Learning to discover: the Higgs boson machine learning challenge

Claire Adam-Bourdarios[a], Glen Cowan[b], Cécile Germain[c], Isabelle Guyon[d], Balázs Kégl[a,c], David Rousseau[a]

[a] LAL, IN2P3/CNRS & University Paris-Sud, France
[b] Physics Department, Royal Holloway, University of London, UK
[c] TAO team, INRIA & LRI, CNRS & University Paris-Sud, France
[d] ChaLearn

10 December 2014, version 2.1

Preamble

The Higgs boson machine learning challenge (HiggsML or Challenge in short) has been set up to promote collaboration between high energy physicists and data scientists. The ATLAS experiment at CERN provided simulated data used by physicists to optimize the analysis of the Higgs boson. The Challenge is organized by a small group of ATLAS physicists and data scientists. It has been hosted by Kaggle at https://www.kaggle.com/c/higgs-boson; the challenge data is now available on https://opendata.cern.ch/education/ATLAS;. This document provides technical background for the Challenge; reading and understanding it is not necessary to participate in the Challenge, but it is probably useful. No prior knowledge of high energy physics is required. This document has been minimally modified to serve as a permanent documentation for the dataset on opendata.cern.ch.

Contents

1 Introduction

The ATLAS experiment and the CMS experiment recently claimed the discovery of the Higgs boson [1, 2]. The discovery was acknowledged by the 2013 Nobel prize in physics given to François Englert and Peter Higgs. This particle was theorized almost 50 years ago to have the role of giving mass to other elementary particles. It is the final ingredient of the Standard Model of particle physics, ruling subatomic particles and forces. The experiments are running at the Large Hadron Collider (LHC) at CERN (the European Organization for Nuclear Research), Geneva, which began operating in 2009 after about 20 years of design and construction, and which will continue operating for at least the next 10 years.

The Higgs boson has many different processes through which it can *decay*. When it decays, it produces other particles. In physics, a decay into specific particles is called a *channel*. The Higgs boson has been seen first in three distinct decay channels which are all boson pairs. One of the next important topics is to seek evidence on the decay into fermion pairs, namely tau-leptons or *b*-quarks, and to precisely measure their characteristics. The first evidence of the H to tau tau channel was recently reported by the ATLAS experiment [3], which, in the rest of this paper, will be referred to as "the reference document" (it should be noted the final ATLAS paper [4] on the subject is being submitted to publication but changes with respect to the reference document are not relevant for the Challenge). The subject of the Challenge was to try and improve on this analysis. Given the broad success of the challenge on the Kaggle forum, and the wish of many participants and others to pursue the study beyond the end of the challenge, the dataset has been made permanently available on `https://opendata.cern.ch/education/ATLAS` [5]. Accompanying software [6] and the present document [7] are available there. The present document is to serve as a permanent documentation for the dataset on opendata.cern.ch; it has been minimally modified w.r.t. the document made available to the Challenge participants, all such modifications contain the key word "opendata.cern.ch".

1.1 The detector and the analysis pipeline

In the LHC, proton bunches are accelerated on a circular trajectory in both directions. When these bunches cross in the ATLAS detector, some of the protons collide, producing hundreds of millions of proton-proton collisions per second. The up to hundreds of particles resulting from each bunch crossing (called an *event*) are detected by sensors, producing a sparse vector of about a hundred thousand dimensions (roughly corresponding to an image or speech signal in classical machine learning applications). From this raw data, the type, the energy, and the 3D direction of each particle are estimated. In the last step of this feature construction phase, this variable-length list of four-tuples is digested into a fixed-length vector of features containing up to tens of real-valued variables.

Some of these variables are first used in a real-time multi-stage cascade classifier (called the *trigger*) to discard most of the uninteresting events (called the *background*). The selected events (roughly four hundred per second) are then written on disks by a large CPU farm, producing petabytes of data per year. The saved events still, in large majority, represent known processes (called *background*): they are mostly produced by the decay of particles which are exotic in everyday terms, but known, having been discovered in previous generations of experiments. The goal of the offline analysis is to find a (not necessarily connected) region in the feature space in which there is a significant excess of events (called *signal*) compared to what known background processes can explain. Once the region has been fixed, a statistical (counting) test is applied to determine the significance of the excess. If the probability that the excess has been produced by background processes falls below a limit,[1] the new particle is deemed to be discovered.[2]

[1] Usually $p = 2.87 \times 10^{-7}$ or its equivalent Gaussian significance $Z = \Phi^{-1}(1-p) = 5$ sigma, where Φ^{-1} is the standard normal quantile.

[2] The real analysis is somewhat more complex since it has to deal with nuisance parameters and unknown properties of the new particle (such as its mass). In this note, and in the Challenge, we assume that these parameters are fixed, so the goal is to test the background only hypothesis against background plus signal with *fixed and known* parameters.

Today, multivariate classification techniques are routinely used to optimize the selection region [3, 8, 9]. The formal objective function is unique and quite different from the classification error or other (e.g., ranking, cost-sensitive classification) objectives used regularly in machine learning. Nevertheless, finding a "pure" signal region corresponds roughly to separating background and signal events, and so classical classification methods proved to be useful in the past, often allowing better discovery sensitivity, compared to more traditional, manual ("cut-based") techniques.

The classifier is trained on simulated background and signal events. Simulators produce weights for each event to correct for the mismatch between the natural (prior) probability of the event and the instrumental probability applied by the simulator (an importance-sampling flavor). The weights are normalized such that in any region the sum of the weights of events falling in the region gives an unbiased estimate of the expected number of events found there for a fixed *integrated luminosity*, which corresponds to a fixed data taking time for a given beam intensity. In our case this corresponds to the data collected by the ATLAS Experiment in 2012. Since the probability of a signal event is usually several orders of magnitudes lower than the probability of a background event, the signal and background samples are usually re-normalized to produce a balanced classification problem. A real-valued discriminant function is then trained on this re-weighted sample to minimize the weighted classification error. The signal region is then defined by cutting the discriminant value at a certain threshold, which is optimized on a held-out set to maximize the sensitivity of the statistical test.

1.2 The goal of the Challenge

The goal of the Challenge is to improve the procedure that produces the selection region. We provide a training set with signal/background labels and with weights, a test set (without labels and weights) (in the opendata.cern.ch dataset these samples are merged and label and weights are provided in all cases) and a formal objective representing an approximation of the median significance (AMS) of the counting test. The objective is a function of the weights of selected events. We expect that significant improvements are possible by re-visiting some of the ad hoc choices in the standard procedure, or by incorporating the objective function or a surrogate into the classifier design.

1.3 The structure of the document

The paper is written for a mixed audience, so we make it clear who the target is of each section. Section 2 describes the problem formally and gives just enough details for understanding the Challenge. Those who wish to participate in the Challenge (both from the physics and the machine learning side) without bothering with the reasoning behind the objective function and the science case can skip the rest of the paper. Section 3 summarizes the physics behind the Challenge, including a more thorough description of the detector and the analysis pipeline (summarized in Section 1.1) and a detailed description of the data and the features. This section is essentially written for non-physicists interested in some of the details of the physics behind the Challenge. Section 4 shows the derivation of the AMS objective used to evaluate the Challenge submissions. The section is based on standard statistics and it is designed both for physicists and computer scientists who are interested in the origins of the formula. Some of the discussion in this section can serve as a basis for designing surrogates which could be incorporated into the algorithmic design. Section 5 details the differences between the Kaggle data sets made available during the Challenge, and the final opendata.cern.ch data set.

2 The formal problem

For the formal description of the Challenge, let $\mathcal{D} = \{(\mathbf{x}_1, y_1, w_1), \ldots, (\mathbf{x}_n, y_n, w_n)\}$ be the training sample, where $\mathbf{x}_i \in \mathbb{R}^d$ is a d-dimensional feature vector, $y_i \in \{\mathrm{b}, \mathrm{s}\}$ is the label, and $w_i \in \mathbb{R}^+$

is a non-negative weight. Let $\mathcal{S} = \{i : y_i = \mathrm{s}\}$ and $\mathcal{B} = \{i : y_i = \mathrm{b}\}$ be the index sets of signal and background events, respectively, and let $n_\mathrm{s} = |\mathcal{S}|$ and $n_\mathrm{b} = |\mathcal{B}|$ be the numbers of simulated signal and background events.

There are two properties that make our simulated training set different from those collected in nature or sampled in a natural way from a joint distribution $p(\mathbf{x}, y)$.[3] First, we can simulate as many events as we need (given enough computational resources), so the proportion $n_\mathrm{s}/n_\mathrm{b}$ of the number of points in the two classes does not have to reflect the proportion of the prior class probabilities $P(y = \mathrm{s})/P(y = \mathrm{b})$. This is actually a good thing: since $P(y = \mathrm{s}) \ll P(y = \mathrm{b})$, the training sample would be very unbalanced if the numbers of signal and background events, n_s and n_b, were proportional to the prior class probabilities $P(y = \mathrm{s})$ and $P(y = \mathrm{b})$. Second, our simulators produce importance-weighted events. Since the objective function (7) will depend on the *unnormalized sum* of weights, to make the setup invariant to the *numbers* of simulated events n_s and n_b, the sum across each set (training, public test, private test, etc.) and each class (signal and background) will be kept fixed, that is,

$$\sum_{i \in \mathcal{S}} w_i = N_\mathrm{s} \quad \text{and} \quad \sum_{i \in \mathcal{B}} w_i = N_\mathrm{b}. \tag{1}$$

The normalization constants N_s and N_b have physical meanings: they are the *expected total number* of signal and background events, respectively, during the time interval of data taking (the year of 2012 in our case). The individual weights are proportional to the conditional densities divided by the instrumental densities used by the simulator, that is,

$$w_i \sim \begin{cases} p_\mathrm{s}(\mathbf{x}_i)/q_\mathrm{s}(x_i), & \text{if } y_i = \mathrm{s}, \\ p_\mathrm{b}(\mathbf{x}_i)/q_\mathrm{b}(x_i), & \text{if } y_i = \mathrm{b}, \end{cases} \tag{2}$$

where

$$p_\mathrm{s}(\mathbf{x}_i) = p(\mathbf{x}_i | y = \mathrm{s}) \quad \text{and} \quad p_\mathrm{b}(\mathbf{x}_i) = p(\mathbf{x}_i | y = \mathrm{b})$$

are the conditional signal and background densities, respectively, and $q_\mathrm{s}(\mathbf{x}_i)$ and $q_\mathrm{b}(\mathbf{x}_i)$ are instrumental densities.

Let $g : \mathbb{R}^d \to \{\mathrm{b}, \mathrm{s}\}$ be an arbitrary classifier. Let the *selection region* $\mathcal{G} = \{\mathbf{x} : g(\mathbf{x}) = \mathrm{s}\}$ be the set of points classified as signal, and let $\widehat{\mathcal{G}}$ denote the *index set* of points that g *selects* (physics terminology) or *classifies as signal* (machine learning terminology), that is,

$$\widehat{\mathcal{G}} = \{i : \mathbf{x}_i \in \mathcal{G}\} = \{i : g(\mathbf{x}_i) = \mathrm{s}\}.$$

Then from Eqs. (1) and (2) it follows that the quantity

$$s = \sum_{i \in \mathcal{S} \cap \widehat{\mathcal{G}}} w_i \tag{3}$$

is an unbiased estimator of the expected number of signal events selected by g,

$$\mu_\mathrm{s} = N_\mathrm{s} \int_{\mathcal{G}} p_\mathrm{s}(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{4}$$

and, similarly,

$$b = \sum_{i \in \mathcal{B} \cap \widehat{\mathcal{G}}} w_i \tag{5}$$

is an unbiased estimator of the expected number of background events selected by g,

$$\mu_\mathrm{b} = N_\mathrm{b} \int_{\mathcal{G}} p_\mathrm{b}(\mathbf{x}) \mathrm{d}\mathbf{x}. \tag{6}$$

[3] We use small p for denoting probability densities and capital P for denoting the probability of random events.

In physics, the estimated number of events s and b selected by g are also usually called signal and background.[4] In machine learning terminology, s and b are true and false positive rates, respectively.[5]

Given a classifier g, a realization of the experiment with n observed events selected by g (positives), the (Gaussian) significance of discovery would be roughly $(n - \mu_b)/\sqrt{\mu_b}$ standard deviations (sigma)[6] since the Poisson fluctuation of the background has a standard deviation of $\sqrt{\mu_b}$. Since we can estimate n by $s + b$ and μ_b by b, this would suggest an objective function of $s/\sqrt{b}$ for training g. Indeed, the first order behavior of all objective functions is $\sim s/\sqrt{b}$, but it is only valid when $s \ll b$ and $b \gg 1$, which is often not the case in practice. To improve the behavior of the objective function in this range, we use the *approximate median significance* (AMS) objective function defined by

$$\text{AMS} = \sqrt{2\left((s + b + b_{\text{reg}})\ln\left(1 + \frac{s}{b + b_{\text{reg}}}\right) - s\right)} \tag{7}$$

where s and b are defined in Eqs. (3) and (5), respectively, and b_{reg} is a regularization term set to a constant $b_{\text{reg}} = 10$ in the Challenge. Eq. (7) with $b_{\text{reg}} = 0$, coming from [10], is used frequently by high-energy physicists for optimizing the selection region for discovery significance; the regularization term b_{reg} was introduced for the Challenge in order to reduce the variance of the AMS. The derivation of the AMS formula is explained in Section 4.

As a summary, the task of the participants is to train a classifier g based on the training data $\mathcal{D}$ with the goal of maximizing the AMS (7) on a held-out (test) data set. On the opendata.cern.ch dataset, the training vs test splitting is left free to the participant.

3 Physics background

This section elaborates on some of the details on the detector and the analysis left out of Section 1.1. Understanding these details is not necessary to participate in the Challenge, but it can provide some background and show to the participants where their contributions may fit into the analysis chain. For further information, we refer the reader to the 21 December 2012 special edition of Science Magazine "Breakthrough of the Year : the Higgs Boson"[7], in particular, a non-specialist account of the discovery of the Higgs boson [11].

3.1 Proton collisions and detection

The LHC collides bunches of protons every 50 nanoseconds within each of its four experiments, each crossing producing a random number of proton-proton collisions (with a Poisson expectation between 10 and 35, depending on the LHC parameters) called events[8]. Two colliding protons produce a small firework in which part of the kinetic energy of the protons is converted into new particles. Most of the resulting particles are very unstable and decay quickly into a cascade of lighter particles. The ATLAS detector measures three properties of these surviving particles (the so-called *final state*): the *type* of the particle (electron, photon, muon, etc.), its *energy*, and the 3D *direction* of the particle. Based on these properties, the properties of the decayed parent particle is inferred, and the inference chain is continued until reaching the heaviest primary particles.

[4] In particle physics, depending on the context, the term *signal* can mean s, μ_s, or s. In this note, since we need all of these three quantities, we use roman s to denote the label and in indices of terms related to signal (e.g., n_s), μ_s (4) for the *expected* number of signal events in a subspace selected by a classifier, and s (3) for the *estimated* number of signal events selected by a classifier. The same logic applies to the three terms b, μ_b (6), and b (5), all usually referred to as background.

[5] s and b are *unnormalized*, more precisely, *luminosity-normalized* (1) true and false positive rates.

[6] See footnote 1 and Section 4.

[7] http://www.sciencemag.org/content/338/6114.toc

[8] Numbers here and later refer specifically to data taken in the year of 2012 in ATLAS. Simulated data provided for the Challenge also corresponds to this period.

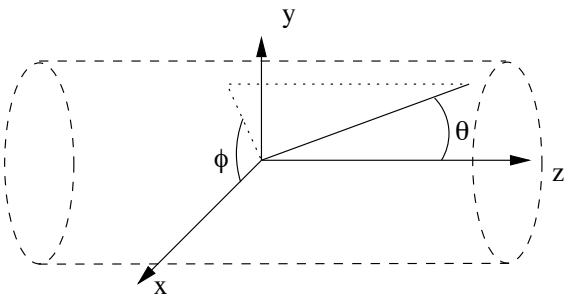

Figure 1: ATLAS reference frame

An online trigger system discards most of the bunch collisions containing uninteresting events. The trigger is a three-stage cascade classifier which decreases the event rate from 20 000 000 to about 400 per second. The selected 400 events are saved on disk, producing about one billion events and three petabytes of raw data per year.

Each event contains about ten particles of interest in the final state, which are reconstructed from hundreds of low-level signals. The different types of particles or pseudo particles of interest for the Challenge are electrons, muons, hadronic tau, jets, and missing transverse energy. Electrons, muons, and taus are the three leptons[9] from the standard model. Electrons and muons live long enough to reach the detector, so their properties (energy and direction) can be measured directly. Taus, on the other hand, decay almost immediately after their creation into either an electron and two neutrinos, a muon and two neutrinos, or a bunch of charged particles and a neutrino. The bunch of hadrons can be identified as a pseudo particle called the hadronic tau. Jets are pseudo particles rather than real particles; they originate from a high energy quark or gluon, and they appear in the detector as a collimated energy deposit associated with charged tracks. The measured momenta (see Appendix A for a short introduction to special relativity) of all the particles of the event is the primary information provided for the Challenge.

We will use the conventional 3D direct reference frame of ATLAS throughout the document (see Fig 1): the z axis points along the horizontal beam line, and the x and y axes are in the transverse plane with the y axis pointing towards the top of the detector. θ is the polar angle and ϕ is the azimuthal angle. Transverse quantities are quantities projected on the $x-y$ plane, or, equivalently, quantities for which the z component is omitted. Instead of the polar angle θ, we often use the *pseudorapidity* $\eta = -\ln\tan(\theta/2)$; $\eta = 0$ corresponds to a particle in the $x-y$ plane ($\theta = \pi/2$), $\eta = +\infty$ corresponds to a particle traveling along the z-axis ($\theta = 0$) direction and $\eta = -\infty$ to the opposite direction ($\theta = \pi$). Particles can be identified in the range $\eta \in [-2.5, 2.5]$. For $|\eta| \in [2.5, 5]$, their momentum is still measured but they cannot be identified. Particles with $|\eta|$ beyond 5 escape detection along the beam pipe. The missing transverse energy is a pseudo-particle which deserves a more detailed explanation. The neutrinos produced in the decay of a tau escape detection completely. We can nevertheless infer their properties using the law of momentum conservation by computing the vectorial sum of the momenta of all the measured particles and subtracting it from the zero vector. In practice, there are measurement errors for all particles which make the sum poorly estimated. Another difficulty is that many particles are lost in the beam pipe along the z axis, so the information on momentum balance is lost in the direction of the z axis. Thus we can carry out the summation only in the transverse plane, hence the name missing *transverse* energy, which is a 2D vector in the transverse plane.

To summarize, for each event, we produce a list of momenta for zero or more particles for each type, plus the missing transverse energy which can always be measured. For the Challenge, we selected only events with one electron or one muon (exclusively), and one hadronic tau. These two particles should be of opposite electric charge. In addition, events with identified b-quark

[9]For the list of elementary particles and their families, we refer the reader to http://www.sciencemag.org/content/338/6114/1558.full.

jets were rejected, which helps to reject some of the background sources[10].

3.2 The physics goal

In the Challenge, the positive (signal) class is comprised of events in which the Higgs boson decays into two taus. This channel is interesting from a theoretical point of view but experimentally challenging. In the Standard Model (SM), the Higgs boson is the particle which is responsible for the mass of the other elementary particles. To verify the Standard Model, it is important to measure the coupling (which can be seen as the strength of the force) of the Higgs boson to other particles and check their consistency with the prediction of the SM. In the original discovery, the Higgs boson was seen decaying into $\gamma\gamma$, WW, and ZZ, which are all boson pairs [1, 2] (bosons are carriers of forces). What about the couplings to fermions, of which matter is made? We know indirectly that the coupling of the Higgs boson to quarks (which are fermions) cannot be very different from what the SM predicts, otherwise the Higgs production cross section (the number of Higgs bosons produced independently of the way it decays) would be significantly different of what has been measured. On the other hand, currently we have little direct information on the coupling of the Higgs boson to leptons (electrons, muons, taus, and their associated neutrinos). For example, given the elusive nature of neutrinos, their minuscule mass, and the way they oscillate between flavors, one could very well imagine that the mass of leptons comes from an entirely different mechanism. Hence the importance of measuring as precisely as possible the coupling of the Higgs to tau (given that the coupling of the Higgs to electrons and muons is beyond the reach of the ATLAS experiment due to the small masses of these two leptons).

The channel of Higgs decaying into two taus is experimentally challenging for essentially two reasons. First, since neutrinos are not measured in the detector, their presence in the final state makes it difficult to evaluate the mass of the Higgs candidate on an event-by-event basis. Second, the Z boson can also decay in two taus, and this happens much more frequently than the decay of the Higgs. Since the mass of a Z (91 GeV) is not very far (within about one standard deviation of the resolution of the mass measurement) from the mass of the Higgs (125 GeV), the two decays produce similar events which are difficult to separate.

We are focusing on one particular topology among the many possible ones: events where one tau decays into an electron or a muon and two neutrinos, and the other tau decays in hadrons and a neutrino. We first extract a fixed number of features (described in Section 3.3 and Appendix B) from the variable-length list of 4-tuples. Then a classifier (trained on simulated events) is used to select a signal-rich region in the feature space. Once the region is fixed, we check whether the number of real events in the selected region is significantly higher than the number of background events predicted by the model, and if the significance is higher than five sigma, the new particle is declared discovered.

The approximate median significance AMS (7) provides a way to evaluate the quality of the classifier rapidly. However, it should be noted that in the final analysis presented in the reference document [3], the evaluation of the significance (and even the determination of the expected significance based on simulations) involves a complex fit of many parameters, taking into account various factors and systematic uncertainties.

3.3 The data

For the Challenge, we provide simulated events using the official ATLAS full detector simulator. The simulator has two parts. In the first, random proton-proton collisions are simulated based on all the knowledge that we have accumulated on particle physics. It reproduces the random microscopic explosions resulting from the proton-proton collisions. In the second part, the resulting particles are tracked through a virtual model of the detector. The process yields simulated events with properties that mimic the statistical properties of the real events with additional information on what has happened during the collision, before particles are measured in the detector.

[10] These two pieces of information are only useful when comparing the Challenge to the reference document.

The signal sample contains events in which Higgs bosons (with fixed mass 125 GeV) were produced. The background sample was generated by other known processes which can produce events with at least one electron or muon and a hadronic tau, mimicking the signal. For the sake of simplicity, only three background processes were retained for the Challenge. The first comes from the decay of the Z boson (with mass 91.2 GeV) in two taus. This decay produces events with a topology very similar to that produced by the decay of a Higgs. The second set contains events with a pair of top quarks, which can have lepton and hadronic tau among their decay. The third set involves the decay of the W boson, where one electron or muon and a hadronic tau can appear simultaneously only through imperfections of the particle identification procedure.

Due to the complexity of the simulation process, each simulated event has a weight (2) which is proportional to the conditional density divided by the instrumental density used by the simulator (an importance-sampling flavor), and normalized for integrated luminosity such that, in any region, the sum of the weights of events falling in the region is an unbiased estimate of the expected number of events falling in the same region during a given fixed time interval. In our case, the weights correspond to the quantity of real data taken during the year 2012. The weights are an artifact of the way the simulation works and so they are not part of the input to the classifier. For the Challenge, weights have been provided in the training set so the AMS (7) can be properly evaluated. Weights were not provided it in the qualifying set since the weight distribution of the signal and background sets are very different and so they would give away the label immediately. However, in the opendata.cern.ch dataset, weights and labels have been provided for the complete dataset.

The $d = 30$ features extracted from the raw date are described in Appendix B after a crash course in special relativity in Appendix A, needed for the formal definition of the features.

4 The statistical model

In this section we describe the basic structure of the statistical test that leads to the criterion (7) presented in Section 2. The goal is to construct a statistical test of the hypothesis that the signal process is absent. The derivation is based on [10].

In Section 4.1 we describe the statistical test and the discovery significance computed on real data. In Section 4.2 we derive the approximate median significance that can be used to estimate[11] the significance on simulated events and to optimize the statistical test, and we explain why we apply a regularization term for not letting b approach zero.

It is not possible to reproduce all of the complexities of the search for the signal process in the Challenge. In particular, some simplifying assumptions are made concerning the treatment of uncertainties in the estimated rates of background processes. Furthermore, the search is carried out by counting the number of events found satisfying certain criteria; one may extend this to multiple counts with different criteria or to use of a more sophisticated likelihood-ratio test. The simplifying assumptions used here preserve the main features of a more complicated analysis and are expected to have only a minor influence on its sensitivity and on the broader question of how machine learning can be used to improve the search.

4.1 Statistical treatment of the measurement

Each proton-proton collision ("event") is characterized by a set of measured quantities, the input variables $\mathbf{x} \in \mathbb{R}^d$. A simple but realistic type of analysis is where one counts the number of events found in a given region in the space of input variables (the "search region", denoted below as $\mathcal{G}$), which is defined by the classifier g, that is, $\mathcal{G} = \{\mathbf{x} : g(\mathbf{x}) = \mathrm{s}\}$. If we fix the classifier g, the

[11] Since the training data is not coming from real observations, rather it is generated by simulators, it may be more appropriate to use the term *approximate*, as in approximating an integral by Monte-Carlo integration. We stick to the term *estimate* to comply with the classical terminology in machine learning.

number of events n found in $\mathcal{G}$ is assumed to follow a Poisson distribution with mean $\mu_s + \mu_b$,

$$P(n|\mu_s, \mu_b) = \frac{(\mu_s + \mu_b)^n}{n!} e^{-(\mu_s+\mu_b)}, \tag{8}$$

where μ_s and μ_b are the expected numbers of events from the signal and background, respectively. To establish the existence of the signal process, we test the hypothesis of $\mu_s = 0$ (the background-only hypothesis) against the alternative where the signal exists, that is, $\mu_s > 0$. From Eq. (8), we construct the likelihood ratio

$$\lambda = \frac{P(n|0, \mu_b)}{P(n|\hat{\mu}_s, \mu_b)} = \left(\frac{\mu_b}{\hat{\mu}_s + \mu_b}\right)^n e^{\hat{\mu}_s} = \left(\frac{\mu_b}{n}\right)^n e^{n-\mu_b}, \tag{9}$$

where $\hat{\mu}_s = n - \mu_b$ is the maximum likelihood estimator of μ_s given that we observe n events in the selection region $\mathcal{G}$.

According to Wilks' theorem [12], given that certain regularity conditions are satisfied, the test statistic

$$q_0 = \begin{cases} -2\ln\lambda & \text{if } n > \mu_b, \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

approaches a simple asymptotic form related to the chi-squared distribution in the large-sample limit. In practice the asymptotic formulae are found to provide a useful approximation even for moderate data samples (see, e.g., [10]). Assuming that these hold, the p-value of the background-only hypothesis from an observed value of q_0 is found to be

$$p = 1 - \Phi\left(\sqrt{q_0}\right), \tag{11}$$

where Φ is the standard Gaussian cumulative distribution.

In particle physics it is customary to convert the p-value into the equivalent *significance* Z, defined as

$$Z = \Phi^{-1}(1 - p), \tag{12}$$

where Φ^{-1} is the standard normal quantile. Eqs. (11) and (12) lead therefore to the simple result

$$Z = \sqrt{q_0} = \sqrt{2\left(n\ln\left(\frac{n}{\mu_b}\right) - n + \mu_b\right)} \tag{13}$$

if $n > \mu_b$ and $Z = 0$ otherwise. The quantity Z measures the statistical significance in units of standard deviations or "sigmas". Often in particle physics a significance of at least $Z = 5$ (a five-sigma effect) is regarded as sufficient to claim a discovery. This corresponds to finding the p-value less than 2.9×10^{-7}.[12]

4.2 The median discovery significance

Eq. (13) represents the significance that we would obtain for a given number of events n observed in the search region $\mathcal{G}$, knowing the background expectation μ_b. When optimizing the design of the classifier g which defines the search region $\mathcal{G} = \{\mathbf{x} : g(\mathbf{x}) = s\}$, we do not know n and μ_b. As usual in empirical risk minimization [13], we estimate the expectation μ_b by its empirical counterpart b from Eq. (5). We then replace n by $s + b$ to obtain the approximate median significance

$$\text{AMS}_2 = \sqrt{2\left((s+b)\ln\left(1 + \frac{s}{b}\right) - s\right)}. \tag{14}$$

AMS_2 can be rewritten as :

$$\text{AMS}_2 = \text{AMS}_3 \times (1 + \mathcal{O}\left(\frac{s}{b}\right)),$$

[12] This extremely high threshold for statistical significance is motivated by a number of factors related to multiple testing, accounting for mismodeling and the high standard one would like to require for an important discovery.

where

$$\mathrm{AMS}_3 = \frac{s}{\sqrt{b}}. \tag{15}$$

The two criteria Eqs. (14) and (15) are practically indistinguishable when $b \gg s$. This approximation often holds in practice and may, depending on the chosen search region, be a valid surrogate in the Challenge.

In preliminary runs it happened sometimes that AMS_2 was maximized in small selection regions $\mathcal{G}$, resulting in a large variance of the AMS. While large variance in the real analysis is not necessarily a problem, it would make it difficult to reliably compare the participants of the Challenge if the optimal region was small. So, in order to decrease the variance of the AMS, we decided to bias the optimal selection region towards larger regions by adding and artificial shift b_{reg} to b. The value $b_{\mathrm{reg}} = 10$ was determined using preliminary experiments.

5 opendata.cern.ch dataset

The data set which was available to Challenge participants on the Kaggle platform (https://www.kaggle.com/c/higgs-boson) during the duration of the Challenge, is now permanently available on the opendata.cern.ch platform (https://opendata.cern.ch/education/ATLAS). Some minimal modifications of the data set were necessary, which are detailed here:.

- The 100000 event training data set and the test data set (this one split into a 100000 events public leaderboard data set and a 450000 private one) have been merged together, as well as a 18238 smaller data set held by the organizers, so that a total of 818238 events are available. It is now the responsibility of the user of this data set to adopt the cross validation method of its choice to avoid overtraining.

- For all events, the weights and labels are available. Weights are normalized such that the whole dataset corresponds to LHC 2012 running. Hence if a subset is defined for example for testing, one should renormalize the weights as follows, where y is the label (s or b), and $\mathbb{1}$ is the indicator function so that for example $\mathbb{1}\{y_i = s\}$ is one for signal events and zero for background events, and $\mathbb{1}\{y_i = b\}$ is zero for signal events and one for background events:

$$w'_j = w_j \frac{\sum_i w_i \mathbb{1}\{y_i = y_j\}}{\sum_{i \in subset} w_i \mathbb{1}\{y_i = y_j\}}. \tag{16}$$

  In words, the weight of signal events of the subset have to be scaled by the fraction of the sum of weights of signal events in the complete data set.

- two additional variables have been made available, **`KaggleSet`** and **`KaggleWeight`**, which allow to recover the original Kaggle training, public and private data set, see Appendix B for details, and to recompute the original public and private Kaggle leaderboard score for any submission.

6 Conclusion

In this note we provided background information for the Higgs boson machine learning challenge. All information for effectively participating in the Challenge is provided by the website mentioned in the preamble; pointers to additional resources are also available there. Any remaining questions should be asked on the forum hosted on the same web site.

Acknowledgments

We would like to thank the ATLAS experiment and the CERN organization for providing the simulated data for the Challenge, LAL-Orsay for serving as the official organizer, the Paris-Saclay

Center for Data Science, Google, and INRIA for providing financial assistance, Kaggle for hosting the Challenge, and the CERN organization again for the permanent post-Challenge hosting of the data set on opendata.cern.ch. BK was supported by the ANR-2010-COSI-002 grant of the French National Research Agency.

References

[1] G. Aad et al., "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys.Lett.*, vol. B716, pp. 1–29, 2012.

[2] S. Chatrchyan et al., "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Phys.Lett.*, vol. B716, pp. 30–61, 2012.

[3] The ATLAS Collaboration, "Evidence for higgs boson decays to tau+tau- final state with the atlas detector", Tech. Rep. ATLAS-CONF-2013-108, November 2013, http://cds.cern.ch/record/1632191.

[4] The ATLAS Collaboration, "Evidence for higgs-boson yukawa couplings in the $h \to \tau\tau$ decay mode with the atlas detector", *being submitted*, 2014.

[5] ATLAS Collaboration, "Dataset from the atlas higgs machine learning challenge 2014", *CERN Open Data Portal*, 2014, http://dx.doi.org/10.7483/OPENDATA.ATLAS.ZBP2.M5T8.

[6] S. Binet, B. Kegl, and D. Rousseau, "Software for the atlas higgs machine learning challenge 2014", *CERN Open Data Portal*, 2014, http://dx.doi.org/10.7483/OPENDATA.ATLAS.DFGK.DB9U.

[7] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyon, B. Kegl, and D. Rousseau, "Learning to discover: the higgs machine learning challenge 2014 - documentation", *CERN Open Data Portal*, 2014, http://dx.doi.org/10.7483/OPENDATA.ATLAS.MQ5J.GHXA.

[8] V. M. Abazov et al., "Observation of single top-quark production", *Physical Review Letters*, vol. 103, no. 9, 2009.

[9] Aaltonen, T. et. al, "Observation of electroweak single top-quark production", *Phys. Rev. Lett.*, vol. 103, pp. 092002, Aug 2009.

[10] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *The European Physical Journal C*, vol. 71, pp. 1554–1573, 2011.

[11] G. Aad et al., "A Particle Consistent with the Higgs Boson Observed with the ATLAS Detector at the Large Hadron Collider", *Science*, vol. 338, pp. 1576–1582, 2012.

[12] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses", *Annals of Mathematical Statistics*, vol. 9, pp. 60–62, 1938.

[13] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

A Special relativity

This appendix gives the very minimal introduction to special relativity for a better understanding of how the Higgs boson search is performed, and what the extracted features mean.

A.1 Momentum, mass, and energy

A fundamental equation of special relativity defines the so-called 4-momentum of a particle,

$$E^2 = p^2c^2 + m^2c^4, \tag{17}$$

where E is the energy of the particle, p is its momentum, m is the rest mass, and c is the speed of light. When the particle is at rest, its momentum is zero, and so Einstein's well-known equivalence between mass and energy, $E = mc^2$, applies. In particle physics, we usually use the following units: GeV for energy, GeV$/c$ for momentum, and GeV$/c^2$ for mass. 1 GeV (10^9 electron-Volt) is one billion times the energy acquired by an electron accelerated by a field of 1 V over 1 m, and it is also approximately the energy corresponding to the mass of a proton (more precisely, the mass of the proton is about 1 GeV$/c^2$). When these units are used, Eq. (17) simplifies to

$$E^2 = p^2 + m^2. \tag{18}$$

To avoid the clutter of writing GeV$/c$ for momentum and GeV$/c^2$ for mass, a shorthand of using GeV for all the three quantities of energy, momentum, and mass is usually adopted in most of the recent particle physics literature (including papers published by the ATLAS and the CMS experiments). We also adopt this convention throughout this document.

The momentum is related to the speed v of the particle. For a particle with non-zero mass, and when the speed of the particle is much smaller than the speed of light c, the momentum boils down to the classical formula $p = mv$. In special relativity, when the speed of the particle is comparable to c, we have $p = \gamma mv$, where

$$\gamma = \frac{1}{\sqrt{1 - (v/c)^2}}.$$

The relation holds both for the norms v and p and for the three dimensional vectors $\vec{v}$ and $\vec{p}$, that is, $\vec{p} = \gamma m\vec{v}$, where, by convention, $p = |\vec{p}|$ and $v = |\vec{v}|$. The factor γ diverges to infinity when v is close to c, and the speed of light cannot be reached nor surpassed. Hence, the momentum is a concept more frequently used than speed in particle physics. The kinematics of a particle is fully defined by the momentum and energy, more precisely, by the 4-momentum (p_x, p_y, p_z, E). When a particle is identified, it has a well defined mass[13], so its energy can be computed from the momentum and mass using Eq. (18). Conversely, the mass of a particle with known momentum and energy can be obtained from

$$m = \sqrt{E^2 - p^2}. \tag{19}$$

Instead of specifying the momentum coordinate (p_x, p_y, p_z), the parameters ϕ, η, and $p_T = \sqrt{p_x^2 + p_y^2}$, explained in Section 3.1, are often used.

A.2 Invariant mass

The mass of a particle is an intrinsic property of a particle. So for all events with a Higgs boson, the Higgs boson will have the same mass. To measure the mass of the Higgs boson, we need the 4-momentum $(p_x, p_y, p_z, E) = (\vec{p}, E)$ of its decay products. Take the simple case of the Higgs boson H decaying into a final state of two particles A and B which are measured in the detector.

[13] neglecting the particle width

By conservation of the energy and momentum (which are fundamental laws of nature), we can write $E_H = E_A + E_B$ and $\vec{p}_H = \vec{p}_A + \vec{p}_B$. Since the energies and momenta of A and B are measured in the detector, we can compute E_H and $p_H = |\vec{p}_H|$ and calculate $m_H = \sqrt{E_H^2 - p_H^2}$. This is called the *invariant mass* because (with a perfect detector) m_H remains the same even if E_H and p_H differ from event to event. This can be generalized to more than two particles in the final state and to any number of intermediate states.

In our case, the final state is a lepton, a hadronic tau, and three neutrinos. The lepton and hadronic tau are measured in the detector, but for the neutrinos, all we have is the transverse missing energy, which is an estimation of the sum of the momenta of the three neutrinos in the transverse plane (explained in Section 3). Hence the mass of the $\tau\tau$ can not be measured; we have to resort to different estimators which are only correlated to the mass of the $\tau\tau$. For example, the *visible mass* which is the invariant mass of the lepton and the hadronic tau, hence deliberately ignoring the unmeasured neutrinos.

A.3 Other useful formulas

The following formulas are useful to compute derived variables from primitives (in Appendix B). For `tau`, `lep`, `leading_jet`, and `subleading_jet`, the momentum vector can be computed as

$$\vec{p} = \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} = \begin{pmatrix} p_\mathrm{T} \times \cos\phi \\ p_\mathrm{T} \times \sin\phi \\ p_\mathrm{T} \times \sinh\eta \end{pmatrix},$$

where p_T is the transverse momentum, ϕ is the azimuth angle, η is the pseudo rapidity, and sinh is the hyperbolic sine function. The modulus of p is

$$p_\mathrm{T} \times \cosh\eta, \tag{20}$$

where cosh is the hyperbolic cosine function. The mass of these particles is neglected, so $E = p$.

The missing transverse energy $\vec{E}_\mathrm{T}^\mathrm{miss}$ is a two-dimensional vector

$$\vec{E}_\mathrm{T}^\mathrm{miss} = \begin{pmatrix} |\vec{E}_\mathrm{T}^\mathrm{miss}| \times \cos\phi_\mathrm{T} \\ |\vec{E}_\mathrm{T}^\mathrm{miss}| \times \sin\phi_\mathrm{T} \end{pmatrix},$$

where ϕ_T is the the azimuth angle of the missing transverse energy.

The invariant mass of two particles is the invariant mass of the 4-momentum sum, that is (still neglecting the mass of the two particles),

$$m_\mathrm{inv}(\vec{a},\vec{b}) = \sqrt{\left(\sqrt{a_x^2+a_y^2+a_z^2} + \sqrt{b_x^2+b_y^2+b_z^2}\right)^2 - (a_x+b_x)^2 - (a_y+b_y)^2 - (a_z+b_z)^2}. \tag{21}$$

The transverse mass of two particles is the invariant mass of the vector sum, the third component being set to zero, that is (still neglecting the mass of the two particles),

$$m_\mathrm{tr}(\vec{a},\vec{b}) = \sqrt{\left(\sqrt{a_x^2+a_y^2} + \sqrt{b_x^2+b_y^2}\right)^2 - (a_x+b_x)^2 - (a_y+b_y)^2}. \tag{22}$$

The pseudorapidity separation between two particles A and B is

$$|\eta_A - \eta_B|. \tag{23}$$

The R separation between two particles A and B is

$$\sqrt{(\eta_A-\eta_B)^2 + (\phi_A-\phi_B)^2}, \tag{24}$$

where $\phi_A - \phi_B$ is brought back to the $[-\pi, +\pi[$ range.

B The detailed description of the features

In this section we explain the list of features describing the events.

Prefix-less variables `EventId`, `Weight`, `Label`, `KaggleSet`, `KaggleWeight` have a special role and should not be used as input to the classifier. The variables prefixed with `PRI` (for PRImitives) are "raw" quantities about the bunch collision as measured by the detector, essentially the momenta of particles. Variables prefixed with `DER` (for DERived) are quantities computed from the primitive features. These quantities were selected by the physicists of ATLAS in the reference document [3] either to select regions of interest or as features for the Boosted Decision Trees used in this analysis. In addition:

- Variables are floating point unless specified otherwise.
- All azimuthal ϕ angles are in radian in the $[-\pi, +\pi[$ range.
- Energy, mass, momentum are all in GeV
- All other variables are unit less.
- Variables are indicated as "may be undefined" when it can happen that they are meaningless or cannot be computed; in this case, their value is -999.0, which is outside the normal range of all variables.
- The mass of particles has not been provided, as it can safely be neglected for the Challenge.

`EventId` An unique integer identifier of the event. Not to be used as a feature.

`DER_mass_MMC` The estimated mass m_H of the Higgs boson candidate, obtained through a probabilistic phase space integration (may be undefined if the topology of the event is too far from the expected topology)

`DER_mass_transverse_met_lep` The transverse mass (22) between the missing transverse energy and the lepton.

`DER_mass_vis` The invariant mass (21) of the hadronic tau and the lepton.

`DER_pt_h` The modulus (20) of the vector sum of the transverse momentum of the hadronic tau, the lepton, and the missing transverse energy vector.

`DER_deltaeta_jet_jet` The absolute value of the pseudorapidity separation (23) between the two jets (undefined if `PRI_jet_num` ≤ 1).

`DER_mass_jet_jet` The invariant mass (21) of the two jets (undefined if `PRI_jet_num` ≤ 1).

`DER_prodeta_jet_jet` The product of the pseudorapidities of the two jets (undefined if `PRI_jet_num` ≤ 1).

`DER_deltar_tau_lep` The R separation (24) between the hadronic tau and the lepton.

`DER_pt_tot` The modulus (20) of the vector sum of the missing transverse momenta and the transverse momenta of the hadronic tau, the lepton, the leading jet (if `PRI_jet_num` ≥ 1) and the subleading jet (if `PRI_jet_num` $= 2$) (but not of any additional jets).

`DER_sum_pt` The sum of the moduli (20) of the transverse momenta of the hadronic tau, the lepton, the leading jet (if `PRI_jet_num` ≥ 1) and the subleading jet (if `PRI_jet_num` $= 2$) and the other jets (if `PRI_jet_num` $= 3$).

`DER_pt_ratio_lep_tau` The ratio of the transverse momenta of the lepton and the hadronic tau.

DER_met_phi_centrality The centrality of the azimuthal angle of the missing transverse energy vector w.r.t. the hadronic tau and the lepton

$$C = \frac{A+B}{\sqrt{A^2+B^2}},$$

where $A = \sin(\phi_{\text{met}} - \phi_{\text{lep}}) * \text{sign}(\sin(\phi_{\text{had}} - \phi_{\text{lep}}))$, $B = \sin(\phi_{\text{had}} - \phi_{\text{met}}) * \text{sign}(\sin(\phi_{\text{had}} - \phi_{\text{lep}}))$, and ϕ_{met}, ϕ_{lep}, and ϕ_{had} are the azimuthal angles of the missing transverse energy vector, the lepton, and the hadronic tau, respectively. The centrality is $\sqrt{2}$ if the missing transverse energy vector $\vec{E}_{\text{T}}^{\text{miss}}$ is on the bisector of the transverse momenta of the lepton and the hadronic tau. It decreases to 1 if $\vec{E}_{\text{T}}^{\text{miss}}$ is collinear with one of these vectors and it decreases further to $-\sqrt{2}$ when $\vec{E}_{\text{T}}^{\text{miss}}$ is exactly opposite to the bisector.

DER_lep_eta_centrality The centrality of the pseudorapidity of the lepton w.r.t. the two jets (undefined if `PRI_jet_num` ≤ 1)

$$\exp\left[\frac{-4}{(\eta_1-\eta_2)^2}\left(\eta_{\text{lep}} - \frac{\eta_1+\eta_2}{2}\right)^2\right],$$

where η_{lep} is the pseudorapidity of the lepton and η_1 and η_2 are the pseudorapidities of the two jets. The centrality is 1 when the lepton is on the bisector of the two jets, decreases to $1/e$ when it is collinear to one of the jets, and decreases further to zero at infinity.

PRI_tau_pt The transverse momentum $\sqrt{p_x^2+p_y^2}$ of the hadronic tau.

PRI_tau_eta The pseudorapidity η of the hadronic tau.

PRI_tau_phi The azimuth angle ϕ of the hadronic tau.

PRI_lep_pt The transverse momentum $\sqrt{p_x^2+p_y^2}$ of the lepton (electron or muon).

PRI_lep_eta The pseudorapidity η of the lepton.

PRI_lep_phi The azimuth angle ϕ of the lepton.

PRI_met The missing transverse energy $\vec{E}_{\text{T}}^{\text{miss}}$.

PRI_met_phi The azimuth angle ϕ of the missing transverse energy.

PRI_met_sumet The total transverse energy in the detector.

PRI_jet_num The number of jets (integer with value of 0, 1, 2 or 3; possible larger values have been capped at 3).

PRI_jet_leading_pt The transverse momentum $\sqrt{p_x^2+p_y^2}$ of the leading jet, that is the jet with largest transverse momentum (undefined if `PRI_jet_num` $= 0$).

PRI_jet_leading_eta The pseudorapidity η of the leading jet (undefined if `PRI_jet_num` $= 0$).

PRI_jet_leading_phi The azimuth angle ϕ of the leading jet (undefined if `PRI_jet_num` $= 0$).

PRI_jet_subleading_pt The transverse momentum $\sqrt{p_x^2+p_y^2}$ of the leading jet, that is, the jet with second largest transverse momentum (undefined if `PRI_jet_num` ≤ 1).

PRI_jet_subleading_eta The pseudorapidity η of the subleading jet (undefined if `PRI_jet_num` $\leq$ 1).

PRI_jet_subleading_phi The azimuth angle ϕ of the subleading jet (undefined if PRI_jet_num $\leq$ 1).

PRI_jet_all_pt The scalar sum of the transverse momentum of all the jets of the events.

Weight The event weight w_i, explained in Section 3.3. Not to be used as a feature. Not available in the Kaggle test sample, but available for all events in the open data.cern.ch dataset

Label The event label (string) $y_i \in \{s, b\}$ (s for signal, b for background). Not to be used as a feature. Not available in the test sample.

KaggleSet Specific to opendata.cern.ch dataset : string specifying to which Kaggle set the event belongs : "t":training, "b":public leaderboard, "v":private leaderboard,"u":unused.

KaggleWeight Specific to opendata.cern.ch dataset : weight normalized within each Kaggle data set according to equation 16.

The events (instances) and the features were used for the training and optimization of the reference ATLAS analysis [3]. However, both the feature list and the events have been simplified for the Challenge in the following way.

- The top sample normally has events with negative weights. These have been removed.
- Only major background sources are included.
- The normalization of the signal and backgrounds (captured in weight) is slightly altered, because correction factors used in the reference analysis [3] have not been applied.
- In the reference analysis [3], manipulated data events are used eventually to evaluate the different backgrounds.

These simplifications allowed us to provide a large sample for possible sophisticated separation algorithms and to provide a relatively simple optimization criterion, while preserving the complexity of the original classification problem. The reference ATLAS analysis can be reproduced reasonably closely (although not exactly) with the provided data.