Warsaw University
of Technology

COURSE OFFERED IN THE DOCTORAL SCHOOL

Code of the course	4606-ES-000000C-0004	Name of the course	Polish	**Analiza strumieni danych. Metody uczenia maszynowego dla strumieni danych.**
			English	**Stream mining. Machine learning methods for data streams.**
Type of the course	Expertise course			
Course coordinator	dr hab. inż. Maciej Grzenda, prof. uczelni/Maciej Grzenda, PhD, DSc			
Implementing unit	Faculty of Mathematics and Information Science	Scientific discipline / disciplines*	Information and communication technology	

Level of education	PhD studies	Semester	Winter/~~summer~~		
Language of the course	English				
Type of assessment:	Graded assessment based on written test outcomes (no exam)	Number of hours in a semester	30	ECTS credits	2
Minimum number of participants	3	Maximum number of participants	15	Available for students (BSc, MSc)	Yes/~~No~~

Type of classes		Lecture	Auditory classes	Project classes	Laboratory	Seminar
Number of hours	in a week	2	-	-	-	-
	in a semester	30	-	-	-	-

* does not apply to the Researcher's Workshop

1. Prerequisites

Theoretical and practical knowledge of standard data mining and machine learning methods dedicated for data sets such as decision trees, random forest, naïve Bayes classifier.

2. Course objectives

In an increasing number of cases, data can be systematically acquired and processed as data streams rather than as periodically collected static datasets with a finite content. Examples include data obtained from measurement devices in industrial environments or vehicle location data streams. The aim of the course is to familiarize participants with the issues of machine learning methods dedicated to data streams, including learning in non-stationary setting (concept drift), with particular emphasis on real conditions such as e.g. limited availability of labels and verification latency.

3. Course content (separate for each type of classes)

Lecture

1. Challenges related to data stream processing (memory limitations, potentially infinite number of examples to be processed and others); stream mining methods vs. analysis of large-scale data (Big Data)
2. Batch machine learning methods versus incremental and online learning.
3. Basics of evaluating models dedicated to data streams with particular emphasis on the evaluation of classification models. Simple reference methods (e.g. NoChange, Majority Class methods).
4. The problem of concept drift, concept drift types, the assessment of drift magnitude, the influence of concept drift on the learning process. Methods of adaptation to concept drift - passive and active approach.
5. Accuracy and other measures to evaluate the model on the example of classification tasks in the batch and stream approach.
6. Adaptation of classical methods for learning and evaluating models with the use of data streams on the example of k Nearest Neighbors methods. Introduction to machine learning methods dedicated to data streams on the example of Hoeffding trees.
7. Machine learning methods dedicated to data streams, including those that take into account concept drift, eg adaptive random forest.
8. Advanced aspects of data stream processing: partially labeled data, delayed labels. Selected advanced aspects in the streaming setting: semi-supervised learning, active learning.

9. Advanced aspects of the assessment of stream methods, including evaluation under conditions of delayed labels. 10. Selected issues of data stream pre-processing, including data reduction. 11. Review of Big Data platforms related to data stream processing on the example of Apache NiFi, Apache Kafka and Apache Spark. 12. Review of reference libraries dedicated to data streams, real stream data and synthetic stream generators. 13. Data stream processing and architectural patterns dedicated to large-scale data storage and analysis environments. During two lecture meetings there will be written test (mid-semester and final).
Laboratory
-

4. Learning outcomes			
	Learning outcomes description	Reference to the learning outcomes of the WUT DS	Learning outcomes verification methods*
Knowledge			
K01	The student knows the basic aspects of machine learning for data streams	SD_W2	written test
K02	The student knows the key methods of detecting and responding to concept drift and exemples of machine learning methods dedicated to learning in non-stationary environments	SD_W3	written test
K03	The student knows selected advanced issues and open topics in the field of machine learning methods dedicated to data streams	SD_W3	written test
Skills			
S01	The student is able to assess the consequences of replacing learning in batch mode with learning in stream mode, identify limitations related to learning in batch mode and challenges related to learning with the use of data streams for machine learning	SD_U2	written test
S02	The student knows how to evaluate stream mining methods.	SD_U2	written test
Social competences			
SC01	The student understands the need for lifelong learning and competence development, including the use of scientific literature on the example of the development of competences related to the analysis of static and finite data sets in terms of the analysis of potentially infinite data streams describing processes that change over time	SD_K1, SD_K2	written test

*Allowed learning outcomes verification methods: exam; oral exam; written test; oral test; project evaluation; report evaluation; presentation evaluation; active participation during classes; homework; tests

5. Assessment criteria
The assessment will be based on two written tests, each marked on a scale of 0-50 points. The final grade depends on the total number of points obtained in both tests and is determined according to the following rules: 0-50 points - 2.0, 51-60 points - 3.0, 61-70 points - 3.5, 71-80 points - 4.0, 81-90 points - 4.5, 91-100 points - 5.0.

6. Literature

Key bibliography:

[1]. Bifet A. et al, Machine Learning for Data Streams with Practical Examples in MOA, MIT Press, 2018

[2]. Kleppmann, M., Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems, O'Reilly, 2017

[3]. Marz N., Warren James, Big Data. Principles and best practices of scalable realtime data systems, Manning, 2015

Additional bibliography:

[4]. Grzenda M., Gomes H. M., Bifet A.: Delayed labelling evaluation for data streams, Data Mining and Knowledge Discovery, 2020, vol. 34, s.1237- 1266. DOI:10.1007/s10618-019-00654-y

[5]. Research papers on stream mining

7. PhD student's workload necessary to achieve the learning outcomes**

No.	Description	Number of hours
1	Hours of scheduled instruction given by the academic teacher in the classroom	30
2	Hours of consultations with the academic teacher, exams, tests, etc.	5
3	Amount of time devoted to the preparation for classes, preparation of presentations, reports, projects, homework	5
4	Amount of time devoted to the preparation for exams, test, assessments	20
	Total number of hours	**60**
	ECTS credits	**2**

** 1 ECTS = 25-30 hours of the PhD students work (2 ECTS = 60 hours; 4 ECTS = 110 hours, etc.)