6. An introduction to classification with CMS data

Author(s)

Name	Institution	Mail Address	Social Contacts
Lucio Anderlini	INFN Sezione di Firenze	Lucio.Anderlini@fi.infn.it	Hangouts: l.anderlini@gmail.com
Vitaliano Ciulli	Università di Firenze	Vitaliano.Ciulli@fi.infn.it	N/A

How to Obtain Support

Mail	Lucio.Anderlini@fi.infn.it
Social	Hangouts: l.anderlini
Jira	N/A

General Information

ML/DL Technologies	Statistical Learning; Forward Neural Networks
Science Fields	High Energy Physics
Difficulty	Intermediate
Language	English
Type	fully annotated

Software and Tools

Programming Language	Python
ML Toolset	scikit-learn; Keras + Tensorflow
Additional libraries	uproot
Suggested Environments	INFN-Cloud VM, bare Linux Node, Google CoLab

Needed datasets

Data Creator	CMS Experiment
Data Type	Simulation
Data Size	< 1 GB
Data Source	Google Drive

Short Description of the Use Case

Classification is a very common problem in High Energy Physics and historically it has been one of the first to be addressed with machine learning techniques.

In this set of notebooks we discuss the basics on classification taking as use case a simulation of Higgs boson decays as they are reconstructed with the CMS detector, at CERN.

We will review in particular:

- *single-variable cut-based selection*
- *Linear multivariate methods (Fisher discriminant and Logistic Regression)*
- *Decision Trees and Forests*
- *Neural Networks*

All modules are discussed with Python code organized in Colaboratory Notebooks and include exercises of increasing difficulty.

This tutorial targets Master Students in Physics approaching data analysis and is expected to require 1 to 3 hours per module to be completed depending on the student skills.

Here, it is provided to share teaching resources among ML practitioners within INFN.

The Standard Model and the Higgs Boson

The Standard Model (SM) is a set of theoretical models that encode the consolidated knowledge on elementary particles and their interactions. According the the SM, elementary particles are quanta of the excitation of fields (naively, a function of space and time), the properties of the fields are reflected into properties of the particles.

Different fields describe the behavior of matter particles and interactions. The interactions described in the SM are the nuclear strong interaction (whose quanta are named *gluons*), the nuclear weak interaction (whose quanta are named *W bosons*) and the electromagnetic interaction. The fields of the weak and electromagnetic interactions are mixed and the neutral quantum excitation of the former mixes with the quantum excitation of the latter resulting into quanta named *Z boson* and *photon*.

Different matter particles (named *fermions* for brevity) interact via different interaction fields. This allows to further divide matter particles into two categories: quarks are sensitive to the nuclear strong interaction, while leptons are not. Quarks are organized in three families of increasing mass: up and down, charm and strange, and top and bottom. Up, charm and top quarks are positively electrically charged (charge: +2/3 e) while the remaining quarks are negatively charged (charge: -1/3 e). Leptons are also divided in three families, each one composed of a negatively charged particle and a light neutral particle named *neutrino*. Charged leptons are the electron, the muon and the tauon. Different neutrinos are named after the charged lepton in their same family: *electron neutrino, muon neutrino and tau neutrino.*

For each fermion, the Standard Model predicts the existence of a corresponding anti-fermion with opposite electric charge, the fermion number (intended as the difference between the number of fermions and the number of anti-fermions) is conserved in all the perturbative interactions.

The quantum field theory describing so successfully the interactions between particles, cannot fit massive quanta of the interaction fields (named *bosons* fo r short) while experimentally the *W* and *Z bosons* are found to be massive. The problem was solved by Brout Englert and Higgs introducing an additional field with which the W and Z boson interact gaining rest energy (which is another name for the mass) in a symmetry breaking mechanism, named BEH mechanisms after the three theorists. If such a field exists, then it would be natural to describe the mass of all particles, including matter particles, as due to interactions with the BEH field rather than with additional free parameters in the model. The interaction of the BEH field with fermions is described by the *Yukawa couplings*. The new field must be associated to an excitation quantum, named *Higgs boson*. Discovering such a predicted particle was for long considered as the smoking-gun evidence of the correctness of the SM and motivated a huge effort in the development and construction of the Large Hadron Collider, at CERN, culminated in 2012 with the discovery of the Higgs boson analysing the data of the two major experiments named *ATLAS* and *CMS.*

The Large Hadron Collider and the CMS experiment at CERN

The Large Hadron Collider (LHC) is the largest particle accelerator currently active in the world. Locate at an average depth of 100 meters underground at the CERN facilities in Geneva, it spans a 27 km circumference crossing the Swiss-French border. It is capable of accelerating beams of protons or heavy ions in two opposite directions, producing proton-proton, ion-ion, and proton-ion collisions. LHC accelerates the proton beams up to maximum energy of 7 TeV.

The beams are made to collide in four points along the circumference, where the four main experiments are located: ATLAS, LHCb, ALICE and CMS. The choice of a collider configuration, as opposed to a fixed target setup, is motivated by the fact that this configuration gives the maximum which allows to explore the production of particles of higher mass.

The CMS experiment, acronym of *Compaact Muon Solenoid*, is one of the four major experiments installed at the LHC. In 2012, together with ATLAS, the CMS Collaboration reported the observation of a new scalar resonance, compatible with the theoretical prediction for the *Higgs boson.* Although this observation was the driving reason behind its construction, CMS is designed to allow a vast range of physics studies. Built with a cylindrical geometry coaxial to the beam pipe, the detector measures approximately 22 m in length and 15 m in diameter. It is comprised of a complex system of subdetectors able to detect photons, electrons, muons and hadronic jets of particles produced in the collisions. The coordinate system in use is a hybrid reference frame composed of three variables:

- the radius, intended as the distance from the LHC beam axis
- the pseudorapidity, which is a monotonic transformation of the polar angle *theta* *[see Wikipedia]*
- the azimuthal angle *phi*

Production of the Higgs boson

When two protons accelerated by the LHC collide, their constituent quarks and gluons have a chance to interact. Because of the internal quantum structure of the proton, its constituents "*carry"* different fractions of the total energy of the proton. Indeed, in contrast with what happens classically, the number of constituents within the proton is not constant and the three "*valence quark*" (two *up* quarks and a *down* quark) interact via the strong interacting emitting gluons and creating from vacuum pairs of quark-antiquarks (named *sea-quark*) that all travel at nearly the speed of light within the accelerated protons. It is impossible to predict, at the time of the collision, what is the exact share of energy of the interacting quarks or gluons, but one can define a *dist ribution* of the probability of finding a given constituent of the proton with a given fraction *x* of its total energy.

When the energy of the collision between the protons becomes very large, the probability of hitting on a high-energy gluon becomes much larger than the probability of hitting on a *valence* or *sea* quark. Hence the dominant mechanism in the production of the Higgs boson is mediated by gluons and is named gluon-gluon fusion. Since gluons are massless and cannot interact directly with the Higgs boson, some higher-order mechanism has to intervene, but this is beyond the scope of this brief introduction.

The take-home message is that the two interacting gluons from which the Higgs boson originates carry a potentially different fraction of the energy of their respective protons. This is a crucial point because it basically washes out most information one can try to obtain from the longitudinal momentum of the produced Higgs: a huge random component on it is due to the fraction of the proton energy carried by the two constitutent particles that randomly interacted to give origin to the Higgs. Still, the Higgs boson is a very heavy particle: in order to make its production kinematically permitted, the collision energy of the two interacting gluons must be at least as large as the Higgs boson mass, which means that they *both* must carry a large fraction of their respective proton energy. This requirement pushes the particles produced in *hard scattering* processes involving the production of heavy particles (among which the Higgs) towards "*large angles"* with respect to the beam axis. In contrast, *soft collisions*, which result from the vast majority of the proton-proton interactions are directed at small angle with respect to the beam axis.

This property of the collisions producing heavy particles is probably the most discriminant feature that allows to identify the collisions of interest: for example, most of the trigger algorithms designed to run in real time discarding the vast majority of collision events where it is unlikely to identify Higgs bosons, require at least one observed particle with a large *transverse momentum* (the component of the momentum orthogonal to the beam axis).

Decays of the Higgs boson

Given that the Higgs boson couples to every massive particle in the SM, its phenomenology is particularly rich. The higher the mass of the particle, the higher the coupling of the Higgs boson with the particle. This reasoning may help understanding naively why the decay of the Higgs to two massive bosons (*ZZ)* or (*WW*) is expected to be one of the most abundant decay modes of the Higgs boson produced in proton-proton collisions at the LHC.

For this exercise we will focus in particular on the decay of the Higgs boson to two charged *W* particles, which lead recently to an important publication on which you may read more here: http://cms.cern/news/examining-how-higgs-boson-shapes

The *W* bosons decay to a charged pair of fermion-antifermion. If the fermions are quarks then because of the strong-nuclear interaction they immediately start to create pairs of quarks-antiquarks until their *color-charge* (equivalent to the electric charge, but for the strong-interaction) is neutralized. This process is named hadronization and leads to the creation of *jets* of particles that flow in the detector in a particularly narrow cone.

Conversely, if the fermion-antifermion pair produced in the W decay are leptons, then one must be a neutrino which is not detectable from CMS and escapes while subtracting the overall energy of the collision event a substantial fraction, possibly breaking the cylindrical symmetry of the system. The other lepton is instead detected very effectively by the CMS detector because its energy is very high so that it is difficult to confuse it with a lepton produced in other processes.

The dataset

This set of exercises is based on a simulated dataset of Higgs decays to two W bosons, both decaying to a charged lepton and a neutrino. In formula,

H WW (l nu) (l nu)

Another simulated dataset of a pair of W boson produced in the same proton-proton collision, but without involving the production of the Higgs is used to estimate the contribution of the most important *background*, i.e. the source of the majority of erroneously selected decay events.

The simulated decay has been reconstructed using the official reconstruction algorithms as for the real data collected with the CMS detector and condensed into few *physics-motivated* discriminant features.:

- the invariant mass of the two leptons
- the missing transverse mass of the higgs
- the number of particle jets reconstructed in the event

- the transverse momentum of the lepton pair
- the R variable of the lepton pairs
- the *improved missing transverse mass*, defined as the invariant mass of the leptons added to the missing transverse momentum
- of the lepton pair;
- between the lepton pair and the missing transverse momentum;
- between the first of the leptons and the missing transverse momentum;
- the missing transverse energy.
- the azimutal angle of the missing energy.

The mathematical definition of most variables is discussed using MathJax in the first notebook.

This is the starting point for most statistical analyses in High Energy Physics: one has to develop an algorithm to classify signal and background events using simulation. That algorithm will then be run on real data and the resulting selected events have to be statistically subtracted for the expected contribution from background events in order to *count* the number of signal events and use them to infer physical properties of the decay.

How to execute it

Getting started (https://colab.research.google.com/drive/1C0-zM3tRRGhrL7XbqolmlUTtrwKaunhQ?usp=sharing)

In this first notebook we discuss the dataset, learn how to plot histograms and use them to define a 1D selection strategy comparing the discriminant power of different variables.

Purity and selection efficiency are discussed and the ROC curve is introduced.

Software-side, we used numpy and matplotlib, only, with explicit implementation of some of the most common operations with dedicated libraries (drawing the ROC curve).

Linear classification in two dimensions (https://colab.research.google.com/drive/1LAaNkpILnPiQL2fhlCYuTeTTVyy1zrP7?usp=sharing)

In this notebook we focus on the multivariate selection based on linear discrimants. We discuss the Fisher Discriminant and the Logistic Regression.

Finally we introduce an approximated likelihood ratio to demonstrate the importance of the hypotheses injected with the choice of the algorithm and introduce the concept of the No Free Lunch Theorem.

Software-side, we introduce scikit-learn, after having discussed the implementation of the various algorithm with pure numpy.

Decision Trees and Forests (https://colab.research.google.com/drive/18NHWmORaYDtpQq74mKJwTb97UZ1R2w0w?usp=sharing)

We move on to discuss decision trees, ensembles of random decision trees to reduce the variance of the method, and then consider two widely adopted boosting algorithms: AdaBoost and Gradient Boosting.

The discussion of the algorithms is performed reducing the dimensionality to the two most-discriminant variables to allow plotting the decision boundaries. In the end a study of the robustness of the decision forests to the curse of dimensionality is given.

Software-side, we play with scikit-learn.

Neural Networks (https://colab.research.google.com/drive/1pKKELV34Hiori7_19Ukbbm6Olw_FxxHz?usp=sharing)

Finally, we extend the Logistic Regression discussed in the Linear-classification notebook modifying the representation through a neural network implemented in plain tensorflow 2, and then in keras.

A brief discussion on overtraining and on regularization techniques is also provided.

Finally, we introduce compound systems of neural networks discussing domain adaptation and proposing semi-supervised learning as an exercise.

Software-side, this notebook is significantly more demanding than the previous ones.

Conclusions

We developed four Colab notebooks to introduce students to the problem of classification taking as an example the signal-background separation in High Energy Physics, and in particular in CMS for a hot topic such as the study of the Higgs boson.

Presentation made on 27 Jul 2020 : https://agenda.infn.it/event/23648/contributions/118758/attachments/74371/94574/Teaching_Material_on_the_KB.pdf