3d Winter School on Data Analytics

November 28-30, 2018

Nizhny Novgorod, Russia

Laboratory of Algorithms and Technologies for Network Analysis of National Research University Higher School of Economics

Keldysh Institute of Applied Mathematics of Russian Academy of Science

JSC Intel A/O

Internet access:

Login: hseguest

Password: hsepassword

Location

HSE building, Rodionova street, 136,
Room 303 (3d floor)

Wednesday, November 28

Room 303 HSE, 136 Rodionova Str.

09:00–09:30 Registration of participants

09:30–09:40 Opening

09:40–10:30 Panos Pardalos

Lecture: Network Analysis, Data Sciences and Control in Computational NeuroScience

10:30–10:50 Coffee break

10:50–11:40 INTEL presentations

Ruslan Israfilov	*Fast Data Analytics with Python and Intel DAAL*
Dmitry Lachinov	*Brain Tumor Segmentation with Deep Learning*

11:50 – 12:40 Mario Guarracino

Lecture: Robust Data Analytics

12:40–14:00 Lunch

14:00–14:50 Mario Guarracino

Lecture: Robust generalized eigenvalue classifier with ellipsoidal uncertainty

15:00–15:50 Arseni Ashukha

Lecture: Introduction to Bayesian Machine Learning

16:00–16:50 Arseni Ashukha

Lecture: Bayesian Deep learning

Thursday, November 29

Room 303 HSE, 136 Rodionova Str.

09:40–10:30 Rostislav Yavorskiy

Lecture: How to build and analyze a simple social graph with yEd Graph Editor

10:30–10:50 Coffee break

10:50–11:40 Rostislav Yavorskiy

Lecture: Different extensions and modifications of the social graph

11:50 – 12:40 Mario Guarracino

Seminar: Supervised classification of network data

12:40–14:00 Lunch

14:00–14:50 Varvara Rasskazova

Lecture: Maximum Independent Set Problem (MISP)

15:00–15:50 Varvara Rasskazova

Lecture: Effective Methods for MISP

16:00–17:30 Student session

Demochkin Kirill: *Decision-Making in Visual Product Recommendation using Neural Aggregation Network and Context Gating.*

Teterina Daria: *Methods of Machine Learning for Censored Demand Prediction*

Guseva Maria*: Macroeconomic Modeling of the USA Economics with the help of BVAR and MSBVAR Models*

Sokolov Artem: *Voice commands recognition in intelligent systems using deep neural networks*

Glazkova Ekaterina, Soboleva Natalia**:** *Deep Learning for EEG Processing*

Miasnikof Pierre: *Three Challenges in Graph Clustering*

Friday, November 30

Room 303 HSE, 136 Rodionova Str.

09:40–10:30 Andrey Savchenko

Lecture: Facial analytics in mobile applications

10:30–10:50 Coffee break

10:50–11:40 Rostislav Yavorskiy

Lecture: Research challenges of corporate networks analysis

11:50 – 12:40 Oleg Prtokopyev

Lecture: A mixed-integer fractional optimization approach to best subset selection

12:40–14:00 Lunch

14:00–14:50 Viktor Lempitsky

Lecture: Generative convolutional networks: a short survey

15:00–15:50 Viktor Lempitsky

Lecture: Generative convolutional networks: a short survey

16:00–16:30 Closing

Arsenii Ashukha

Samsung AI Center, Moscow, University of Amsterdam.

ars.ashuha@gmail.com

Lecture 1: Introduction to Bayesian Machine Learning

In this lecture, we will consider the Bayesian machine learning approach, its basic concepts, as well as the pros and cons in comparison to the classical maximum likelihood approach or MAP inference.

Lecture 2: Bayesian Deep learning

In recent years, Deep Learning models have been successfully integrated into the Bayesian Inference framework. That allows us to start a new line of machine learning research, that is related to generative modeling and new regularization techniques. In this lecture, we will go throw basic concepts of scalable variational inference and also will discuss several examples of Bayesian deep learning models e.g. variational autoencoder and variational dropout.

Mario Guarracino
ICAR-CNR, Italy
mario.guarracino@cnr.it

Lecture 1: Robust Data Analytics,

Real measurements involve errors and uncertainties. Dealing with data imperfections and imprecisions is one of the modern data mining challenges. The term “robust” has been used by different disciplines such as statistics, computer science, and operations research to describe algorithms immune to data uncertainties. The lecture will give some introduction to this important field.

Lecture 2: Robust generalized eigenvalue classifier with ellipsoidal uncertainty

In supervised learning, uncertainty affects classification accuracy and yields low quality solutions. For this reason, it is essential to develop machine learning algorithms able to handle efficiently data with imprecision. In this paper we study this problem from a robust optimization perspective. We consider a supervised learning algorithm based on generalized eigenvalues and we provide a robust counterpart formulation and solution in case of ellipsoidal uncertainty sets. We demonstrate the performance of the proposed robust scheme on artificial and benchmark datasets from University of California Irvine (UCI) machine learning repository and we compare results against a robust implementation of Support Vector Machines.

Mario Guarracino

ICAR-CNR, Italy

mario.guarracino@cnr.it

Seminar: Supervised classification of network data

Networks represent a convenient model for many scientific and technological problems. From power grids to biological processes and functions, from financial networks to chemical compounds, the representation of case studies with graphs enables the possibility to highlight both topological and qualitative characteristics. In this work, we are interested in the supervised classification models for data in form of networks. Given two or more classes whose members are networks, we want to build a mathematical model to classify them. We focus on networks with labeled nodes and weighted edges. We define distances between networks and we build a classification model. We provide empirical results on datasets of biological interest providing details on graphical model selection

Ruslan Israfilov

Joint Stock Company Intel A/O

Presentation: Fast Data Analytics with Python and Intel DAAL

The growth of data analytical and machine learning applications made Python programming language very popular especially for data scientists. For now many mathematical packages reached the level of maturity and became de-facto standard for scientific computing in Python community. The packages provide a convenient interface, but in terms of performance they often cannot catch up the native C/C++ code. We will talk about software performance challenges arisen in data science and overview the solutions to mitigate them.

Dmitry Lachinov

Joint Stock Company Intel A/O

Presentation: Brain Tumor Segmentation with Deep Learning

MRI analysis takes central position in brain tumor diagnosis and treatment, thus its precise evaluation is crucially important. However, its 3D nature imposes several challenges, so the analysis is often performed on 2D projections that reduces the complexity, but increases bias. On the other hand, time consuming 3D evaluation, like segmentation, is able to provide precise estimation of a number of valuable spatial characteristics, giving us understanding about the course of the disease. Recent studies focusing on the segmentation task, report superior performance of Deep Learning methods compared to classical computer vision algorithms. But still, it remains a challenging problem. Here we will discuss novel deep cascaded approach for automatic brain tumor segmentation in the scope of the BraTS 2018 challenge. We will review modifications of the 3D UNet architecture and augmentation strategy than enable efficient handling of the multimodal MRI input. Besides this, we will discuss approach to enhance segmentation quality with context obtained from models of the same topology operating on downscaled data.

Viktor Lempitsky

Samsung AI Center, Skolkovo Institute of Science and Technology

Lectures 1-2: Generative convolutional networks: a short survey

Convolutional networks (ConvNets) are most commonly used for image recognition. In recent years, however, there is a growing interest in generative ConvNets, and in particular in ConvNets that are trained to generate or to transform images. As it turns out, ConvNets are as good at generating images, as they are at recognizing them. The lecture will overview the mechanisms behind such generative ConvNets as well as some important details specific to them. In particular, the lecture will cover the so-called perceptual loss functions and the adversarial learning approach, which both enhance the realism of ConvNet outputs greatly. I will also discuss some innate priors inside generative ConvNet architectures that make them particularly well suited for image generation.

Panos Pardalos

University of Florida, USA and NRU HSE Nizhny Novgorod

http://www.ise.ufl.edu/pardalos

Lecture: Network Analysis, Data Sciences and Control in Computational NeuroScience.

The human brain is probably one of the most complex objects in nature. In recent years many network models have been proposed to analyze brain dynamics and study certain neurological disorders. In nearly every study conducted on human brain networks the questions asked were what are the hubs of the network, e.g. the nodes with highest degree? There is however another important network characteristic set of nodes, arising from network controllability theory, which for the time being remained beyond the attention of researchers: identify a minimum set of driver nodes, providing controllability of the network. In this talk we are going to discuss a spectrum of problems in computational neuroscience. whose solution needs tools from data sciences and control.

Oleg Prokopyev

University of Pittsburgh, USA and NRU HSE Nizhny Novgorod

prokopyev@gmail.com

Lecture: A mixed-integer fractional optimization approach to best subset selection

We consider the best subset selection problem in linear regression, i.e., finding a parsimonious subset of the regression variables that provides the best fit to the data according to some predefined criterion. We show that, for a broad range of criteria used in the statistics literature, the best subset selection problem can be modeled as a mixed-integer fractional optimization problem. Then we show how to exploit underlying submodularity in the problem to strengthen the formulations, and propose to tackle the problem by solving a sequence of mixed-integer quadratic optimization problems. The proposed approach can be implemented with off-the-shelf solvers and, in our computations, it outperforms existing alternatives by orders of magnitude.

Varvara Rasskazova,
MAI (Moscow Aviation Institute)
varvara.rasskazova@mail.ru

Lecture 1. Maximum Independent Set Problem (MISP)

The first lecture will be devoted to the classical NP-hard problem on the maximum independent set of vertices in an undirected graph (MISP). As a review, the justification of computational complexity and practical applications of this problem will be observed. As is known, some classes of graphs are polynomial solvable with respect to the MISP. The lecture will provide an overview of these classes as well as corresponding algorithms. Finally, there are adapted exact algorithms for solving the MISP, which will also be observed in frame of lecture.

Lecture 2. Effective Methods for MISP

The second lecture will be devoted to effective methods for solving MIPS, such as approximate and (meta)heuristic algorithms. Polynomial schemes are the most popular methods among approximate algorithms. However, theoretical estimations are often far from global optima. So, fast heuristic and metaheuristic algorithms become increasingly popular nowadays. VNS seems the most successful in this direction. That’s why significant part of the lecture will be devoted to this method. As a rule, the quality of heuristic solutions is determined by strong lower bound and computation time. However, the upper bound is also of interest and can serve as an additional criterion for fast algorithms. In this regard, effective algorithms for strong upper bound are also relevant. These algorithms often use the chromatic number of the graph and will be also discussed in frame of lecture.

Andrey Savchenko
National Research University Higher School of Economics
avsavchenko@hse.ru

Lecture: Facial analytics in mobile applications

In this talk we give a brief presentation of typical facial analysis tasks. We will particularly focus on the automatic extraction of persons and their attributes (gender, year of born) from album of photos and videos. We discuss the two-stage approach, in which, firstly, the convolutional neural network simultaneously predicts age/gender from all photos and additionally extracts facial representations suitable for face identification. In the second stage, extracted faces are grouped using hierarchical agglomerative clustering techniques. The born year and gender of a person in each cluster are estimated using aggregation of predictions for individual photos. The experimental results in facial clustering and video-based age/gender recognition are presented.

Rostislav Yavorskiy

National Research University Higher School of Economics

https://www.hse.ru/en/staff/ryavorsky

Lecture 1: How to build and analyze a simple social graph with yEd Graph Editor

Lecture 2: Different extensions and modifications of the social graph

Lecture 3: Research challenges of corporate networks analysis

Student session

Demochkin Kirill, HSE NN, bachelor student (4-th year of study).

Decision-Making in Visual Product Recommendation using Neural Aggregation Network and Context Gating.

Abstract: In this paper we focus on the problem of user interests' classification in visual product recommender systems. We propose the two-stage procedure. At first, the image features are learned by fine-tuning the convolutional neural net-work, e.g., MobileNet. At the second stage, we use such learnable pooling techniques as neural aggregation network and context gating in order to compute a weighted average of image features. As a result we can capture the relationships between the products images purchased by the same user. We provide an experimental study with the Amazon product dataset. It was shown that our approach achieves a F1-measure of 0.90 for 15 recommendations, which is much higher when compared to 0.66 F1-measure classification of traditional averaging of the feature vector.

Student session

Teterina Daria, HSE Perm, master student**.**

Methods of Machine Learning for Censored Demand Prediction

In our study, we analyze a new approach for demand prediction in retail taking into account data censorship and using machine learning methods. One of the significant gaps in demand prediction by machine learning methods is the unaccounted data censorship. Econometric approaches to modeling censored demand are used to obtain consistent and unbiased estimates of parameters. These approaches can also be applied to various classes of machine learning models to reduce the prediction error of sales volume. In this study we construct two ensemble models for demand prediction with and without demand censorship. Both models are based on the predictions aggregations of machine learning methods such as Ridge regression, Lasso and Random Forest. Having estimated the predictive properties of both models, we empirically test the best predictive power of the model that takes into account the censored nature of demand.

Student session

Guseva Maria, HSE NN (PhD student). Macroeconomic Modeling of the USA Economics with the help of BVAR and MSBVAR Models

In the present research, questions related to the period, the structure of fluctuations in economic cycles and the interaction between macroeconomic variables are investigated. For the analysis, two vector autoregressive models were constructed: Bayesian vector autoregression (BVAR) and Bayesian vector autoregression with Markov switching (MSBVAR). The model parameters were estimated on the basis of an a priori independent normal - the Wishart inverse distribution (a generalization of the Minnesota a priori distribution). The specification of the BVAR and MSBVAR models included two variables: real US GDP and US employment. The period of model evaluation is from the 1st quarter of 1953 to the 3rd quarter of 2015. Based on the results of the estimation of the two-dimensional model, with Markov-Switching for GDP growth and employment, it was calculated that the average GDP growth rates were 3.35% in 1 mode and 7.35% in mode 2, where mode 1 regime of slow growth and mode 2 is a regime of rapid growth. The expected duration of the weak growth is 4.35 quarters and for the strong growth mode it is 4.24 quarters.

A graph of the probability of finding each regime for the MSBVAR model was constructed, from which it was concluded that the model is acceptable for describing the USA economy, since it describes the probability of being in this or that regime at different time periods with high accuracy. The model correctly reveals the crises of 1955, 1975 and 2008. The adequacy of the BVAR and MSBVAR models was evaluated on the basis of comparison of impulse response functions. The impact of GDP shocks and unemployment was quantified. At the last stage, the predictive properties of the above autoregressive models were revealed, an out-of-sample comparison of the two models was carried out. To obtain the forecast, a recursive regression experiment was used. Based on the results of the experiments, the RMSE values were calculated. On the basis of these values, it was concluded that it is more appropriate to use the BVAR model for short term forecasting.

Student session

Artem Sokolov, HSE NN, PhD student

Voice commands recognition in intelligent systems using deep neural networks

In this article, we focus on the isolated voice command recognition for autonomous man-machine systems or intelligent robotic systems. We propose to create a grammar model for small test command set with self-loops for each state to return blank symbols for noise and out-of-vocabulary words. In addition, we use single arc connected beginning and end of the grammar in order to filter unknown commands. As a result, the grammar is resistant to distortion and unexpected words near or inside of command. We implemented the proposed approach using Finite State Transducers from the Kaldi framework and examined it using self-recorded noised data with various level of signal-to-noise ratio. We compared recognition accuracy and average decision-making time of our approach with the state-of-the-art continuous speech recognition engines based on language models. It was experimentally shown that our approach is characterised by up to 60% higher accuracy than conventional offline speech recognition methods based on language models. The speed of utterance recognition is 3 times higher than speed of traditional continuous speech recognition algorithms.

Student session

Ekaterina Glazkova, Soboleva Natalia, HSE Moscow, bachelor student (4-th year of study)

Deep Learning for EEG Processing

Implementation of brain-computer interfaces is one on the most fascinating tasks of scientific world and at the same time one of the most challenging. The peculiarity of nonstationary electroencephalography (EEG) signals is that even short signals, for example, intentions and emotions, are represented in high dimensional space. To simplify the communication between computer and brain we need to be able to extract all meaningful information from initial datasets. Such presentation might be useful for scientific and practical purposes. For example, accurate classification of EEG signals can provide the solution for medical researches on detecting brain behavior. In this study we are looking at this task from slightly another angle – extraction of valuable information of EEG in relatively small embeddings. We design a joint of convolutional and recurrent neural networks with the usage of autoencoder to compress high dimension representation of the initial data.

Pierre Miasnikof, University of Toronto, Canada,

Three Challenges in Graph Clustering