Heavy Flavour Data Mining workshop

Thursday 18 February 2016 - Saturday 20 February 2016

University of Zurich, Irchel Campus

Book of Abstracts

Contents

1

Welcome

Data Science Applications / 2

Classifiers for centrality determination in proton-nucleus and nucleus-nucleus collisions

Author: Igor Altsybeev[1]

Co-author: Vladimir Kovalenko [1]

[1] *St. Petersburg State University (RU)*

Corresponding Author: igor.altsybeev@cern.ch

Centrality, as a geometrical property of the collision, is crucial for the physical interpretation of proton-nucleus and nucleus-nucleus experimental data. However, it cannot be directly accessed in event-by-event data analysis.
Contemporary methods of the centrality estimation in A-A and p-A collisions usually rely on a single detector (either on the signal in zero-degree calorimeters or on the multiplicity in some semi-central rapidity range). In the present work, we develop an approach for centrality determination that is based on machine-learning techniques and utilizes information from several detector subsystems simultaneously. Different event classifiers are suggested and evaluated for their selectivity power in terms of the number of nucleons-participants and the impact parameter of the collision.

The authors acknowledge Saint-Petersburg State University for a research grant 11.38.242.2015.

HEP challenges / 3

Summary of «Flavors of Physics» Challenge

Corresponding Author: andrey.ustyuzhanin@cern.ch

Data Science Applications / 4

Classifier output calibration to probability

Corresponding Author: tatiana.likhomanenko@cern.ch

Data Science Applications / 6

Data Fusion Surogate Modeling on Incomplete Factorial Design of Experiments

This work concerns a construction of surrogate models for a specific aerodynamic data base. This data base is generally available from wind tunnel testing or from CFD aerodynamic simulations and contains aerodynamic coefficients for different flight conditions and configurations (such as Mach number, angle-of-attack, vehicle configuration angle) encountered over different space vehicles mission. The main peculiarity of aerodynamic data base is a specific design of experiment which is a union of grids of low and high fidelity data with considerably different sizes. Universal algorithms can't approximate accurately such significantly non-uniform data. In this work a fast and accurate algorithm was developed which takes into account different fidelity of the data and special design of experiments

Machine Learning tools & tutorials / 7

Boosting applications for HEP

Corresponding Author: alex.rogozhnikov@cern.ch

Machine Learning tools & tutorials / 9

An introduction to machine learning with Scikit-Learn

Corresponding Author: g.louppe@cern.ch

https://github.com/glouppe/tutorial-scikit-learn

Machine Learning tools & tutorials / 10

Reproducible Experiment Platform & Everware

Corresponding Authors: alex.rogozhnikov@cern.ch, andrey.ustyuzhanin@cern.ch

Data Science Applications / 11

Mathematics of Big Data

Data Science Applications / 12

Efficient Elastic Net Regularization for Sparse Linear Models in the Multilabel Setting

HEP challenges / 13

Data Doping solution for ”Flavours of Physics” challenge

HEP challenges / 14

Transfer Learning solution for ”Flavours of Physics” challenge

Machine Learning tools & tutorials / 15

OpenML: Collaborative machine learning

Summary:

Today, the ubiquity of the internet is allowing new, more scalable forms of scientific collaboration. Networked science tools allow scientists to share and organize data on a global scale, build directly on each other’s data and techniques, reuse them in unforeseen ways, and mine all data to search for patterns. OpenML.org is a place for researchers to analyse data together, building on shared data sets, machine learning code and prior experiments. Integrated in many machine learning environments, it helps researchers win time by automating reproducible sharing, reuse and experimentation as much as possible. It also helps scientists and students across scientific fields to explore the latest and most relevant open data sets and machine learning techniques, find out which are most useful in their work, collaborate with others online, and gain more credit for their work by making it more visible and easily reusable.

HEP challenges / 16

Data Science at LHCb

Corresponding Author: tim.head@cern.ch

17

Closing Remarks

Corresponding Authors: andrey.ustyuzhanin@cern.ch, marcin.jakub.chrzaszcz@cern.ch

Data Science Applications / 18

Summary of open space discussions

Corresponding Author: andrey.ustyuzhanin@cern.ch

Machine Learning tools & tutorials / 19

TensorFlow introduction & tutorial #1

Author: Rafal Jozefowicz[1]

[1] *Google*

Introduction into deep learning, hands-on tutorial and demonstration of TensorFlow using HiggsML challenge dataset.

HEP challenges / 20

Pitfalls of evaluating a classifier's performance in high energy physics applications

Corresponding Author: g.louppe@cern.ch

Data Science Applications / 22

Optimized Methods to Apply Neural Networks in HEP

Author: Lev Dudko[1]

[1] *M.V. Lomonosov Moscow State University (RU)*

Different steps of NN application in HEP are considered. Possible optimization methods for each of the steps are discussed. The proposed methods were applied for the single top quark analysis in CMS and corresponding examples are presented in the talk.

Data Science Applications / 25

Automatic Tuning of Hyperparameters

Author: Alexander Fonarev[1]

[1] *Skoltech*

The training process of a machine learning algorithm includes tuning of hyperparameters, such as the regularization coefficient of a linear model or the depth of a decision tree. Unfortunately, it usually is conducted manually, what is very expensive to be done on a regular basis. Moreover, the growing number of hyperparameters in modern complex machine learning methods additionally complicates this problem. In our talk, we overview methods to make the process of hyperparameters tuning more autonomous, i.e. make it less requiring help of experts.

Machine Learning tools & tutorials / 26

Calibration curves as tool to test for over- and underfitting

Author: Artem Vorozhtsov[1]

[1] *Yandex*

We present a simple approach to test correctness of bias or regularization strength, or other hyperparameters.
The main idea is to fit hyperparameters so that test and train calibration curves after applying proper isotonic regression should intersect at diagonal.

Data Science Applications / 27

Fast multimodal clustering: searching for optimal patterns

Author: Dmitry Ignatov[1]

[1] *HSE*

In Machine Learning, we usually deal with object-attribute tables. However, underlying objects may have other modalities than attributes only. For instance, an object may have a certain attribute only under specific conditions. The real examples came from gene expression data, where a gene can be active (expressed) in particular situations at a certain moment of time, implying ternary relation with triples (g,s,t). One more example came from resource sharing systems like Flickr or Bibsonomy, i.e. a user u can assign a certain tag t to a resource r. One may ask how to find homogeneous patterns, groups of genes with similar properties or communities in such data.
This talk presents several definitions of "optimal patterns"in triadic data and results of experimental comparison of five triclustering algorithms on real-world and synthetic datasets. The evaluation is carried over such criteria as resource efficiency, noise tolerance and quality scores involving cardinality, density, coverage, and diversity of the patterns. An ideal triadic pattern is a totally dense maximal cuboid (formal triconcept). Relaxations of this notion under consideration are: OAC-triclusters; triclusters optimal with respect to the least-square criterion; and graph partitions obtained by using spectral clustering. We show that searching for an optimal tricluster cover is an NP-complete problem, whereas determining the number of such covers is #P-complete. Our extensive computational experiments lead us to a clear strategy for choosing a solution at a given dataset guided by the principle of Pareto-optimality according to the proposed criteria. In the end on the talk, we will outline future prospects of multimodal triclustering and its relationship with tensor factorisation.

Data Science Applications / 28

Deep Learning for event reconstruction

Corresponding Author: amir.farbin@cern.ch

Machine Learning tools & tutorials / 30

TensorFlow introduction & tutorial. Continuation

Machine Learning tools & tutorials / 31

Nvidia tutorial

Author: Alison Lowndes[1]

[1] *NVIDIA*