Our overview talk Machine Learning with ROOT/TMVA
Monday 4 Nov 11:15 at Hall G (Track 6) by Stefan Wunsch

TMVA in modern ML ecosystem
- TMVA: Toolkit for Multivariate Analysis, first released in 2005
- Machine Learning toolkit in ROOT
- ML Landscape is fast evolving, TMVA needs to adapt
- Focus on a fast and robust ML inference system

1. Fast Inference Engine for Decision Trees

   Initiated as CERN Summer Student Project by Luca Zampieri
   - Decision-tree ML algorithms widely popular in HEP and in data science
   - Special need for application in HEP
     - Low-latency inference critical for some use cases i.e. HLT
     - Focus on event-loop inference rather than batch inference

   Just-in-time compilation
   - With Cling, the interactive compiler in ROOT
   - Compiles hard-coded evaluation logic parsed from the model
   - This allows us to exploit compiler optimization dynamically

   “Branchless” representation of trees
   - Unroll the tree from linked nodes into a long sequential array
   - Fill in missing values in sparse trees to create full binary trees in the array representation
   - Tree traversal is now a maths operation - cheaper than if branch

   Branchless implementation assumes shallow, nearly-full trees
   - Most decision-tree ML algorithms produce these
   - Branchless implementation to be integrated for deep trees

   Tree ordering
   - Trees evaluated in order of feature & cut value of root node
   - Improve dynamic branch prediction, reduces branch misses

   Loop nest optimization
   - Chunk iteration space (over trees & events) into small blocks
   - Improve data/instruction locality, reduces cache misses

   Loop nest optimization improves non-JITted implementations only

2. Inference of ONNX Deep Learning Models

   ONNX [3] is an open format for DL models
   - Supports most popular DL operators/layers
   - Converters available from major DL framework to ONNX
   - ONNX runtime: an open source inference engine
     - Supports by industry in fast development
     - Highly optimized for low-latency inference
     - Multiple backends and optimization methods supported

   Development in TMVA
   - An inference interface for ONNX models, designed for HEP applications
   - ONNX operator-based infrastructure
     - ONNX model exploration and manipulation
     - Allow potential customized optimization
   - Interface with open-source ONNX runtime
     - Convenient interoperability with ROOT data
     - Support for implicit multi-threading inference
   - Code generation from ONNX model

   Work in Progress

Acknowledgement and References
4. We gratefully acknowledge the support of the Marie Skłodowska-Curie Innovative Training Network fellowship of the European Commission Horizon 2020 Programme, under contract number 765770-INSIGHTS. https://root.cern