms. The course covers both foundational concepts and advanced topics, equipping students with the skills to discover valuable knowledge from large datasets. Topics include data types, graph and high-dimensional data analysis, kernel methods, dimensionality reduction techniques, association rule mining, sequence mining, graph pattern mining, distance and similarity measures, various clustering algorithms, and cluster evaluation.

List of lectures

* Introduction to data mining: This lecture introduces the core concepts of data mining, its importance in extracting knowledge from large datasets, and the data mining process.
* Data types: This lecture explores the different types of data, including quantitative, qualitative, and specific types like numerical, nominal, and text data, and their relevance in data mining
* Graph data & high-dimensional data: This lecture covers the representation and analysis of graph data, the challenges of high-dimensional data, and dimensionality reduction techniques

* Kernel methods
* Dimensionality reduction (PCA, SVD, TSNE and UMAP): This lecture focuses on techniques for reducing the dimensionality of data while preserving essential information, including PCA, SVD, TSNE, and UMAP.
* Association rules I (frequent itemsets, association rules, apriori): This lecture introduces association rule mining, focusing on frequent itemsets, association rules, and the Apriori algorithm.
* Association rules II (prefix-span, evaluation of rules, interestingness measures): This lecture builds upon association rules, covering the PrefixSpan algorithm, rule evaluation, and interestingness measures.
* Sequence mining
* Graph pattern mining
* Distance and similarity: This lecture covers methods for measuring distance and similarity between data points, which are fundamental to many data mining algorithms.
* Representative-based clustering (k-means/medoids, kernel-k-means, EM clustering): This lecture explores clustering algorithms that use representative points, such as k-means, k-medoids, kernel k-means, and EM clustering.
* Hierarchical & density-based clustering: This lecture introduces hierarchical and density-based clustering methods for grouping data points based on their relationships.
* Spectral & graph clustering
* Clustering evaluation: This lecture focuses on methods for evaluating the quality and effectiveness of clustering results.

Teaching methods

Lecture: multimedia presentation illustrated by examples given on the blackboard and using Python notebooks.
Laboratory: independent work on the basis of examples provided by the instructor, tutorials, quizzes, tasks to be carried out independently, 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	5,00
Classes requiring direct contact with the teacher	62	2,50
Student's own work (literature studies, preparation for laboratory classes/ tutorials, preparation for tests/exam, project preparation)	63	2,50