CS281 Project Proposal

use Machine to learn variables for signal selection in High Energy Physics

Baojia Tong, Jonah Philion

This project will eventually target as an application paper. In high-energy physics, machine learning tools have been successfully applied to a wealth of problems. The large dataset (25 GB/s) with highly correlated observables provides an ideal environment for employing machine learning tools.

Problem Statement The search project is to try to measure the triboson production for the first time at the LHC, specifically, to $W^{\pm}W^{\pm}W^{\mp}$ and then decay to three charged leptons e, μ and neutrinos. This measurement probes the quartic coupling that is predicted by the SM yet quiet sensitive to new physics models, which could completely change our understanding of fundamental nature. With higher center of mass energy and more data, this process could be observed in two years.

The challenge is the small signal yield and the enormous and complicated background. Different other electroweak process ($W^{\pm}W^{\mp}$, ZZ, $W^{\pm}Z$) has about 1000 times higher rate and can also generate signatures that are similar to the signal. Other SM process like $t\bar{t}$ has about 100000 times higher rate, and can generate decay products that fake the signal signatures in the detector. All these add difficulties in the discovery process. In the 2015 analysis, a cut-based selection is used, and sensible results are achieved.

Therefore this is an ideal challenge for testing machine learning methods. The problem essentially is an **identification problem**, with fixed input variables. Some of the most useful variables are the three lepton's momentum and identification information from the physics detector. Additional hadronic activities will be measured as jets. The weakly-interacting neutrinos cannot be measured directly, and hence only the sum of their transverse momentum can be inferred from the conservation of momentum. These will form the input of the identification problem.

Evaluation The evaluation metric will be the signal significance. The previous analysis' selection will be the honest baselines, which can be reproduced and then compared with.

Approach Both signals and backgrounds will be simulated, normalized to the expected yield of current data size. All inputs will be based on simulation and converted to numpy arrays. Standard cleaning and normalization will be applied after checking. Linear regression will be tested, but more interesting options will be decision trees and neutral networks. The ROC curve will be useful in choose a good selection point, where enough distinguish power is achieved while maintaining reasonable measurement Poisson uncertainties.

Milestone 1.0 The proof of concept test could be done using signal and only one background (WZ). It is expected that machine learning methods will do much better than previous selections. This requires baseline framework building, mostly in python, and utilizing machine learning packages, mostly from **(tensorflow/pytorch ??)**. This is estimated to be finished before November.

Milestone 2.1 From there, more backgrounds could be added. There will be around ten different kinds of backgrounds, and it is challenging to have identifiers that will simultaneously reject all backgrounds while keeping the signal acceptance. Lots of optimization will be necessary. This is estimated to be finished by December.

Milestone 2.2 Parallel to Milestone 2.1, the optimizer's output needs to be interpretable. In high energy physics, and in science in general, the "black-box"type of machine learnt output is disfavored, not only because it is hard to interpret, but also because it is hard to exam and validate the method. Hence, a challenging part of this project will be map the neural network/decision tree outputs to variables that can not only be computed analytically, but also be easily validated in the future with data. We can try to understand the selections by reducing the number of inputs or employ a data planing method. This is also estimated to be finished by December.

Collaboration Plan Tony will be in charge of sample production. Jonah will work hard and do all the work. :)

Double-dipping This result, if successful, will be documented as an ATLAS internal note, and will be used by the official analysis.