Report from the Heavy Flavor Data Mining Workshop

Marcin Chrząszcz
mchrzasz@cern.ch

Universität Zürich,
Institute of Nuclear Physics, Polish Academy of Science

Zurich meeting, CERN
March 4, 2016

Credits where they belong!


P. Koppenburg


T. Blake


M. Bettler


F. Dettori


A. Ustyuzhanin


T. Likhomanenko


T. Head

Workshop sponsors

⇛ Main sponsors:

University of Zurich[UZH]

intel

⇛ Additional support:

CERN

Yandex

Some details

- $50+$ participants!
- A good mixture between the ML and Physics community.
- Live discussions!

Useful information

- Indico: `https://indico.cern.ch/event/433556/`
- Full of interesting talks! Please check them out
- Let me try to give you a overview:

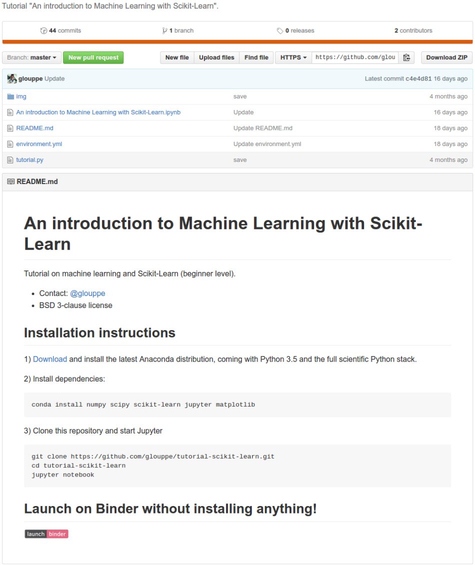
$\Rrightarrow$ Physics Prize winners:

Alexander Rakhlin, Vicens Gaitan

Tutorials

⇛ We had four ML tutorials in our workshop:

- Gilles Louppe (New York Uni), gave a super tutorial on Scikit-Learn.
- The tutorial was physics oriented $\rightarrow$ examples of training with weights.
- The material is available on the indico page with Binder.
- Highly recommend to check it out!



Tutorials

⇛ We had four ML tutorials in our workshop:

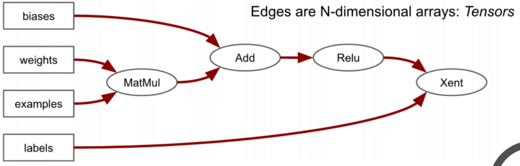
- Rafal Jozefowicz (Google Brain), gave a excellent tutorial on the Googles TensorFlow!
- Since TensorFlow is rather new and most physicists didn't have a chance to hear about it, the tutorial starts from basics!
- The material is available on the indico page.
- Please check it out, this looks like something that we as physicist could really benefit from!

TensorFlow Introduction

TensorFlow

Presentation by Rafal Jozefowicz, Google Brain

TensorFlow is an open source software library for numerical computation using data flow graphs.



Tutorials

⇛ We had four ML tutorials in our workshop:

- Alison B Lowndes(nVidia), gave a super tutorial on nVidia solutions towards neural networks trainings.
- Very impressive stuff is done by nVidia.
- So impressive that some guys from UZH want to try it out in LHCb!
- Personal view: More effort should be put inside the experiments towards using those cards.

Tutorials

⇛ We had four ML tutorials in our workshop:

- Andrey Ustyuzhanin, Aleskei Rogozhnikov (Yandex), gave a great tutorial on using rep.
- All experiments suffer from reproducibility!
- A mature solution is proposed.

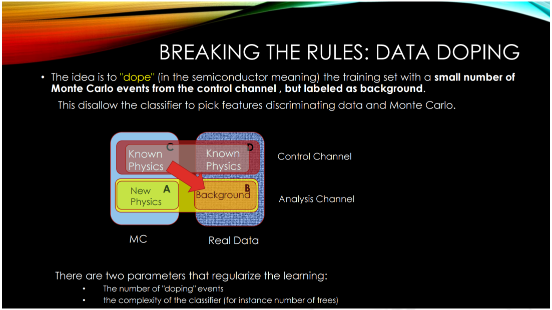
Reproducible Experiment Platform

- Python-based (numpy, pandas, ...), Jupyter-friendly
- Unified scikit-learn-like API to many ML packages (Sklearn, XGBoost, uBoost, TMVA, Theanets, ...)
- Meta-algorithms pipelines («REP lego»)
- Configurable interactive reporting & visualization to ensure model quality (e.g. check for overfitting)
- Pluggable quality metrics
- Parallelized training of classifiers & grid search (IPython parallel)
- Demo server: https://lhcb-rep.cern.ch, password: '`rep`'
- Github: https://github.com/yandex/rep

Kaggle Winning solutions

⇛ We have rewarded Kaggle Physics prize on the workshop:

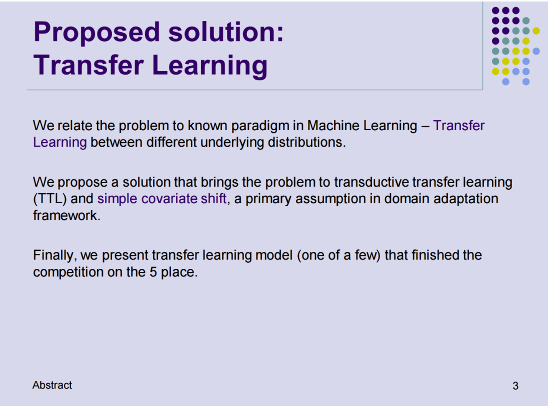
- Vicens Gaitan (Grupo AIA), presented his winning solution.
- He used so called data-doping technique to reduce data-MC agreement.
- Vincens did his PhD with LEP experiments so he understands the two worlds.
- Please check it out as it might be useful for you!
- Talk



Kaggle Winning solutions

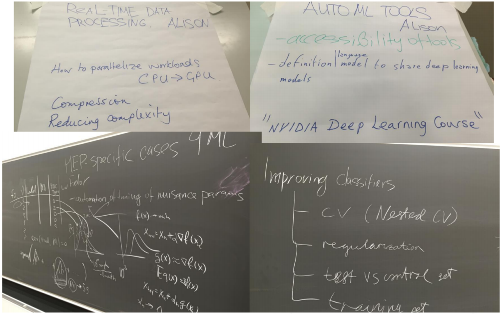
$\Rrightarrow$ We have rewarded Kaggle Physics prize on the workshop:

- Alexander Rakhlin, presented his winning solution.
- He used so called Transfer learning
- Transfer learning is method that can be used in the trainig if some of the underlying distributions are not well known.
- I think I don't need to convince anyone that might be usefull in physics ;)
- Talk



Open-Space discussion

- As an experiment we had a Open-Space discussion.
- It turns out that one can have meetings without conveners ;)
- Summary Talk



Other interesting talks

- Automatic Tuning of Hyperparameters
- Classifier output calibration to probability
- Classifiers for centrality determination in proton-nucleus and nucleus-nucleus collisions
- Data Fusion Surogate Modeling on Incomplete Factorial Design of Experiments
- Mathematics of Big Data
- OpenML: Collaborative machine learning
- Boosting applications for HEP
- Efficient Elastic Net Regularization for Sparse Linear Models in the Multilabel Setting
- Deep Learning for event reconstruction

Summary

⇛ I hope I interest you enough that you check out the workshop!

⇛ The workshop was a success and future events like this should happen!

Backup