AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY

Code:	UBPJO-233	Module name:	Fundamentals of Data Science		
Academic year:	2017/2018	Semester:	Spring	ECTS credits:	5
Programme:	Physics and Applied Computer Science				
Course homepage:	http://home.agh.edu.pl/~slukasik/#teaching	Lecture language:	English		
Responsible teacher:	dr inż. Łukasik Szymon (slukasik@agh.edu.pl)				
Academic teachers:	dr inż. Łukasik Szymon (slukasik@agh.edu.pl)				

Module summary

This course aims at presenting fundamental problems of contemporary data science, namely: data reduction, outlier detection, cluster analysis and classification along with their real-world instances.

Description of learning outcomes for module

MLO code	Student after module completion has the knowledge/ knows how to/is able to	Method of learning outcomes verification (form of completion)
Social competence		
M_K001	The student is able to plan and perform work in a team that is responsible for creative activities	Project, Execution of a project
Skills		
M_U001	Student is able to select proper technique of data analysis and establish suitable parameters	Execution of laboratory classes, Execution of a project
M_U002	Student is able to apply basic procedures of data analysis and critically assess obtained results	Execution of laboratory classes, Execution of a project
Knowledge		
M_W002	Student has basic knowledge about procedures of data analysis	Project, Execution of laboratory classes
M_W003	Student has elementary knowledge about practical issues related to the application of data mining procedures	Execution of a project, Completion of laboratory classes

FLO matrix in relation to forms of classes

MLO code	Student after module completion has the knowledge/ knows how to/is able to	Form of classes										
		Lectures	Auditorium classes	Laboratory classes	Project classes	Conversation seminar	Seminar classes	Practical classes	Fieldwork classes	Workshops	Others	E-learning
Social competence												
M_K001	The student is able to plan and perform work in a team that is responsible for creative activities	-	-	-	+	-	-	-	-	-	-	-
Skills												
M_U001	Student is able to select proper technique of data analysis and establish suitable parameters	-	-	+	+	-	-	-	-	-	-	-
M_U002	Student is able to apply basic procedures of data analysis and critically assess obtained results	-	-	+	+	-	-	-	-	-	-	-
Knowledge												
M_W002	Student has basic knowledge about procedures of data analysis	+	-	-	-	-	-	-	-	-	-	-
M_W003	Student has elementary knowledge about practical issues related to the application of data mining procedures	+	-	-	-	-	-	-	-	-	-	-

Module content

Lectures

1. Introduction to Data Science – history and methodological background.
2. Typical workflow of data analysis.
3. Data preprocessing – data reduction, cleaning, handling missing elements.
4. Unsupervised learning – outlier detection and cluster analysis.
5. Classification and regression.
6. Recommender systems and text mining.
7. Big Data – brief overview of issues and computational methods.

Laboratory classes

Laboratory exercises illustrating selected problems of data analysis:
1. Outlier detection
2. Data dimensionality reduction.
3. Cluster analysis.
4. Classification.

5. Recommender systems and text mining.

Project classes

Group projects aimed at familiarizing with real-world data science problems.

Method of calculating the final grade

Weighted average of laboratory exercises' final grade (weight 2/3) and project's grade (weight 1/3).

Prerequisites and additional requirements

Programming in one of the following programming languages: C/C++, Python, Java, MATLAB, R.

Recommended literature and teaching resources

- Jiawei Han, Jian Pei,Micheline Kamber, "Data Mining: Concepts and Techniques", Elsevier, 2011.
- Jake VanderPlas, "Python Data Science Handbook: Essential Tools for Working with Data". O'Reilly Media, 2016.
- Hadley Wickham, "R for Data Science: Import, Tidy, Transform, Visualize, and Model Data", O'Reilly Media, 2017.
- Jeffrey Solka, Angel R. Martinez, "Exploratory Data Analysis with MATLAB", Chapman & Hall, 2017.
- UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/
- Kaggle Competitions, https://www.kaggle.com/competitions

Scientific publications of module course instructors related to the topic of the module

- S. Łukasik, A. Moitinho, P.A. Kowalski, A. Falcão, R.A. Ribeiro, P. Kulczycki, „Survey of Object-Based Data Reduction Techniques in Observational Astronomy", Open Physics, vol. 14, pp. 578-586, 2016.
- D. Domańska, S. Łukasik, "Handling high-dimensional data in air pollution forecasting tasks", Ecological Informatics, vol. 34, pp. 70-91, 2016.
- M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Łukasik, "Discrimination of Wheat Grain Varieties Using X-Ray Images", in: Information Technologies in Biomedicine, E. Pietka, P. Badura, J. Kawa, W. Wieclawek (eds.), Springer-Verlag, Berlin-Heidelberg, 2016, pp. 39-50.
- P. Kulczycki, S. Łukasik, "An Algorithm for Reducing Dimension and Size of Sample for Data Exploration Procedures", International Journal of Applied Mathematics and Computer Science, vol. 24, pp. 133-149,
- P. Kulczycki, M. Charytanowicz, P.A. Kowalski, S. Łukasik, "Exemplary Applications of the Complete Gradient Clustering Algorithm in Bioinformatics, Management and Engineering", in: „Issues and Challenges of Intelligent Systems and Computational Intelligence", L.T. Kóczy, C. Pozna, J. Kacprzyk (eds.), Springer, pp. 119-132, 2014.
- S. Łukasik, P. Kulczycki, "Using Topology Preservation Measures for Multidimensional Intelligent Data Analysis in the Reduced Feature Space", Lecture Notes in Artificial Intelligence, vol. 7895, pp. 184-193, 2013.
- S. Łukasik, M. Haręza, M. Kaczor, "Document content mining for authors' identification task", Technical Transactions: Automatic Control, vol. 1-AC, pp. 3-15, 2013.
- P. Kulczycki, M. Charytanowicz, P.A. Kowalski, S. Łukasik, "The Complete Gradient Clustering Algorithm: Properties in Practical Applications", Journal of Applied Statistics, vol. 39, pp. 1211-1224, 2012.

Additional information

Laboratory classes are obligatory. One unjustified absence is allowed in the case of these classes. Absences (also justified) in laboratory classes need to be reworked in the form and time agreed with the instructor. Half of the unjustified classes result in a lack of credit. From that decision the student teacher may appeal to the instructor and/or the Dean.

Student workload (ECTS credits balance)

Student activity form	Student workload
Participation in lectures	15 h
Participation in laboratory classes	15 h
Completion of a project	30 h
Preparation of a report, presentation, written work, etc.	15 h
Participation in project classes	15 h
Contact hours	15 h
Preparation for classes	30 h
Summary student workload	135 h
Module ECTS credits	5 ECTS