Karlsruher Institut für Technologie

Moderne Methoden der Datenanalyse Computerpraktikum
Bearbeitungszeitraum: 23.06 – 13.08.2023

Project: Multivariate Analysis - Higgs Challenge

This sheet is not a regular exercise, but rather a little "Project" to enhance your skills acquired during the exercises. The project is dedicated specifically to the use of multivariate analysis techniques.

A (rather complicated) data set is provided — it is a subset of data released by the ATLAS experiment at the Large Hadron Collider LHC at CERN under the title "Learning to discover: the Higgs boson machine learning challenge". The related document is provided on our web page; the full information is given at http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014, a web page hosted by CERN, which makes LHC data available to the public for educational projects and research.

The MVA project will be centred around methods to search for the very rare signal of Higgs-production in the data of an LHC experiment. The released "data" in fact consist of simulated data of a complex signal, decays of a Higgs boson to two τ leptons, and the sum of all background reactions. Among the many combinations of possible decay modes of a pair of τ leptons, the "Higgs challenge" concentrates on the dominant one:

- one τ decays to hadrons and a neutrino (τ_{h}),
- and the other one to an electron or muon and two neutrinos (τ_{ℓ}).

The short document atlas-higgs-challenge-2014.pdf describes the problem and gives background-information on the input data. Roughly speaking, these fall into two categories:

1. primary or PRImitive variables (indicated by the prefix PRI) are raw quantities of the identified objects in the event, and
2. DERived variables (prefix DER) are quantities derived from these by the ATLAS physicists to be used in their own multivariate analysis, the final results of which were published in spring 2015.

The challenge is, of course: try to beat them!

For the purpose of this data-analysis project, no deep understanding of the underlying physics is required, simply take the data as an example of a complex "multivariate data-set".

The original data set consists of 250'000 simulated events as the "training sample" for multivariate algorithms classified as signal (i.e. a true $\mathrm{H}\to\tau\tau\to\tau_{\mathrm{h}}\tau_{\ell}$ event) and background reactions. Weights are provided with each event such that the sum of weights corresponds to the same total number of events observed by the ATLAS experiment in the year 2012. An independent "validation sample"with 450'000 events to measure the performance of any proposed algorithm is also provided. To avoid "cheating", originally there was no classification of these events. This was added later and now allows everybody to compare the performance of any newly developed multivariate classifier.

To set the scale: a record number of 1'785 teams participated in the Higgs challenge conducted by Kaggle. The winner achieved a score of 3.81; there were only tiny differences in the top league, rank four still scored 3.72. The real competition is over now, but we will use the data for the final data classification project or the course "Moderne Methoden der Datenanalye", and it is your chance to combine what you've learned in the course with your own creativity!

Project requirements
This project is meant to be "open" to stimulate your own creativity to exercise your skills on a real-life example. For recognition, a document with about two pages of text and an appendix with graphics documenting your findings is expected. Please work in teams of maximal four people to discuss ideas, solve technical problems, or share some of the needed development work. A single document with four names on it will be accepted and accounted for all team members.

- **Step 1.1:**
  Familiarize yourself and work through the provided ipython-Notebook HiggsChallenge.ipynb. It contains code to read ROOT trees, reconstruct simple variables, make some quick plots and to turn the ROOT trees into numpy arrays which can be used with machine-learning frameworks in python. It also implements some example models and demonstrates how to train and evaluate them.

  A word on the "score", the measure of performance, is advisable here: for the optimisation of a search, we use the measure of significance given by the number of signal events, s, divided by the expected background uncertainty, $\sqrt{b}$. As this is not very stable in the limit $s, b \to 0$, a regularisation b_0 was introduced, so that the significance becomes $S = s/\sqrt{(b+b_0)}$. For the Higgs Challenge, $b_0 = 10$ has been imposed, thus providing a common measure of performance for all participants. For studies with the reduced data set, you may temporarily set this to $b_0 = 3$.

- **Step 1.2:** **obligatory**
  Now it is time to work on improvements. Inspect the variables and their correlations. Which of them show the most prominent differences between signal and background? Which of them are not so relevant? Summarise your ideas and collect some of the graphs for your documentation.

  You should implement code to produce some figures of the distribution of one or more highly-discriminating variables (e.g. the reconstructed Higgs boson mass `DER_mass_MMC`), after applying a cut on the classifier corresponding to the best performance. A Higgs-Signal may then become visible in a single distribution.

- **Step 1.4:** Be Creative! **obligatory**
  This is the most challenging, but also most rewarding part of the project. Develop, discuss, test and improve ideas to increase the performance of your classification. Consults the documentations of the different ML-libraries (`scikit-learn` or `keras`) to get inspiration on options to try out, or clever combinations of variables to play with. It's really a creative process, without any further prescription...

- **Step 1.4:**
  If you are convinced of your improved classifier, you may run it on the full data set of the challenge, prepared for you in a compatible `.root` file:
  https://gitlab.etp.kit.edu/Lehre/dataanalysisexercises_forstudents/-/blob/main/Higgschallenge/atlas-higgs-challenge-2014-v2.root

It will take some time, but after some hours of invested CPU time, you will be rewarded with your personal score for the Higgs Challenge 2014!