TAFC — Advanced Topics in Computational Physics

October 24, 2017 — AD and JS

Submission deadline: Sunday, November 5, 2017 at 23:59 UTC

Particle Physics module problem set

For this module there will be a single problem.

Goals

- To understand aspects of $H \to \tau\tau$ signal events and how they differ from background events.
- To get acquainted with machine learning techniques.
- To go through the `train-test-validate` loop in a realistic environment.
- To participate in a real data challenge.

Higgs Boson Machine Learning Challenge

This data challenge was one of the first initiatives to open the LHC data to machine-learning professionals and specialists from other areas. The challenge was delivered using the `Kaggle` platform.

The goal of the competition was to find the best possible function (a classifier) that distinguishes between two classes of events:

- $H \to \tau\tau$ signal, and
- Background processes, where something else (a jet, a few pions, etc) is misidentified as a τ lepton.

The better one can distinguish between these two, the easier it will be to discover a Higgs boson.

Make sure to peruse the documentation, the HEP for ML people, and the ML for HEP people. There is also an interesting series of slides and videos from after the competition. A similar event was held at NIPS14: slides and videos. The web also has interesting reports and recent explorations.

The workflow is rather simple and there are several starter kits. More advanced tools, hints, tips, and discussion can be found in the `Kaggle` discussion forums, and follow-up workshops. Basically

- You are given a `training` sample of many events. Each event has several quantities (features, $\vec{f}$) and the true label (signal or background, l).
- You then use that sample to construct a function that predicts the label of each event based on the values of the features. This is the "learning" process and the outcome is a classifier, a function of the features, $C(\vec{f})$. This learning is done against the approximate median significance (AMS); the higher this value, the more significant a Higgs signal would be and the better your classifier is in separating "wheat from chaff".
- You are then given a `test` sample of events of which you only know the features, not the labels. For this `test` sample you calculate (predict) an estimate of the label, $\hat{l} = C(\vec{f})$.
- You submit your prediction for the `test` sample to `Kaggle` and it is scored.

We encourage you to **try multiple classifiers**, from the most naive to the most advanced. You can submit up to 3 for evaluation; our advice is that you try, in this order, a cut-based analysis on important variables, a BDT like `XGBoost`, a (deep) neural network from `scikit-learn` or `Keras`. We also encourage you to **try different parameters** in each algorithm (different cut values, BDT boosting type, NN number of neurons and layers, etc).

You will have to register an account with `Kaggle` and make a team named `TAFC17 xx`, where `xx` is your group number. The data files are available from the `Kaggle` data section.

Yes, we know that everything is now released in the CERN Open Data Portal. If you cheat, we'll find out. Speaking of cheating, there are some interesting musings by one of the original challenge participants in his blog.

Submission

You must submit:

- An exposé with a maximum of **5 pages of text** that:
  - Discusses the key physics aspects of the $H \to \tau\tau$ decays, including the most relevant differences between the signal and the background.
  - Discusses the features (variables) provided: their distributions, joint distributions, etc.
  - Discusses the way in which the competition gauges the best submissions, both in terms of the metric used as well as in terms of the need for two leaderboards (private and public).
  - Reports on the classifiers used, the variety, the underlying algorithms/libraries, and their implementation.
  - Reports on variations done to the algorithms' parameters in the search for a better performance.

  - Includes a link to your `Kaggle` user profiles, i.e. something like this.

- The code you used for the `Kaggle` submission(s), including:
  - the submission file(s), and
  - the means to execute all steps, from training, to evaluation of the submission file(s) created by your code.

The submission should be made available via **a link to a single file** (from `DropBox`, `Google Drive`, `OwnCloud`, `OneDrive`, etc) that is to be sent via email (this link includes the email envelope).

The submission link must be valid until November 6, 23:59.

Evaluation

The grade will be in the 0–20 points range, with the following components being separately evaluated:

- Discussion of $H \to \tau\tau$ decays (2 points)
- Discussion of the variables (features) available (2 points)
- Discussion of the metric used to assess what is a good result (AMS) (2 points)
- Report on the algorithm(s) tried and their implementation (9 points)
- Report on the algorithm variations tried (2 points)
- Originality of the implementation(s) (2 points)
- Best score obtained (1 point)

Malus Submissions received after the deadline will have a maximum grade that is reduced by 1 point for each day of delay. I.e., submissions received on November 6 (7) will have a maximum grade of 19 (18).