2nd Winter School on Data Analytics (DA 2017)

November 3 – 4, 2017
Nizhny Novgorod, Russia

Russian Science Foundation (RSF), 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

Internet access:

Wireless network: **HSE**

Login: **hseguest**

Password: **hsepassword**

Friday, November 3

Room 402 HSE, 136 Rodionova Str.

09:30–**10:00** Registration of participants

10:00–**10:50** Panos Pardalos

Lecture: Data Uncertainty and Robust Machine Learning Optimization Models

10:50–**11:10** Coffee break

11:10–**12:00** Ivan Oseledets

Lecture: Tensor networks and deep neural networks

12:10–**13:00** Mikhail Burtsev

Lecture: iPavlov: Conversational Intelligence Project

13:00–**14:30** Lunch

14:30–**15:20** Marcello Pelillo

Lecture 1: Game-Theoretic Methods in Machine Learning

15:30–**16:20** Marcello Pelillo

Lecture 2: Game-Theoretic Methods in Machine Learning

16:20–**16:40** Coffee break

16:40–**17:30** Student session

Anastasya Popova. *Emotion recognition in sound*

Anastasya Sokolova. *Cluster analysis of facial video data in video surveillance systems using deep learning.*

Angelina Kharchevnikova. *The Video-Based Age and Gender Recognition with Convolution Neural Networks*

Saturday, November 4

Room 402 HSE, 136 Rodionova Str.

10:00–**10:50** Mario Guarracino

Lecture: Data Analytics using WEKA

10:50–**11:10** Coffee break

11:10–**12:00** Theodeore Trafalis

Lecture: Robust Learning with Approximate Dynamic Programming and Kernel Methods

12:10–**13:00** Oleg Prokopyev

Lecture: Exact approaches for finding maximum quasi-cliques and dense clusters/subgraphs

13:00–**14:30** Lunch

14:30–**15:20** Sergey Fedorov

Lecture: Modeling of nonstationary time series using nonparametric methods

15:30–**16:20** Sergey Fedorov

Lecture: Analyzing D2D Mobility in 5G Network

16:20–**16:40** Coffee break

16:40–**17:30** Round table: new trends in data analytics

Panos Pardalos
University of Florida, USA; NRU HSE, Russian Federation

Data Uncertainty and Robust Machine Learning Optimization Models

This talk presents robust chance-constrained support vector machines (SVM) with second-order moment information and obtains equivalent semidefinite programming (SDP) and second-order cone programming (SOCP) reformulations. Three types of estimation errors for mean and covariance matrix are considered and the corresponding formulations and techniques to handle these types of errors are presented. A method to solve robust chance-constrained SVM with large scale data is proposed based on a stochastic gradient descent method.

Ivan Oseledets
Skolkovo Institute of Science and Technology, Russian Federation

Tensor networks and deep neural networks

Approximation of multivariate functions plays a crucial role in physics, chemistry, biology, data analysis. In this talk we discuss known and new results about the approximation of multidimensional arrays (tensors) with tensor decompositions, with deep neural networks and show connections and differences between these approaches.

Mikhail Burtsev
Moscow Institute of Physics and Technology, Russian Federation

iPavlov: Conversational Intelligence Project

iPavlov is one of the leading projects of National Technological Initiative of Russia. The main goal of the project is R&D in the field of deep learning architectures for the conversational intelligence. The project is expected to deliver two technological outcomes. The first one is opensource deep learning NLP library DeepPavlov, and the second is AI platform DeepReply implementing NLP services on top of DeepPavlov library for the chat-bot and dialogue systems products. The talk will cover current deep learning methods for NLP and dialogue systems as well as roadmap of iPavlov project.

Marcello Pelillo
University of Venice, Italy

Game-Theoretic Methods in Machine Learning

The development of game theory in the early 1940's by John von Neumann was a reaction against the then dominant view that problems in economic theory can be formulated using standard methods from optimization theory. Indeed, most real world economic problems typically involve conflicting interactions among decision-making agents that cannot be adequately captured by a single (global) objective function, thereby requiring a different, more sophisticated treatment. Accordingly, the main point made by game theorists is to shift the emphasis from optimality criteria to equilibrium conditions. As it provides an abstract theoretically-founded framework to elegantly model complex scenarios, game theory has found a variety of applications not only in economics and, more generally, social sciences but also in different fields of engineering and information technologies. In particular, in the past there have been various attempts aimed at formulating problems in machine learning and related areas from a game-theoretic perspective and, with the recent development of algorithmic game theory, the interest in these communities around this topic is growing at a fast pace. The goal of these lectures is to offer an introduction to the basic concepts of game theory and to provide an overview of recent work on the use of game-theoretic models in some classical machine learning problems. Applications in areas such as computer vision and natural language processing will be discussed. I shall assume no pre-existing knowledge of game theory by the audience, thereby making the lectures self-contained and understandable by a non-expert.

Mario Guarracino
High Performance Computing and Networking Institute, National Research Council, Italy

Data Analytics using WEKA

In this lecture we present two classes of algorithms for data analysis: supervised classification and clustering. Examples of algorithms will be shown for both classes. Then, some techniques will be detailed to reduce the number of variables. Performance evaluation and statistical validation tests will be introduced and explained. Some real world problems will be addressed through examples, which use data in high dimensional spaces.

Theodore Trafalis
University of Oklahoma, USA

Robust Learning with Approximate Dynamic Programming and Kernel Methods

Decision making is ubiquitous in science and engineering. Reinforcement Learning (RL) and Approximate Dynamic programming (ADP) algorithms provide a structured approach to tackle the different decision making challenges. Due to the problem of the "curse of dimensionality", finding the exact strategy or policy might be challenging. Several methods, such as linear approximate dynamic programming (LADP), have emerged as an alternative to solving dynamic programming (DP) problems. The promising results obtained using those methods make DP more appealing to solve decision making problems for complex systems. Nevertheless, even state-of-the-art DP algorithms are not efficient in terms of computing time. Consequently, most of the algorithms require off-line calculations. In practical applications the system usually changes dynamically; new states might become part of the system and some actions that were not accounted for before might become available. Therefore, a policy that was optimal previously might not be optimal anymore or might even lead to disastrous results. I present a new class of dynamic programming algorithms that combine supervised and unsupervised learning to mitigate the "curse of dimensionality". Machine-Learning and kernel methods techniques are used to cluster similar states and reduce the dimensionality of the problem. We also consider uncertainties within the transition probability matrices and the cost function and use robust optimization formulations to mitigate the effect of uncertainty within the transition probability matrices. An example of option pricing is discussed. I also discuss how to solve RL problems in an on-line framework using the algorithms presented by considering an example of a city evacuation plan.

Oleg Prokopyev
University of Pittsburg, USA; NRU HSE, Russian Federation

Exact approaches for finding maximum quasi-cliques and dense clusters/subgraphs

The concept of a clique is used in a number of application areas due to its elegance and inherent ability to logically represent cohesive subgroups and clusters, of "tightly knit" elements (i.e., nodes). However, in many real-life applications, using cliques for discovering large cohesive clusters is impractical due to the fact that the definition of a clique is rather idealistic and, thus, can be too limiting. Consequently, a number of clique relaxation ideas and models have appeared in recent years. In particular, the concept of a quasi-clique is perhaps one of the most popular models in the literature. In this talk, we consider two versions of the quasi-clique concept, namely, an edge-based and a degree-based quasi-clique. In this talk we briefly discuss related computational complexity issues, and then focus on exact integer programming based approaches for finding maximum quasi-cliques. The developed approaches also allow handling functional generalizations of the considered clique relaxation models.

Sergey Fedorov
Keldysh Institute of Applied Mathematics, Russian Federation

Modeling of nonstationary time series using nonparametric methods.

Analysis of nonstationary random data is part of the problem of so-called Big Data. The problem of modeling nonstationary time series that arise in many areas of human activity has now great practical importance. There are a large number of examples of data series that require modeling taking into account non-stationary properties of observed values. Such are the exchange series of transaction prices for financial instruments, cardiograms and encephalograms in medicine, seismograms, temperature curves and meters of radioactivity counters, sequences of symbols in texts and genome chains. We propose the method of non-stationary time-series trajectory generation in accordance with Fokker-Plank equation for the empirical distribution function density. Parameters of trend and diffusion are estimated on the samples of time-series.

Analyzing D2D Mobility in 5G Network.

Fifth generation (5G) cellular systems are expected to rely on the set of advanced networking techniques to further enhance the spatial frequency reuse. Device-to-device (D2D) communications is one of them allowing users to establish direct connections. The use of direct communications is primarily determined by the signal-to-interference ratio (SIR). However, depending on the users movement, the SIR of an connection is expected to drastically fluctuate. This lecture shows how the technique of nonstationary random trajectories generation could help solve this problem. We develop an analytical framework allowing to predict the channel quality between two moving entities in a filed of moving interfering stations. Assuming users movement driven by Fokker-Planck equation we obtain the empirical probability density function of SIR. The proposed methodology can be used to solve problems in the area of stochastic control of D2D communications in cellular networks.

Student session

Anastasya Popova

Emotion recognition in sound

In this paper we consider the automatic emotions recognition problem, especially the case of digital audio signal processing. We consider and verify an straightforvard approach in which the classification of a sound fragment is reduced to the problem of image recognition. The waveform and spectrogram are used as a visual representation of the image. The computational experiment was done based on Radvess open dataset including 8 different emotions: "neutral", "calm", "happy," "sad," "angry," "scared", "disgust", "surprised". Our best accuracy result 71% was produced by combination “melspectrogram + convolution neural network VGG-16”.

Anastasya Sokolova

Cluster analysis of facial video data in video surveillance systems using deep learning.

In this paper we propose the approach of structuring information in video surveillance systems by grouping the videos, which contain identical faces. Firstly, the faces are detected in each frame and features of each facial region are extracted at the output of preliminarily trained deep convolution neural networks. Secondly the tracks that contain identical faces are grouped using face verification algorithms and hierarchical agglomerative clustering. In the experimental study with the YTF dataset we examined several ways to aggregate features of individual frame in order to obtain descriptor of the whole video track. It was demonstrated that the most accurate and fast algorithm is the matching of normalized average feature vectors.

Angelina Kharchevnikova

The Video-Based Age and Gender Recognition with Convolution Neural Networks

The paper reviews the problem of age and gender recognition methods for video data using modern deep convolutional neural networks. We present the comparative analysis of classifier fusion algorithms to aggregate decisions for individual frames. We implemented the video-based recognition system with several aggregation methods to improve the age and gender identification accuracy. The experimental comparison of the proposed approach with traditional simple voting using IJB-A, Indian Movies and Kinect datasets is provided. It is demonstrated that the most accurate decisions are obtained using the geometric mean and mathematical expectation of the outputs at softmax layers of the convolutional neural networks for gender recognition and age prediction, respectively.