satRday Gdańsk 2019

17-18:05:2019

Book of abstract

Organizing Committee

Olgun Aydin - Tricity R Users Group, trigeR
Michał Maj - Tricity R Users Group, trigeR
Anna Rybińska-Fryca - Tricity R Users Group, trigeR
Marcin Kosiński - Why R? Foundation
Patryk Jasik - Gdańsk University of Technology, Faculty of Applied Physics and Mathematics
Karol Flisikowski - Gdańsk University of Technology, Faculty of Management and Economics

POLITECHNIKA
GDAŃSKA
WYDZIAŁ FIZYKI TECHNICZNEJ
I MATEMATYKI STOSOWANEJ

WHY R?
FOUNDATION

trige(R)

WYDZIAŁ ZARZĄDZANIA I EKONOMII
POLITECHNIKA GDAŃSKA

Sponsors

The organizers would like to thank the following companies for sponsoring this meeting:

Partners

The organizers would like to thank the following institutions and initiatives for supporting this meeting:

Koło Naukowe
Metod
Ilościowych

ClickMeeting

SmarterPoland.pl

aiBAY

Program

Friday, May 17th **Workshops**			
15:30–16:00	**Registration** (Hall)		
16:00–20:00	**Modern and beautiful dashboards: building Shiny apps using SemanticUI components** **Dominik Krzemiński** (Room 463)	**Introduction to Deep Learning in R with Keras** **Michał Maj** (Room 252)	**Introduction to Tidyverse** **Jan Wasilewski** **Jakub Borkowski** (Room 202)

Saturday, May 18th **Conference**	
HOUR	**PROGRAM**
8:00-9:00	**Registration** (Hall)
9:00-9:30	**Few words from organizers** (Room 252)
9:30-10:00	**Sponsors session** (Room 252)
10:00-11:00	**Joint modeling and dynamic predictions with applications to cancer research using R package frailtypack**

	Agnieszka Król (Room 252)	
11:00-11:30	**Coffee break** (Hall)	
11:30-12:00	**predPCR: an automated classification of qPCR curves** **Michał Burdukiewicz** (Room 252)	**AmyloGram: analysis of proteins in R** **Jarosław Chilimoniuk** (Room 211)
12:00-12:30	**Nonnegative Matrix Factorization as a Tool to Segment Respondents in a High Dimensional Survey** **Marcin Kosiński** (Room 252)	**A Case Study on Machine Learning Classification Algorithms in R** **Olgun Aydin & Ezgi Nazman** (Room 211)
12:30–13:30	**Machine Learning meets Design, Design meets Machine Learning** **Przemysław Biecek & Hanna Piotrowska** (Room 252)	
13:30-15:00	**Lunch break** (Hall)	
15:00-15:30	**Tuning & Bootstrapping Performance of ML Model** **Monika Nawrocka** (Room 252)	**Sit, relax, monitor: How to maintain models and how R can help?** **Natalia Reszka** (Room 211)
15:30-16:30	**Drawing ROC curves** **Błażej Kochański** (Room 252)	
16:30-17:00	**Coffee break** (Hall)	

17:00-17:30	**A shiny application enabling facial attractiveness evaluation for purposes of plastic surgery** **Lubomír Štěpánek** (Room 252)	**Elasticsearch and R - deal with it!** **Bartłomiej Staszkiewicz** (Room 211)
17:30-18:00	**Bayesian inference in big data analysis** **Katarzyna Sidorczuk** **HaDeX - analysis of HDX-MS data** **Weronika Puchała** **drake: reproducible workflow management in R** **Dominik Rafacz** (Room 252)	
18:00-18:15	**Farewell**	
18:30	**Guided tour**	
20:00	**Evening paRty!**	

Invited Speakers

Joint modeling and dynamic predictions

Agnieszka Król

In the medical research different kinds of patient information are gathered over time together with clinical outcome data such as overall survival (OS): Joint models enable analysis of correlated data of different types such as individual repeated data together with OS: The repeated data may be recurrent events e:g:, appearance of new lesions or a longitudinal outcome called biomarker e:g:, tumor size: Moreover, joint models are useful for individual dynamic predictions of death using a patient’s history: The talk will introduce joint frailty models for recurrent events and a terminal event as well as present a multivariate joint model for longitudinal marker, recurrent events and survival in the context of treatment evaluation in cancer clinical trials: Presentation of implementing these methods will be given using the R package frailtypack:

Machine learning meets design, design meets machine learning

Przemysław Biecek, Hanna Piotrowska

We are surrounded by machine learning models: They decide which articles we see, which ads are displayed, what kind of shop offers, movies, clips, medical therapies are recommended for us: But, do we understand how these decisions are being made? Machine learning models are complex: A single decision is sometimes calculated based on millions of parameters: How we can possibly understand them? We desperately need an effective language to explore, explain, debug and validate complex models: During this talk we will present a visual language designed to explore predictive models: We will show visual elements of this language, describe how they were designed, and how they can be used:

Drawing ROC curves

Błażej Kochański

Sometimes you want to draw a ROC curve while not having a classification model (yet): For example: you consider investing in a new credit scoring system: You hear you may increase your AUROC (area under ROC) by 5 percentage points: In order to translate it into financial results, you need to draw a nonexistent ROC curve: I will review available options of modelling ROC curves for credit scoring and show which fit the real data best:

Regular Talks

predPCR: an automated classification of qPCR curves

Michał Burdukiewicz

Quantitative Real-Time PCR (qPCR) is a simple and high-throughput method common in research, forensics, and medicine. The popularity of qPCR resulted in the plethora of software devoted to the analysis of qPCR data, including R packages. Despite that fact, there are no methods for automated assessment of qPCR results. Thus, we propose the predPCR tool for reproducible classification of qPCR data. Our model benefits not only from advancements in machine learning, as Bayesian optimization of hyperparameters, but also from theoretical advancements in the field of qPCR analysis. A combination of these two factors results in a powerful (AUC ~ 0.95) yet simple to interpret model. We present the complete analytical workflow, along with supplementary tools explaining the decision-making process of predPCR tailored for users less fluent in machine learning. The machine learning approach enabled reliable, scalable, and automated qPCR curve classification with broad potential clinical and epidemiological applications. A web server is available from http://www.smorfland.uni.wroc.pl/shiny/predPCR/.

AmyloGram: analysis of proteins in R

Jarosław Chilimoniuk

The structure and therefore function of proteins are encoded in the linear sequence of amino acids. Our toolkit, the biogram R package, provides a set of useful tools for encoding protein sequences into features understandable by machine learning algorithms. Our software, inspired by natural language processing, extracts n-grams of amino acids from proteins and selects only the most informative ones using developed by us Quick Permutation Test (QuiPT). We present advantages of our approach using AmyloGram, an R package and shiny server (link) for prediction of amyloids, proteins associated with the number of clinical disorders (e.g., Alzheimer's, Creutzfeldt-Jakob's and Huntington's diseases). Amyloid proteins are extremely diverse sequence-wise, but all of them can undergo a unique self-aggregation. AmyloGram effectively recognizes patterns responsible for this behavior (AUC = 0.8972) outperforming existing amyloid-predicting software. Moreover, predictions of AmyloGram were verified experimentally as our tool led to the discovery of a novel amyloid protein, MspA, produced by Methanospirillum hungatei JF-1. www.smorfland.uni.wroc.pl/shiny/AmyloGram/

Nonnegative Matrix Factorization as a Tool to Segment Respondents in a High Dimensional Survey

Marcin Kosiński

From the current segmentation one requires them to follow following features: it should be balanced, segments should be distinctive, the discovered over and under indexed features within segments should create a meaningful story, and in the best case the amount of differentiate factors that drives segmentation should be small. The last requirement often is a bottleneck in the scenario of a survey where respondents are asked enormous amount of question. The solution, one from many, to this use case can be the nonnegative matrix factorization that in one attempt segments respondents and their features! I'll present concept of the NMF decomposition and I'll present applications in R, with the explanation of diagnostic plots. Working with high dimensional data? Often facing the need to group observations? That's a good presentation for you.

A case study on machine learning classification algorithms in R

Olgun Aydin

With the continuous development of machine learning packages in R, statistical classification algorithms have been widely preferred to learn the performance of discrete label output (groups) with the help of these packages in many fields. On the other hand, researchers can confront with some data sets which have no groups while it is available to be grouped by many other statistical methods. Data Envelopment Analysis (DEA), is a non-parametric method based on linear programming principles, can generate the grouped data as efficient and inefficient. It is also important to evaluate this grouping results with classification algorithms. As it is known, Logistic Regression, Decision Trees, Artificial Neural Networks, Support Vector Machine the algorithms are most widely used classification methods in literature. With this respect, we aimed to compare these classification performances on DEA grouping result. First, we obtained data set from Social Progress Index includes data from only 123 non-OECD countries on 7 indicators which consist of undernourishment (% of population; 5 signifies ≤5), depth of food deficit (calories/undernourished person), deaths from infectious disease (deaths/100.000), traffic deaths (deaths/100.000), greenhouse gas emission (CO2 equivalents per GDP), maternal mortality deaths (deaths/100.000 live births), and life expectancy (at years). The undernourishment, depth of food deficit, deaths from infectious disease, and traffic deaths were considered as related indicators on life expectancy. Second, we constructed an input oriented Charnes-Cooper-Rhodes (CCR) model (Charnes, Cooper, and Rhodes, 1978), where the life expectancy is output and other indicators are inputs. As a result, efficient and inefficient countries were obtained in terms of these indicators using Benchmarking package in R. Finally, we evaluated the grouping results in terms of accuracy rate using several classification algorithms such as Logistic Regression, Decision Tree, Artificial Neural Networks, Support Vector Machine on this grouped data set using glm, keras, caret packages in R.

Tuning & Bootstrapping Performance ML Model

Monika Nawrocka

Building Machine Learning models seems to be simple. We have many easily applicable libraries for different purpose functions. However, it does not always turn out that our models bring the expected results. Sometimes it even happens that the learned model on the training set is evaluated high, and after testing on new sets, it incorrectly assesses the reality. Tuning hyperparameters and modeling by re-sampling the sample data and performing inferences on the sample from the re-tested data may be a solution to improve the quality of already built models. The main advantage of such bootstrap samples is the measurability of inference based on a real sample from the re-tested data.

Sit, relax, monitor. How to maintain models and how R can help?

Natalia Reszka

In today's financial institutions, analytical models are high-value strategic assets. As models are needed to run the business and keep up with regulations, they must be managed for optimal performance once in production. Model performance can go down over time no matter how good model is. In this talk, I will discuss best practices for preventing the output disaster from a data science perspective.

A Shiny application enabling facial attractiveness evaluation for purposes of plastic surgery

Lubomír Štěpánek

The ways how to evaluate facial attractiveness complexly and how to make comparisons between facial images of patients before and after facial plastic surgery procedure are still unclear and require ongoing research. In this study, we have developed a web-based shiny application providing facial image processing, both manual and automated landmarking, facial geometry computations and machine-learning models allowing to identify geometric facial features associated with an increase of facial attractiveness after undergoing rhinoplasty, common facial plastic surgery. Patients' facial image data were processed, landmarked and analysed using the application. Facial attractiveness was measured using Likert scale by a board of independent observers. Machine-learning built-in approaches were performed to select predictors increasing facial attractiveness after undergoing rhinoplasty. The shiny web framework enables to develop a complex web interface including HTML, CSS and javascript front-end and R-based back-end bridging C++ library dlib which performs image computations. In addition, the connected shinyjs package offers a user-server clickable interaction useful for the landmarking.
keywords: shiny, R, machine learning, facial attractiveness, plastic surgery

Elasticsearch and R – deal with it!

Bartłomiej Staszkiewicz

As larger quantities of data are being stored and managed by enterprises of all kinds, NoSQL storage solutions are becoming more popular. Elasticsearch is a popular, high-performance NoSQL data storage option, but it is often unfamiliar to end users and difficult to navigate for day to day analytic tasks. It provides a distributed full-text search engine with a HTTP web interface and schema-free JSON documents. This presentation will briefly discuss the benefits and disadvantages of Elasticsearch on Amazon Web Services (Amazon Elasticsearch Service) and describe in detail and with examples, how efficiently transfer data between ES and R. Three packages designed for this work - elastic, elasticsearchr and uptasticsearch (R and similar on Python) will be introduced. In addition, methods to deal with nested data (from JSON) and its conversion to data frame will be presented.

BAYESIAN INFERENCE IN BIG DATA ANALYSIS

Katarzyna Sidorczuk

The ‘large p small n’ problem occurs when the number of available covariates is significantly larger than sample size. Such short fat data is a common issue in medical studies, where the availability of patients is limited. Frequentist methods fail in such cases as the correction for multiple testing results in very high p-values. In consequence, it is difficult to distinguish significant effects from noise. To address these problems, we used Stan software for performing Bayesian modeling and inference. The R interface to Stan is available as the rstan package allowing to easily fit models from R and access its outputs. The Bayesian approach provides a widespread value distribution over, e.g., the effect size that is more interpretable than p-values. We applied this method for analysis of data derived from peptide arrays - collections of short protein fragments used in research and diagnostics. They allow testing thousands of peptides simultaneously, bringing advantage to search for new biomarkers used for clinical diagnoses of different diseases, including cancers. We compared frequentist and Bayesian methods and present advantage of the latter in big data analysis.

HaDeX - analysis of HDX-MS data

Weronika Puchała

Complex and precise experiments like Mass Spectrometry generate an enormous amount of data. Such datasets require manual pre-processing, which due to their size, is tedious, time-consuming and error-prone. To automatise these steps and also provide a whole analytic workflow, we present HaDeX, an R package for analysis and visualization of Hydrogen/Deuterium Exchange Mass Spectrometry (HDX-MS) data. As we do not want to limit our tools to users familiar with programming, HaDeX is also available as a Shiny web server. It facilitates complete data analysis, including quality control and Bayesian framework for differential analysis. The sheer volume of data requires highly efficient data processing which is ensured by the data.table package. Our tool also provides a collection of data visualizations that comprehensively summarize HDX-MS results. In addition to that, our analytic methodology is discussed in-depth in the package vignette. The package is available on GitHub: https://github.com/michbur/HaDeX.

drake: reproducible workflow management in R

Dominik Rafacz

drake (Landau 2018, https://cran.r-project.org/package=drake) is an R package for reproducible data analysis workflows. Inspired by the GNU Make build system, drake controls all tasks involved in the data analysis. The abstraction system of drake incorporates all analytic steps, from data acquisition to report generation. The package orchestrates the execution of the workflow and automatically caches the results of every step. Additionally, drake monitors dependencies between these tasks pointing out when the alteration of a single task affects the others. It reduces the time needed to rerun the workflow, as drake evaluates only altered step while relying on cached results of the non-altered. Finally, drake simplifies accessing the parallelization backends and streamlines running the workflow on multicore systems or even supercomputers. drake pipelines, designed directly in R and exported as R objects further enhance the workflow reproducibility. During my presentation, I show how with the help of containers and the drake package reach the pinnacles of reproducibility in R.