General Math, TMVA, RooFit

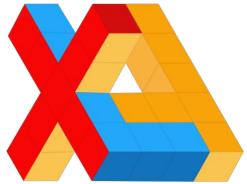
S. Wunsch, S. An (CERN, EP-SFT) for the ROOT team

TMVA

Modernisation of TMVA: Goals and Pitfalls

- Since our last Plan of Work, we have continued to work on modernisation of TMVA to adapt to the rapidly evolving landscape of ML/DL.

- Main trend 2019-2020 in the outside world:
  a. Popularisation of **new DL architecture**, such as Graph Neural Network and Transformers
  b. Continued **development and improvement of DL Frameworks** (TF 2.0 and PyTorch)
  c. A shift of focus to **provide better support for deployment in production**



Modernisation of TMVA: Goals and Pitfalls

- We have refined our strategy in modernisation of TMVA:
  - **1 Going out: ROOT -> External**
    - Provide a batch generator of ROOT data for use in external ML/DL frameworks
      - Allows for batching and shuffling of data on the fly
      - Allows easy training on very large datasets
      - **Manpower constraint:** This is **already proposed last year** and was noted to “expect significant effort”
      - **Technical challenge:** Batch mode of RDF? Complications with Xrootd?
        - Strong interaction with RDF

Example ML workflow loading batches

```
df = ROOT.RDataFrame("Events", "http://file.root")
generator = TMVA.BatchGenerator(df, cols, batchSize)
for step in gradientSteps:
    x = generator()
    model.fit(x)
```

Modernisation of TMVA: Goals and Pitfalls

- We have refined our strategy in modernisation of TMVA:
  - **2 Going in: External -> ROOT**
    - Provide a **Fast Inference System** for easy application of externally trained ML (xgBoost etc.) and DL (TF, pyTorch) models to ROOT data
    - Already discussed in CHEP2019
    - Already in development
      - BDT fast inference core: done
      - DL fast inference core (code generation):
        - ONNX based so generic models supported
        - change in specification due to concurrent external efforts
        - major structure of system close to completion
          - technical challenge: clean type erasure, static typing…
        - designed with long-term maintainability in mind
        - expect a demonstrator (DNN & CNN) soon
    - Nice to have
      - A new, modern (Pythonic?), uniform interface for ML/DL training and inference

Modernisation of TMVA: Goals and Pitfalls

- We have refined our strategy in modernisation of TMVA:
  - **3 Staying put: Improve the experience of users of legacy functionalities**
    - Best example - BDT
    - What we could do
      1. Provide a modern inference for both training and inference, preserving core algorithms and old factory interface for backwards compatibility
      2. Modernisation of TMVAGui (visualisation of training results)
         - Interplay with new ROOT7 GUI development?

```
x = ROOT.RDataFrame("Events", "http://filex.root")
y = ROOT.RDataFrame("Events", "http://filey.root")
model = TMVA.RBDT(nTrees = 800, maxDepth = 3, nCuts = 20, separationType = "GiniIndex")
model.fit(x)
model.save("BDT_v2", "bdt_v2_model.root")

model2 = TMVA.RModel("GNN_v3", "gnn_v3.onnx")
input = ROOT.RDataFrame("Events", "http://input.root")
pred = model2.predict(input) #JIT from generated code
```

Modernisation of TMVA: Goals and Pitfalls

- Other points you have raised…
  a. TMVA could benefit from PyRDF, specifically with regards to potential support for Spark Data Pipeline
  b. (from last year) decorrelation/anti-sculpting algorithm support

Received comments

- **Better interoperability with external ML ecosystem** and how TMVA could be benefit from new ROOT developments such as PyRDF (investigate how it could work alongside a Spark Data Pipeline?)
- Promote what was already done from point of view of **modern high-level interfaces** (e.g. RTensor+ RDF)
- **Investigate clad integration** (it could be a good Technical student or IRIS-HEP Fellow project?)
- TMVA should focus strongly on the **fast ML inference** part. The interfaces and example implementations for legacy TMVA and general BDTs are in experimental but have to be brought into production with long-term support.
- The support of generic NN models is crucial to get people moving to new ROOT ML i**nterfaces for fast inference**.
- **IO for ML** is another big topic and covered below in the IO section, but could also be treated as a part of TMVA.

Received comments

- TMVA: if I understand correctly we acknowledge that most **ML training will happen outside of TMVA/ROOT** (e.g. in Tensorflow/PyTorch), but we still expect users to want to **use those trained models** to perform inference at analysis time (possibly using ROOT). As far as I know, we don't have **performant interfaces** for this scenario, nor high-level, nice Pythonic ones, so that would be where to invest IMHO. Plus top-notch docs.
- It's probably just that it's not my area of expertise, but I also have the feeling that (especially for TMVA) **we are out of touch with the userbase**. Do we know what people want?
- While I'm still working on modernisation, there are still a small but non-significant group of **users using TMVA for its traditional functionalities**, i.e. BDTs. We should also try to **improve their experience**, for e.g. by modernising the interfaces to these old functionalities, and by bringing TMVAGui up to a similar standard of the new ROOT web GUI.

RooFit/Math

Main focus

- **RooFit/modeling/fitting is essential part of ROOT.** Need to continue investing in it by providing support, modernize it and provide new features as requested by users.

- Need to continue **improving CPU efficiency and memory usage**

Detailed work items

New improvements to be done next year (in detail):

- Math: Make Minuit2 default minimizer, deprecate TMinuit
- RooFit: New C++ HistFactory implementation
- RooFit: CUDA support
- RooFit: Prototype usage of AD
- RooFit/RooStats/TMVA: Better python API

Additional comments

- Roofit/Math: **Investigate clad integration** (it could be a good Technical student or IRIS-HEP Fellow project?)
- RooFit: we need **nice Pythonic interfaces**
- It's probably just that it's not my area of expertise, but I also have the feeling that (especially for TMVA) **we are out of touch with the userbase**. Do we know what people want?