Yandex Data Science tools for Science

Andrey Ustyuzhanin[1,2] on behalf of Yandex team

[1]CERN LHCb, [2]Yandex School of Data Analysis

Services landscape

- Web search
- Image search
- Speech recognition
- Car traffic prediction
- Mail and spam filtering
- Natural language translation
- Market (shopwindow for internet shops)
- Yandex Data Factory (`yandexdatafactory.com`)
- Yandex School of Data Analysis (member of LHCb since Dec'15)

Detector Operation & Data Quality

- WebPresenter aka Monet, LHCb data quality monitoring
  - up and running in 2016 data taking
- LHCb anomaly detection & prediction
  - under development within CERN openlab framework

Triggers Optimization

Triggers are event selection procedures that should filter out uninteresting events

- Topological trigger with MatrixNet formula optimized for speed
  - improvement up to 60% percent signal efficiency increase in RunII compared to RunI
  - used in 60% of LHCb papers

Infrastructure Optimization

Events are stored in the LHCb grid for a longer term

- Data Storage optimization
  - up to 40% disk storage save, under development within CERN openlab framework
- Event metadata indexing
  - run-event number access
  - used for optimization of streams

Data Analysis Tools

- Reproducible Experiment Platform (`https://github.com/yandex/rep`)
- hep_ml (`https://github.com/arogozhnikov/hep_ml`)
  - reweighting
  - uniform boosting
- Matrixnet-as-a-Service
- everware - service for managing Jupyter-based research environment using Github and Docker (`http://everware.xyz`)

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
- Paralleled training of classifiers & grid search (IPython parallel)
- Open-sourced, Apache 2.0:
  `https://github.com/yandex/rep`
- Supported by Yandex

REP: Meta Machine Learning (REL-Lego)

- Factory
- Grid Search
  - GridOptimalSearch
  - Folding Scorer
  - Various Optimization algorithms
- Interface of parameter optimizer
- Folding `https://github.com/yandex/rep/blob/master/howto/04-howto-folding.ipynb`
- Stacking

REP: Reporting

- Draws set of reports upon model training completion. Supported libraries:
  - Matplotlib
  - ROOT
  - Bokeh (Javascript)
  - plot.ly (going to be deprecated due to limitations)
- Extensible!

`https://github.com/yandex/rep/blob/master/howto/02-howto-Factory.ipynb`

HEP ML package

ML-inspired tools for HEP

- UGBoost `http://bit.ly/uBoost`
- GBReweighting `http://bit.ly/GBReweight`

Everware. Sharing Research. Reproducible

- Jupyter-based
- Docker-empowered
- github-backed

`http://everware.xyz`

Collaboration workflow

versioning and continuous ...

- testing
- integration
- publishing

Outreach

- Flavours of physics Kaggle challenge
  `https://kaggle.com/c/flavours-of-physics`
- Machine Learning for HEP summer schools
  `http://bit.ly/mlhep2016`, `http://hse.ru/mlhep2015`
- Conference on Machine Learning
  - `https://yandexdataschool.com/conference`
- Workshop on ML applications in HEP at NIPS'15
  - `http://yandexdataschool.github.io/aleph2015/`
- Workshop on Machine Learning in Zurich
  - `http://indico.cern.ch/event/433556/`

Tipping Water

to be published at ACAT'16

- LHCb flavour tagging

Other projects include (work in progress)

- Particle and jet identification
- "Alternative" tracking, long-lived particles
- CRAYFIS - smartphone-based cosmic rays observatory
- COMET.tracking improved ROC AUC from 88.3% to 99.9%
  `https://inclass.kaggle.com/c/comet-track-recognition-mlhep-2015`

See you at CHEP'16!

We Love Data

> Crunching data for food... :)

...sharing is caring:

- REP
- hep_ml
- everware

Thank You!