About: ML@Atlas/CMS, present and future

Cécile Germain

TAO (Apprentissage & Optimisation)

Université Paris Sud – INRIA - CNRS

Challenges: what we learned with HiggsML

What tool for what problem and *Learn how-where to apply what MML for HEP*

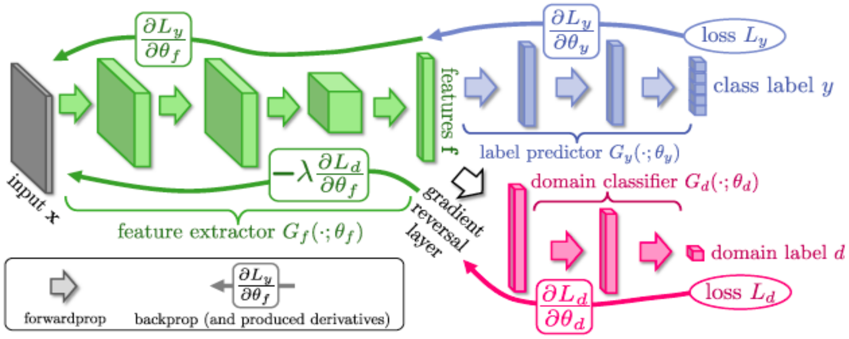
- Crowdsourcing
  - Finding a tool might be very difficult (1800+ submissions NIPS2015, 400+ accepted, 20+ workshops)
  - Formulate your task as a machine learning problem, with adequate data. Easier said than done, highly non trivial, requires a small, balanced and dedicated team.
  - Then a lot of people are eager to solve the problem
- Positive results
  - Gradient boosting considered of interest as a method and a tool
  - Also raised awareness about validation eg k-fold
- To improve
  - Core ML-research involvement eg would be most beneficial for all image-related tasks
  - Challenges with a classical (proxy) objective function will be easier to convert to a benchmark.

The tools approach

- Can we do better with raw data ?
    - The promise of deep learning: learning internal representation in place of feature engineering
    - From the limited experience of HiggsML, not exactly the case
    - Best observables not necessarily minimally correlated
- Anomaly detection -> novelty detection = learning from positive and unlabeled examples. Unlabeled data definitely help
- Cluster splitting Naive question: what about classical clustering (with very large datasets)?
- Deep learning for event classification: what advantage do you expect (over other classification methods)?
- Parallel tracking: muti-objective optimization – one figure of merit, but you might want to ask for more information
- Systematics. Ongoing work at Orsay + IC on a principled ML formalization, with practical methods to follow.

The tools approach

Systematics. Ongoing work at Orsay + IC on a principled ML formalization, with practical methods and benchmark dataset (worked out from the HiggsML one) to follow soon: domain adaptation


$\frac{\partial L_y}{\partial \theta_f}$
$\frac{\partial L_y}{\partial \theta_y}$
loss L_y
class label y
features $\mathbf{f}$
label predictor $G_y(\cdot;\theta_y)$
$-\lambda \frac{\partial L_d}{\partial \theta_f}$
domain classifier $G_d(\cdot;\theta_d)$
input $\mathbf{x}$
gradient reversal layer
feature extractor $G_f(\cdot;\theta_f)$
domain label d
loss L_d
$\frac{\partial L_d}{\partial \theta_d}$
forwardprop
$\frac{\partial L_y}{\partial \theta_f}$
backprop (and produced derivatives)

From Ganin et al., NIPS 2015

Data at training and test time come from similar but different distributions

For effective domain transfer to be achieved, predictions must be made based on a data representation that cannot discriminate between the training (source) and test (target) domains. Data representation = good motivation for DNN

Vision

ML: input all information and let machine learn

- Basically true for supervised and non supervised learning– but active learning critical for many real-world tasks, eg anomaly detection. Experience in HEP?
- Reinforcement learning is about providing external feedback

And beyond Can we combine ML and physics input in a smart way?

Integration of a priori knowledge

- Most useful on an ad-hoc basis: identify/describe what is already encoded in the simulation data, how it is encoded, and the residual. Would greatly help for contributing to systematics analysis
- Bayesian approaches – ad hoc, of course we need it
- In the long run, domain scientist in the loop: within reinforcement learning, preference learning (related: apprenticeship learning)

Vision

We are physicits, not data scientists. We want to focus on physics. We want to make "optimal" use of our data.

- Just like Facebook, Google, all the finance industry, health industry, et al. Data Scientist shortage is so terrible that IBM sells MLaaS – machine Learning as a Service, on the Cloud…

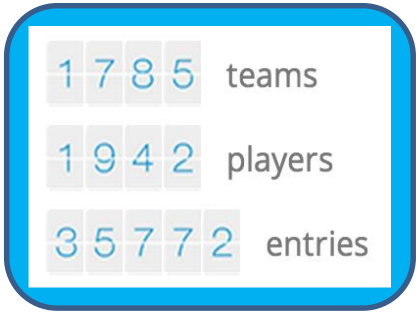
In the next years, how do you entice the best and the brightest young ML?

Possibly, partially, tentatively: by giving them an opportunity to demonstrate they actually are.

- LHC name is great, but ordinary results with impact on the real world are not enough per se
- Fortunately, HEP real problems are strongly related to fundamental and active ML questions, eg (a few)
  - Systematics: infomation geometry: define/estimate/use « good » distance between distributions
  - Model selection and parameter tuning, quality (questions about upper bounds on discrimination): beyond asymptotic analysis, data-dependent complexity estimates
  - Your exotic learning objective functions: NP-learning, ranking
  - Not knowing enough on the other themes, earger to learn about them.

Conclusion

Vision + tools = benchmarks



1785 teams

1942 players

35772 entries

$13,000