DOCTORAL (PHD) STUDIES

COURSE UNIT DESCRIPTION

Course unit title	Scientific areas	Faculty	Institute, department
Nonlinear statistical models for massive data analysis (7 ECTS, 190 h)	Informatics (09 P), Mathematics (01P), Engineering of Informatics (07T)	Faculty of Mathematics and Informatics	Institute of Data Science and Digital Technologies

Study method	Number of credits	Study method	Number of credits
Lectures	1 ECTS (30 h)	Consultations	1 ECTS (30 h)
Individual works	4 ECTS (100 h)	Seminars	1 ECTS (30 h)

Summary

The aim of the subject is to supplement students' knowledge of machine learning with knowledge of nonlinear statistical modelling, emphasizing critical statistical thinking.

In the context of the course, the Big Data is understood as data collected in massive volume and without a specific purpose and it is called massive data. The latter data is, as a rule, heterogeneous from simple small text entries to minutely stock information or whole genome data. The data can be analysed by treating them as Black Box and using machine learning methods. On other hand it is possible to apply statistical modelling and, taking into account data generation process, make assumptions about the data distribution and test them. The more specific (rigorous) assumptions are in line with the data, the more meaningful and subtle the interpretation and inference of the results obtained. Therefore, it is meaningful to combine machine learning methods with nonlinear (parametric and nonparametric) statistical modelling methods.

Main topics:
• Classic and Robust Linear Methods (Agresti)
• Generalized linear models (GLM), model selection and statistical inference (Agresti)
• Repeated measurements, random factors and longitudinal observations (Faraway)
• Mixed factor models, not normal response variable (Faraway)
• Bayesian hierarchical modeling, Monte Carlo Markov chains (Madigan)
• Regularization of GLM (Agresti)
• Non-parametric regression (Faraway), generalized additive models and smoothing methods (Agresti, Faraway)

Student skills. During the course, students, after analysing indicated literature, are able to realize the generalized linear and nonparametric regression models for the selected data. When needed, student is able to adapt the methods of regularization for massive data analysis, respectively Bayesian methods for the investigation of complex structures. Students who have completed the course are aware of and are able to evaluate the uncertainty and reliability of the data and their analysis results. The methods are implemented in R language or another program which is mutually acceptable to both the student and teacher.

Assessment methods
The assessment of the course consists of two parts - problem solving and a scientific report. Both parts are graded on a 10-point scale. The solution of problems is 40% final evaluation.
Using the accumulated knowledge, the student solves the following problems from Wasserman's book:
• Chapter 15.6 from Task 4;
• Task 7 of Chapter 21.7;
• Tasks 3 and 6 of Chapter 22.13

• Problem 5 of Chapter 24.7. After receiving a positive evaluation from the solution of the tasks, the student writes a scientific paper. The paper should consist of the following parts: 1. Brief presentation of the dissertation in preparation (why are nonlinear models important in your dissertation?). 2. Analysis of scientific articles from the field of the dissertation, in which nonlinear models are applied (which models and for which problems to apply?) 3. The empirical part. Selection and implementation of nonlinear models for the selected data set (data description, exploratory analysis, selection and implementation of models, description of the obtained results and discussion). The selected data array and the software code implementing the models must be available to the course instructors.
Main literature
Agresti, A., 2015. Foundations of linear and generalized linear models. John Wiley & Sons. http://bayanbox.ir/view/7443147326514856944/Foundations-of-Linear-and-Generalized-Linear-Models-Wiley-Series-in-Probability-and-Statistics-Alan-Agresti-2015.pdf
Faraway, J.J., 2016. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models (Vol. 124). CRC press.
Madigan, D., Ridgeway G., 2002. Bayesian Data Analysis for Data Mining. https://pdfs.semanticscholar.org/58be/87292ce8c4a0eefca6dd5430368f4af4e177.pdf
Wasserman, L., 2004. All of statistics: a concise course in statistical inference (Vol. 26). New York: Springer.
Complementary literature
Alpaydin, E., 2010. Introduction to Machine Learning. The MIT Press. ISBN-10: 0-262-01243-X, ISBN-13: 978-0-262-01243-0. http://cs.du.edu/~mitchell/mario_books/Introduction_to_Machine_Learning_-_2e_-_Ethem_Alpaydin.pdf
Han, J., Kamber, M., 2006. Data Mining Concepts and Techniques. 2nd ed. CA: Morgan Kaufmann Publishers is an imprint of Elsevier. http://ccs1.hnue.edu.vn/hungtd/DM2012/DataMining_BOOK.pdf https://pdfs.semanticscholar.org/02e0/bc77460469aefec5bd794ee6c4efc15e6adb.pdf

Lecturer(s) (name, surname)	Science degree	Main publications
Audronė Jakaitienė	PhD.	http://www.elaba.mb.vu.lt/dmsti/?aut=Audronė+Jakaitienė
Marijus Radavičius	PhD.	http://www.elaba.mb.vu.lt/dmsti/?aut=Marijus+Radavičius
Olga Kurasova	PhD.	http://www.elaba.mb.vu.lt/dmsti/?aut=Olga+Kurasova