Additional Functionality in TMVA (Software R&D)
Regression and Boosted Decision Trees

Student: Andrew Carnes, University of Florida
Supervisor: Sergei Gleyzer, University of Florida

Abstract:

In high energy physics, machine-learning based regression algorithms have been used extensively to improve detector resolution for measurements of photons, electrons and b-jets. Due to their robustness and simplicity, boosted decision trees (BDTs) have been used more frequently in the regression context. One current limitation of the commonly-used Toolkit for Multivariate Analysis (TMVA) package is the use of a single hard-coded loss function for the regression algorithm. This project plans to implement an array of the most popular loss functions for use with the BDTs and other regression algorithms through creation and inheritance of a general class. The user will be able to choose the loss function that fits their regression goals the best or implement their own. The project further proposes to speed up the BDT implementation with multithreading/multiprocessing techniques. New functionality will be documented in Doxygen and the project will deliver an executable Jupyter notebook demonstrating new functionality for at least two different benchmark examples from different LHC experiments. The new features will be subsequently advertised in the LHC machine-learning community through all available fora.

Software R&D Proposal for TMVA Regression

Student: Andrew Carnes, University of Florida
Supervisor: Sergei Gleyzer, University of Florida
Period: Summer/Fall 2016

Introduction

The Toolkit for Multivariate Analysis (TMVA) is a commonly used ROOT-integrated software package in High Energy Physics (HEP). Recently, TMVA has undergone an upgrade in both functionality and performance[1]. The changes center around greater modularity and flexibility in design, improvements in memory handling and the performance of individual methods, parallelization, additional features and interfaces. The following proposal focuses on the improvement of TMVA functionality in the context of regression, in particular for boosted decision trees (BDTs).

Machine learning regression techniques have been used in HEP to improve detector resolution and energy measurements of photons, electrons, b-jets [2, 3]. This technique is currently used in the CMS trigger system to estimate muon momenta [4].

TMVA has several methods available for regression, such as BDTs, neural networks and others. Due to their robustness and simplicity, BDTs have been used more frequently in the regression context. One current limitation is the use of a single hard-coded loss function for regression.

Andrew has developed a standalone BDT package for regression capable of implementing a variety of loss functions, called BDTLib [5]. CMS trigger system currently uses BDTLib to predict the muon momenta, showing a performance improvement over previous non-machine-learning-based methods. These improvements are in large part due to the choice of the loss function. At the Inter-experimental Machine-Learning Working Group meeting [6], there was a broad consensus in the HEP Machine-Learning community of a need to incorporate BDTLib's flexible loss function capability into TMVA.

Deliverables

The proposed project will deliver a full integration of flexible loss-functions in TMVA, implemented as a general class, independent of the regression algorithm is used. The user will be able to choose from an array of useful loss functions or have a possibility to implement their own loss function to be used during regression. The project further proposes to speed up the BDT implementation with additional parallelization. Integrating loss function versatility into TMVA will provide continued support for this functionality in the HEP machine-learning eco-system. New functionality will be documented with Doxygen. Andrew will also deliver an executable Jupyter notebook example demonstrating the new functionality for at least two different benchmark examples, and a dedicated test for the TMVA test suite.

Timeline

Part I

Week 1: Analysis of TMVA regression code, initial benchmarks between TMVA/BDTlib
Week 2-3: First implementation of the TMVA::LossFunction class

Week 4-5: Tests of the code, detailed performance benchmarks, addition of regression tests to the TMVA test suite
Midterm report (presentation of results for part I)

Part II

Week 6-8: Additional parallelization of BDTs in the context of regression
Week 9-10: Benchmarks of the parallelization
Week 11-12: Updates to the Documentation, creation of additional executable Jupyter Notebook Examples, presentation of new functionality to the community
Final Report including results of parts I & II

Dissemination of Results

During the project, Andrew will present regular updates in TMVA Developer meetings and to the broader Machine-Learning in HEP community during both midterm evaluations and after project completion. Andrew will additionally present new functionality in all relevant HEP-Machine-learning forums including IML[7] and Diana-HEP[8]. The results will be included in the new TMVA release later this year and presented at CHEP 2016, as part of new TMVA functionality developed this year.

[1] S. V. Gleyzer, L. Moneta, O. A. Zapata Mesa, Development of Machine-Learning Tools in ROOT, Proceedings of the XVIIth International Workshop on Advanced Computing and Analysis Techniques in Physics Research, ACAT 2016, (submitted)
[2] CMS collaboration, S. Chatrchyan et al., Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. Phys. Lett. B716 (2012) p 30-61
[3] CMS collaboration, S. Chatrchyan et al., Search for the Standard Model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks, Phys. Rev. D89, 012003 (2014)
[4] CMS collaboration, Tapper A. et al., CERN-LHCC-2013-011, CMS-TDR-12, Technical Design Report for Level-1 Trigger Upgrade (2013) p.15
[5] https://github.com/acarnes/bdt
[6] https://indico.cern.ch/event/495175/
[7] http://iml.cern.ch
[8] http://diana-hep.org