JMLR: Workshop and Conference Proceedings 42:19-55, 2015 HEPML 2014

The Higgs boson machine learning challenge

Claire Adam-Bourdarios claire.bourdarios@cern.ch
LAL, IN2P3/CNRS & University Paris-Sud, France

Glen Cowan G.Cowan@rhul.ac.uk
Physics Department, Royal Holloway, University of London, UK

Cécile Germain Cecile.Germain@lri.fr
LRI, University Paris-Sud & CNRS & INRIA, France

Isabelle Guyon guyon@chalearn.org
ChaLearn and ClopiNet, Berkeley, California, USA

Balázs Kégl kegl@lal.in2p3.fr
LAL, IN2P3/CNRS & TAO team, INRIA & LRI, CNRS && University Paris-Sud, France

David Rousseau rousseau@lal.in2p3.fr
LAL, IN2P3/CNRS & University Paris-Sud, France

Editor: Cowan, Germain, Guyon, Kégl, Rousseau

Abstract

The Higgs Boson Machine Learning Challenge (HiggsML or the Challenge for short) was organized to promote collaboration between high energy physicists and data scientists. The ATLAS experiment at CERN provided simulated data that has been used by physicists in a search for the Higgs boson. The Challenge was organized by a small group of ATLAS physicists and data scientists. It was hosted by Kaggle at https://www.kaggle.com/c/higgs-boson; the challenge data is now available on http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014. This paper provides the physics background and explains the challenge setting, the challenge design, and analyzes its results.

Keywords: high energy physics, Higgs boson, statistical tests, machine learning.

1. Introduction

This paper reports results of a unique experiment in high energy physics: using the power of the "crowd" to help solving difficult physics problems. This experiment took the form of a data science challenge organized in 2014: the Higgs Boson Machine Learning Challenge (HiggsML) [1]. Its results were discussed at a workshop of the Neural Information Processing Systems conference (NIPS 2014).

For the first time, the ATLAS experiment at CERN publicly released a portion of the simulated data used by physicists to optimize the analysis of the Higgs boson. The challenge was organized by an interdisciplinary team of ATLAS physicists and computer scientists, LAL (Université Paris Sud and CNRS / IN2P3), LRI (Université Paris Sud and CNRS, INRIA, Royal Holloway University of London and ChaLearn, a non-profit group dedicated to the organization of challenges in Machine Learning.

An essential component of the analysis of the experimental data is a procedure for the selection of a region of interest in the space of measured features, i.e., the variables for each particle collision or "event". In the past, the region of interest was designed by human expertise (naïve-Bayes-like "cut-based" techniques). Today, multivariate classification techniques are routinely used to optimize the selection region (The ATLAS Collaboration, 2013; V. M. Abazov et al., 2009; Aaltonen, T. et. al, 2009). The immediate goal of the Challenge was to *explore the potential of advanced classification methods to improve the statistical significance of the experiment.* In addition, the Challenge promoted collaboration between high energy physics and machine learning.

No knowledge of particle physics was necessary to participate. The Challenge, posted on Kaggle, attracted an unprecedented number of participants over a short period of time (May 12, 2014 to Sept 15, 2014) and it was one of the most popular Kaggle competitions. The software developed by the participants was made available freely, and the data were released at http://opendata.cern.ch/ after the end of the Challenge.

The success of the Challenge can be attributed in part to the choice of the subject (the Higgs boson discovery), which is of high interest to the public, and to the support of CERN. Also important were design choices, including the simplification of the problem setting, which allowed us to reach both computer scientists and physicists. We also stimulated participation by providing a starting kit, responding promptly to questions in the online forum, where participants were also helping each other, and through wide advertising. An additional incentive was provided in the form of prizes for the three winners and an invitation to visit CERN to discuss their results with high-energy physicists.

The outcomes of the Challenge have been interesting to high-energy physicists in several respects. The winning method of Gábor Melis, which used an ensemble of deep neural networks, was computationally expensive but performed better than the runner-up by a statistically significant margin. The success of the winner is attributable in part to the very careful way in which he conducted cross-validation to avoid overfitting. A special "High Energy Physics meets Machine Learning" award was given to team Crowwork (Tianqi Chen and Tong He). They had a slightly lower score but provided a Boosted Decision Tree method that is a good compromise between performance and simplicity, which could improve tools currently used in high-energy physics.

[1] https://www.kaggle.com/c/higgs-boson

Given the broad success of the Challenge and the wishes of many participants and others to pursue the study beyond the formal end of the competition, the dataset consisting of 800,000 simulated events provided by the official ATLAS simulator, has been made permanently available[2], together with accompanying software (Binet et al., 2014) and documentation (Adam-Bourdarios et al., 2014).

The rest of the paper analyzes the Challenge settings and results. We start with a description of the scientific goals of the Challenge. Readers only interested in a brief overview of the problem can direct their attention to Section 2. This section presents an introduction to the relations between the supervised learning context and the *learning to discover* question of physics. Section 3 provides a more detailed description of the underlying physics, and Section 4 elaborates on the formal settings of the statistical learning setup. Then, we proceed to the analysis of the Challenge. Section 5 describes its organization, the datasets and the features. Section 6 presents the results and the lessons learned. Some open questions are discussed in Section 7 before the conclusion.

2. Learning to discover

Viewed in a simplified way, the analysis leading to the discovery of a new particle starts with the acquisition of experimental data; then a classifier is used to determine whether events of interest (signal) occurred in addition to background events; finally a statistical test is applied to see if one can reject the hypothesis that the event sample contains only background events. If the probability, assuming background only, of finding data as indicative of signal or more so falls below a given limit, then the background-only hypothesis is rejected and the new particle is deemed to be discovered. This section summarizes each of these steps and discusses their relations with the supervised learning framework.

2.1. Physics motivation

The ATLAS experiment and the CMS experiment recently claimed the discovery of the Higgs boson (Aad et al., 2012a; Chatrchyan et al., 2012). The discovery was acknowledged by the 2013 Nobel prize in physics given to François Englert and Peter Higgs. This particle was predicted to exist theoretically almost 50 years ago as part of the mechanism by which other elementary particles have mass. Its importance is considerable because it is the final ingredient of the Standard Model of particle physics, ruling subatomic particles and forces. Without confirmation of its existence, the fundamental principle on which our current Standard Model of elementary particles is based would collapse. The discovery relies on experiments being carried out 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 to run for at least the next 10 years.

The Higgs boson has many different processes through which it can disintegrate or *decay*. Beyond the initial discovery, the study of all modes of decay increases confidence in the validity of the theory and helps characterize the new particle. When a particle decays, it

[2] http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014 (ATLAS Collaboration, 2014)

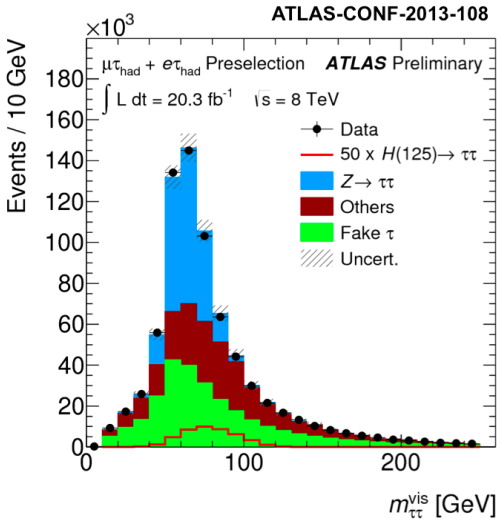


Figure 1: Distribution of the visible mass of Higgs to tau tau events at preselection stage The ATLAS Collaboration (2013). The colored stacked histograms show the estimated contribution of different background processes. The red line shows the scaled expected contribution of the signal.

produces other particles, and these are classified as being of one of two fundamental types: *fermions* or *bosons*, which differ in their amount of intrinsic angular momentum or "spin".

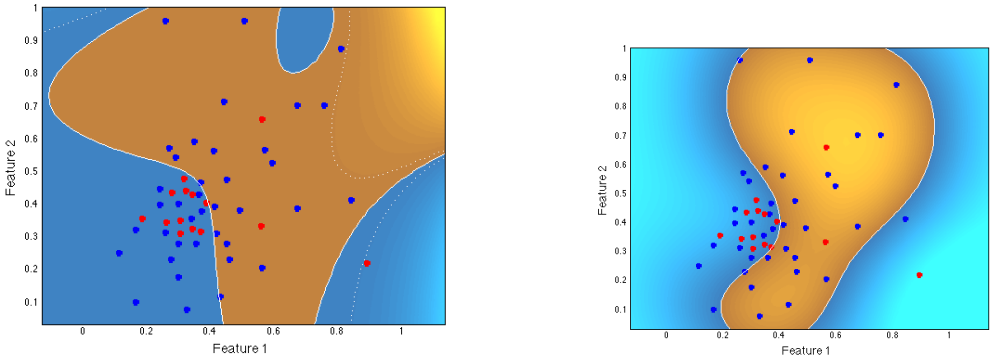
The decay into specific particles is called a *channel* by physicists. The Higgs boson was first seen in three distinct decay channels which are all boson pairs. One of the next important topics is to seek evidence on the decay into fermion pairs, namely tau-leptons or b-quarks, and to precisely measure their characteristics. The subject of the Challenge was to study the H to tau tau channel. The first evidence of the H to tau tau channel was recently reported by the ATLAS experiment (The ATLAS Collaboration, 2013). We refer to this paper as the "reference document"[3].

Figure 1 shows why the problem is difficult. The expected signal has a broad distribution which is masked by much more abundant backgrounds, in particular the Z to tau tau decay which produces a very large peak at a slightly lower mass.

2.2. Classification for selection

From the machine learning point of view the problem can be formally cast into a *binary classification problem*. Events generated in the collider are preprocessed and represented as a *feature vector*. The problem is to classify events as *signal* (that is, an event of interest, in our case a H to tau tau decay) or *background* (an event produced by already known processes). More precisely, the classifier is used as a *selection method*, which defines a

[3] It should be noted the final ATLAS paper (Aad et al., 2015) on the subject is now published, but changes with respect to the reference document are not relevant to the Challenge



(*a*) Polynomial fit (*b*) Radial basis function fit

Figure 2: Toy example of classification/selection. We represent signal events as red dots and background events as blue dots. (a) and (b) show decision boundaries obtained with two different classifiers. Points in the turquoise zone are classified as signals while points in the orange zone are classified as background.

(not necessarily connected) signal-rich region in the feature space. Figure 2 provides a toy example in an arbitrary two-dimensional feature space (the challenge data used 30 features).

In the rest of the paper, the following terminology will be used.

- *Event*: an elementary record. In classification vocabulary, an event is an instance.
- *Signal event*: an event in which a Higgs boson decays to a pair of tau leptons. In classification vocabulary, a signal event is a member of the positive class.
- *Background event*: any event other than the signal type. In classification vocabulary, a background event is a member of the negative class.
- *Selected event*: an event that a selection apparatus deems a candidate for being signal. In classification vocabulary, a selected event is a predicted positive.
- *Selected background*: an non-signal event that, because of the physical statistical fluctuations, has properties close to those of signal. In classification vocabulary, a selected background event is a false positive.
- *Selected signal*: In classification vocabulary, a true positive.

Because the problem is the discovery of a new phenomenon, labeled examples of actual signal events in the real data are not available. Rather, events are simulated using an elaborate simulator that artificially generates events following the Standard Model and a model of the detector, taking into account noise and possible artifacts. The classifiers are evaluated using these extensive simulations. The goal of the Challenge was to propose new and better classifiers.

From the machine learning point of view, the problem presents several difficulties.

- In the real data, the classes are very imbalanced (approximately two signal events in a thousand events[4] after preselection); for this reason the simulated data provided in the Challenge are enriched in signal events. To compensate for this bias, all events are weighted with importance weights reflecting their probability of occurrence.
- The classes completely overlap; in fact the background class is a big "blob" inside which the signal class is a small "blip".
- The objective, called the *Approximate Median Significance* (AMS) is unusual.
- Though the number of training examples is relatively large (250,000), the AMS depends only on the number of selected events which is much smaller, and so it is prone to overfitting.

Most classification methods calculate a discriminant value $f(x)$, which is a score taking small values for the negative class (background events) and large values for the positive class (signal events). By putting a threshold on the discriminant value, classification decisions can be reached: an event is selected if the discriminant value is larger then the threshold and predicted negative otherwise.

Figure 3 illustrates the class overlap issue based on the discriminant function only, computed with the MultiBoost classifier that was provided as a starting kit for the Challenge. Figure 3(a) shows the classwise score distributions: the classes are separated quite neatly. However, when the distributions are normalized to their prior probabilities (Figure fig 3(b)), the signal is dominated by the background (every individual event is more likely to be background than signal). This suggests that classification accuracy is a very poor measure of success in this case. Indeed, the AMS is not a function of classification accuracy.

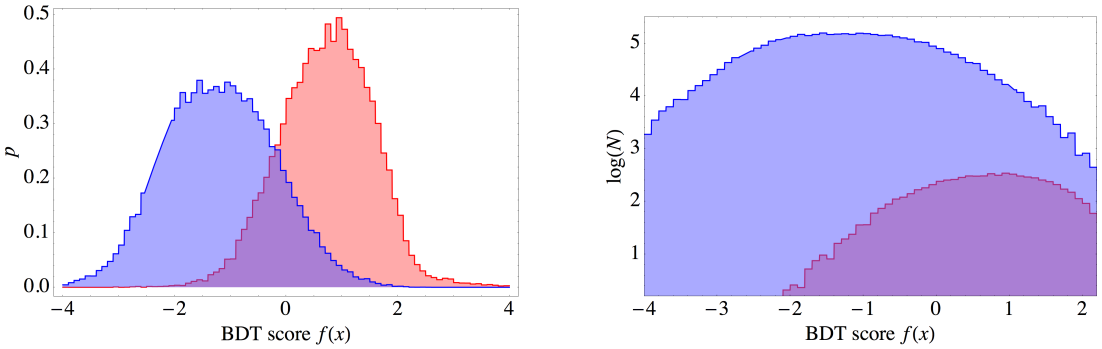
Section 2.3 motivates the AMS in the framework of discovering the existence of the signal process and Section 2.4 discusses it in the context of classical ML objective functions.

2.3. Discovery and the AMS

The Approximate Median Significance (AMS) that we define below is an objective function used to determine a region in the feature space where one expects an enhanced number of signal events. In the real experiment, one simply counts the number of events n found in this region. This value will follow a Poisson distribution with a mean of $s+b$, where s and b are the mean numbers of events from signal and background processes, respectively. If n is found much greater than b, then the background-only hypothesis is rejected. This statistical test is quantified using the p-value of the background-only hypothesis or equivalently through its significance, as described in more detail in Sec. 4.3.

For purposes of planning the statistical analysis, one would like to maximize the discovery significance (significance with which one rejects the background-only hypothesis) that is to be expected *if the signal process is present*. The general approach for such a maximization, initially described by (Dempster and Schatzoff, 1965) and (Joiner, 1969) in a purely statistical framework, is that the discovery significance itself has a certain sampling distribution. In our case, assuming the presence of signal, it is approximately Gaussian.

[4] DR check sum of weights of signal over total sum of weights



(*a*) Unweighted class distribution

(*b*) Weighted class distribution. Note the log scale on the vertical axis

Figure 3: Background (blue) and signal (red) distributions of the BDT discriminant scores. The classifier is AdaBoost with $T = 3000$ trees of $N = 2$ inner nodes (three leaves) each. Both T and N were optimized using 10-fold cross validation.

For technical reasons, it is convenient to work with the median rather than the mean significance, and therefore we take "expected" significance here to refer to the median.

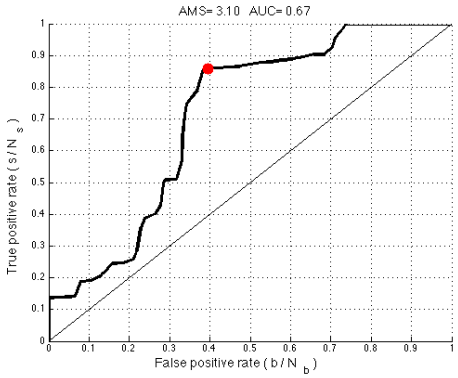
The objective function we present below is an approximation for this median significance (the AMS). In the following, we derive a simple approximation (AMS_3) of the AMS that is useful for understanding the basic principles. More precise approximations of the AMS including the one that was used in the Challenge are described in Section 4.

Assume that the occurrence of background events is a Poisson process (in the original feature space as well as in any selection region). Over a given time window during which events are recorded, if the expected number of selected background events is μ_b, its variance is also μ_b (since Poisson distributed variables have an equal mean and variance). Approximating the Poisson distribution with the Normal law, the statistic $\text{AMS}_3 = (n - \mu_\text{b})/\sqrt{\mu_\text{b}}$ (n being the number of events ending up in the selection region), distributed according to $N(0, 1)$, can serve as test statistic for anomalous fluctuations (detection of signal events). A fluctuation is considered sufficiently large to claim discovery of the signal process if it exceeds 5 sigma, that is if $\text{AMS}_3 > 5$, which corresponds to a p-value of the one-sided Z-test of 3×10^{-7}.

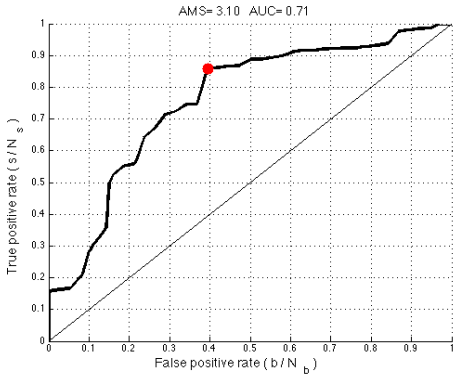
We distinguish two use cases for $\text{AMS}_3 = (n - \mu_\text{b})/\sqrt{\mu_\text{b}}$:

- *As test statistic:* At utilization time, to conduct discovery tests with real (unlabeled) events, μ_b is a constant determined by theoretical considerations. A large deviation of n (the number of events detected in the selection region) from μ_b (at least 5 sigma) indicates a discovery.

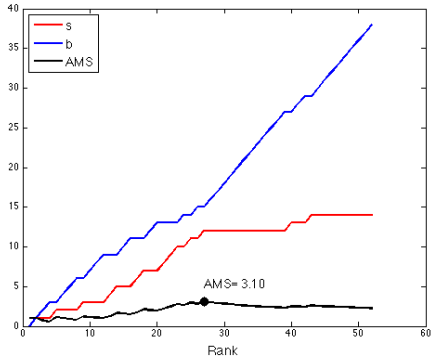
- *As objective function and test performance metric:* For the purpose of the Challenge (for training and testing using labeled simulated data, which include both signal and background events), we estimate n by $s + b$ and approximate μ_b by b. In this way, the test statistic becomes $\text{AMS}_3 = s/\sqrt{b}$. The goal of the challenge is to improve the



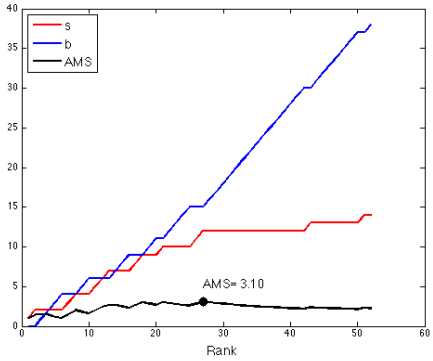
(*a*) ROC curve for polynomial



(*b*) ROC curve for radial



(*c*) AMS for polynomial



(*d*) AMS for radial

Figure 4: (a) and (b) show the ROC curves, obtained by varying the threshold. The red dot indicates the position of the AMS obtained when the threshold is optimized. (c) and (d) show the AMS curve as a function of the rank of the example, when examples are ordered by the discriminant value of the classifier, and the value of the optimum AMS. All results are computed on the training data.

significance of the discovery test by training a classifier that optimizes the selection region. This is achieved by maximizing the AMS on the training data with respect to the parameters and hyper parameters of the classifier. The performance of the classifier is then assessed with the AMS computed on test data.

2.4. Interpreting the AMS objective function

Optimizing for the AMS has two components, ranking and threshold selection. The discriminant function produces a ranking of all events in order of decreasing score from signal to background. It is well known (Dempster and Schatzoff, 1965; Clémençon et al., 2005) that ranking is equivalent to an AUC (Area Under ROC) optimization problem. Thus, the role of the AMS is related to the role of the ROC or the precision-recall curve in "learning to

rank" problems. In fact, the idea is rather close: in a ranked set of events (candidate signal events coming first) we are interested in estimating with confidence the fraction of falsely discovered events in the top ranking events. For instance, in bioinformatics, the significance of a discovery is evaluated with a statistical test assessing whether the false discovery rate is small enough. The AMS is a test statistic similar in spirit to the false discovery rate.

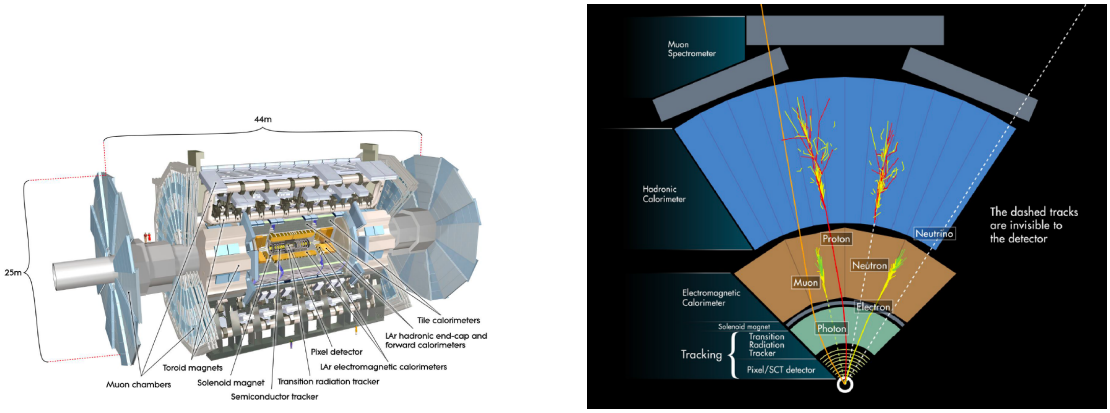
However, our toy example reveals that optimizing for the AUC and the AMS is not equivalent. Figure 4 shows how the AMS and the area under the ROC (AUC) curve can differ. In the example, the classifier is a kernel ridge regression classifier with either a polynomial or a radial basis kernel. The AMS is optimized with respect to the bias value only, for simplicity. We see that the same value of the optimal AMS is achieved in both cases (Figures 4(c) and 4(d)) for different AUC values (Figures 4(c) and 4(d)). This is not surprising since the AUC integrates over all values of the bias (or threshold) while the AMS considers a single (optimal) point. Conversely two solutions may have identical AUC values and different AMS.

A comparison of the AMS and the classical weighted accuracy objective, on the Challenge dataset, is deferred to Section 4, where the necessary formalization is introduced.

3. Physics motivation

This section elaborates on some of the details on the detector and the analysis. Understanding these was not necessary to participate in the Challenge, but it helps understanding both the importance of the challenge and the difficulty of the learning problem. For further information, we refer the reader to the 21 December 2012 special edition of Science Magazine "Breakthrough of the Year : the Higgs Boson"[5], in particular, a non-specialist account of the discovery of the Higgs boson (Aad et al., 2012b).

3.1. Proton collisions and detection



(a) The Atlas detector (b) An event

Figure 5: The Atlas experiment

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

The LHC collides bunches of protons every 50 nanoseconds within each of its four experiments. Each bunch crossing results in a random number of proton-proton collisions (with a Poisson expectation between 10 and 35, depending on the LHC conditions) called events[6].

Two colliding protons produce a small "explosion" 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 (Figure 5(*a*)) 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 (Figure 5(*b*). From these quantities, the properties of the decayed parent particle is inferred, and the inference chain is continued until reaching the heaviest primary particles.

An online trigger system discards the vast majority of bunch collisions, which contain 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 taus, jets, and missing transverse energy, which are explained below. Electrons, muons, and taus are the three leptons[7] 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 collimated 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.

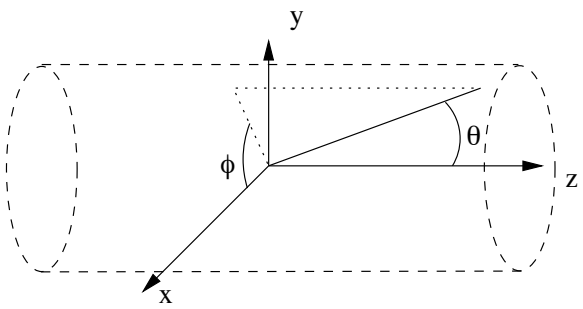


Figure 6: ATLAS reference frame

[6] 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.

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

We will use the conventional 3D direct reference frame of ATLAS throughout the document (Figure 6): 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 onto 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 around 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 plane perpendicular to the z axis.

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 jets were rejected, which helps to reject some of the background sources[8].

3.2. The physics goal

In the Challenge, the positive (signal) class consists of events in which the Higgs boson decays into two taus. This channel is interesting from a theoretical point of view but experimentally it is very challenging. In the Standard Model (SM), the Higgs boson is the particle which is responsible for the mass of the other elementary particles. To test 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 whether the results are consistent with the predictions of the SM.

In the original discovery, the Higgs boson was seen decaying into $\gamma\gamma$, WW, and ZZ, which are all boson pairs (Aad et al., 2012a; Chatrchyan et al., 2012) (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

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

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. It is therefore important to measure as precisely as possible the coupling of the Higgs to the charged leptons (electron, muon, tau). The coupling of the Higgs to electrons and muons is beyond the reach of the ATLAS experiment due to their small masses, leaving the measurement using tau leptons as the only realistic possibility.

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 into two taus, and one expects far more tau pairs from events of this type than from Higgs decays. 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.

In the analysis considered here, we focus 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.

3.3. The challenge data set

For the Challenge, we provide simulated events using the official ATLAS full detector simulator. The simulator yields simulated events with properties that mimic the statistical properties of the real events of the signal type as well as several important backgrounds.

The signal sample contains events in which Higgs bosons (with fixed mass 125 GeV) were produced. The background sample contains events corresponding to 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 products. 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.

4. The formal model

4.1. The learning problem

For the formal description of the Challenge, let $\mathcal{D} = \big\{(\mathbf{x}_1, y_1, w_1), \ldots, (\mathbf{x}_n, y_n, w_n)\big\}$ 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[9] .

There are two properties that make our simulated set different from those collected in nature or sampled in a natural way from a joint distribution $p(\mathbf{x}, y)$.[10] First, we can simulate as many events of the signal class 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 $P(y = \mathrm{s})/P(y = \mathrm{b})$. As explained in Section 3, 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, the simulator produces 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) is 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 numbers* 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,

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

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

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

[9] We use roman s to denote the label and in indices of terms related to signal (e.g., n_{s}), and s (3) for the *estimated* number of signal events selected by a classifier. The same convention applies to the terms related to background.

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

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, i.e.,

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

In machine learning terminology, s and b are are *unnormalized*, or more precisely, *luminosity-normalized* (1) true and false positive rates.

Given a classifier g, the AMS objective function used for the Challenge (AMS_c) is defined by

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

where s and b are defined in Eqs. (3) and (5), respectively, and b_{reg} is a regularization term set to a constant $b_{\mathrm{reg}} = 10$ in the Challenge. The derivation of this formula is explained in Section 4.3.

In summary, the task of the participants was to train a classifier g based on the training data $\mathcal{D}$ with the goal of maximizing the AMS_c (7) on a held-out (test) data set.

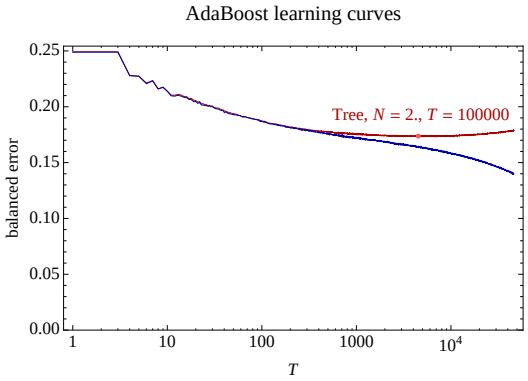
4.2. Optimizing for the AMS

Comparing N_{b} and N_{s} to the number of simulated signal and background events n_{s} and n_{b}, it is clear that the typical weight of a signal event in the training sample is about 300 times smaller than the typical weight of a background event. Training a classifier with this original weight distribution would mean that the optimization would concentrate on the extreme end of the ROC curve with no false positives, and would mostly ignore misclassified signal events. The objective is not the (weighted) classification accuracy but the AMS (7), and running some preliminary experiments, it is clear that the AMS is *not* optimized in this extreme corner of the ROC curve. Taking this into consideration, the usual approach adopted by the high-energy physics community consists of two steps:

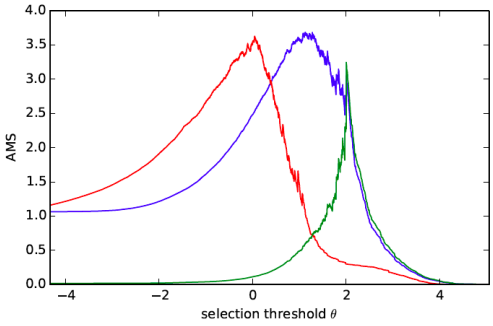
1. Training a real-valued discriminant function $f : \mathbb{R}^d \to \mathbb{R}$ to minimize the *balanced* classification error

$$R(f) = \sum_{i=1}^{n} w'_i \mathbb{I}\left\{ \mathrm{sign}\big(f(\mathbf{x}_i)\big) \neq y_i \right\},^{11} \tag{8}$$

[11] The indicator function $\mathbb{I}\{A\}$ is 1 if its argument A is true and 0 otherwise. From now on, to simplify notation, we replace the label s by $y = +1$ and the label b by $y = -1$.



(a) Learning curves (blue: training error; red: test error) of AdaBoost with trees of three leaves (BDT)



(b) The AMS versus the selection threshold with balanced weighting (red) and calibrated weighting (blue) . The green curve is the refined AMS_1 (20) with $\sigma_\text{b} = 0.1b$.

Figure 7: Optimizing for the AMS

where the weights w'_i are normalized in both the signal and background classes to $N'_\text{b} = N'_\text{s} = 0.5$, that is,

$$w'_i = w_i \times \begin{cases} \frac{1}{2N_\text{s}} & \text{if } i \in \mathcal{S} \\ \frac{1}{2N_\text{b}} & \text{if } i \in \mathcal{B}. \end{cases} \tag{9}$$

2. Optimizing the AMS with respect to the selection threshold θ in the the classifier $g(\mathbf{x}) = \text{sign}\big(f(\mathbf{x}) - \theta\big)$ on a held-out validation set using the original weights.

To illustrate this procedure, we walk the reader through the details of the MultiBoost benchmark. We first trained AdaBoost with $T = 3000$ trees of $N = 2$ inner nodes (three leaves) each on 90% of the training set using balanced weighting $N'_\text{b} = N'_\text{s}$ (9). Both T and N were optimized using 10-fold cross validation. The balanced test error rate (8) of the resulting classifier is about 17.5%. Figure 7 shows the learning curves and the classwise score distributions.

We then optimized the AMS with respect to the selection threshold on the remaining 10% of the training set. Figure 7(*b*) shows the AMS curve (blue). It is maximized at $\theta = 1.28$ indicating that the selection region $\mathcal{G} = \{\mathbf{x} : f(\mathbf{x}) > \theta\}$ is a small subset of the positive region $\{\mathbf{x} : f(\mathbf{x}) > 0\}$ defined by the balanced classifier $\text{sign}(f(\mathbf{x}))$. The region contains about 3600 signal points and 250 background points, with sum-of-weights $s = 229$ and $b = 3688$, and $\text{AMS} \approx s/\sqrt{b} = 3.69$.

From Figure 7(*b*) is is also clear that, if the goal is to maximize the AMS, equal weighting is not optimal: the AMS is maximized at a classification threshold that implies a much larger false negative rate than false positive rate. A plausible (although, for the time being, heuristic) way to improve the match between the two objectives (the AMS (7) and the weighted classification error (8)) is to reweight the data in a such way that the maximum

of the AMS is taken at approximately $f(x) = 0$. However, this approach did not provide a significant improvement.

4.3. The statistical setup

This section describes the basic structure of the statistical framework that leads to the criterion (7). The derivation is based on (Cowan et al., 2011).

In Section 4.3.1 we describe the statistical test and the discovery significance computed on real data. In Section 4.3.2 we derive the approximate median significance that can be estimated using simulated events[12], which is used to optimize the selection procedure, and we explain why we apply a regularization term for not letting b approach zero.

4.3.1. Statistical treatment of the measurement

Each proton-proton collision or "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 number of events n found in $\mathcal{G}$ is assumed to follow a Poisson distribution with mean $\mu_\mathrm{s} + \mu_\mathrm{b}$,

$$P(n|\mu_\mathrm{s}, \mu_\mathrm{b}) = \frac{(\mu_\mathrm{s} + \mu_\mathrm{b})^n}{n!} e^{-(\mu_\mathrm{s}+\mu_\mathrm{b})}, \tag{10}$$

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 $\mu_\mathrm{s} = 0$ (the background-only hypothesis) against the alternative where the signal exists, that is, $\mu_\mathrm{s} > 0$. From Eq. (10), we construct the likelihood ratio

$$\lambda = \frac{P(n|0, \mu_\mathrm{b})}{P(n|\hat{\mu}_\mathrm{s}, \mu_\mathrm{b})} = \left(\frac{\mu_\mathrm{b}}{\hat{\mu}_\mathrm{s} + \mu_\mathrm{b}}\right)^n e^{\hat{\mu}_\mathrm{s}} = \left(\frac{\mu_\mathrm{b}}{n}\right)^n e^{n-\mu_\mathrm{b}}, \tag{11}$$

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

According to Wilks' theorem (Wilks, 1938), given that certain regularity conditions are satisfied, the test statistic

$$q_0 = \begin{cases} -2 \ln \lambda & \text{if } n > \mu_\mathrm{b}, \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

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., (Cowan et al., 2011)). Assuming that these hold, the p-value of the background-only hypothesis from an observed value of q_0 is found to be

$$p = 1 - \Phi\left(\sqrt{q_0}\right), \tag{13}$$

[12] 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.

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

$$Z = \Phi^{-1}(1-p), \tag{14}$$

where Φ^{-1} is the standard normal quantile. Eqs. (13) and (14) lead therefore to the simple result

$$Z = \sqrt{q_0} = \sqrt{2\left(n \ln\left(\frac{n}{\mu_\mathrm{b}}\right) - n + \mu_\mathrm{b}\right)} \tag{15}$$

if $n > \mu_\mathrm{b}$ 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}.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.

4.3.2. The median discovery significance

Equation (15) 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 $\mathcal{G} = \{\mathbf{x} : g(\mathbf{x}) = \mathrm{s}\}$, we do not know n and μ_b. As usual in empirical risk minimization (Devroye et al., 1996), 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

$$\mathrm{AMS}_2 = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)}. \tag{16}$$

AMS_2 can be rewritten as

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

where

$$\mathrm{AMS}_3 = \frac{s}{\sqrt{b}}. \tag{17}$$

The two criteria Eqs. (16) and (17) 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.

Our preliminary MultiBoost benchmark revealed a potential pitfall. 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 have made it difficult to compare reliably the participants of the Challenge if the optimal region was small. Therefore, 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. Challenge organization

5.1. The datasets

The datasets are lists of events (instances). The same events and features were used for the training and optimization of the reference ATLAS analysis (The ATLAS Collaboration, 2013). Both the feature list and the events have been simplified for the Challenge, but so slightly that the reference ATLAS analysis can be reproduced reasonably closely (although not exactly) with them. Using the reference ATLAS data had two advantages. Most importantly, best relevance to the physics application and community; and technically, preventing the risk of leakage *i.e.,* exposing inadvertently data that should not be available for modeling, which is not uncommon (Rosset et al., 2010).

The events were allocated to the training set (250K events), the validation set (100K events used in the public leaderboard) and the test set (450K events used in the private leaderboard). Special care has been devoted to making the three sets identically distributed. Around one third of the events in each set were signals.

The weights were provided in the training set so that the AMS_c (7) could be properly evaluated. Weights were not provided in the test set since the weights distribution of the signal and background sets are so different that they would have given away the label immediately.

The proper normalization of weights repetitively raised many questions[13] at the beginning of the competition, and numerous postings from the organizers were needed to make clear that, when a subset of the training set is considered (*e.g.,* for cross-validation), the weights should be renormalized along Eq. (1). The precise formula is given in Appendix C.

5.2. The features

This section highlights the most important characteristics of the $d = 30$ features. They are described individually in Appendix B.

The features prefixed with PRI (for PRImitives) are "raw" quantities about the bunch collision as measured by the detector, essentially the momenta of particles. Those prefixed with DER (for DERived) are quantities computed from the primitive features. These quantities were defined by the ATLAS physicists in the reference document (The ATLAS Collaboration, 2013) either to select regions of interest or as features for the Boosted Decision Trees used in this analysis. For all but one, the computation is simply an algebraic formula (see Appendix A.1). As detailed in A.2, the invariant mass (DER_mass_MMC) feature is estimated. Are these features *necessary* and/or *sufficient*? We defer the discussion to Section 6.

Table 1 shows the frequency of missing values amongst the examples (events). Overall, more than 70% of the events lack at least one feature. The missing values do not result from flaws in the simulation process, that would be the equivalent of unobserved measurements, but are structurally absent (Chechik et al., 2008). For instance, in events where there is no jet (`PRI_jet_num` $= 0$), there is no such a thing as a "leading jet", thus the associated primitive quantities (`PRI_jet_leading_pt`, `PRI_jet_leading_eta`, `PRI_jet_leading_eta`)

[13] *e.g.,* https://www.kaggle.com/c/higgs-boson/forums/t/9815/ams-metric-drastically-different-from-training-and-

	Signal	Background
No jet	81002 (29.6%)	239067 (45.3%)
One jet	88189 (32.3%)	159071 (30.2%)
Total jet	169191(61.9%)	398138 (75.5%)
Higgs Mass	9011 (3.3%)	114925 (21.8%)

Table 1: Missing values, number of events and percentage of the class. The first column is the cause.

are structurally undefined, and the features derived from these as well. Such data intrinsically live in a lower dimensional subspace of the feature space, determined by its own actual features. All missing features are related to `PRI_jet_num`, except `DER_mass_MMC`.

5.3. The competition

The organization followed the usual challenge methods. The challenge lasted four months (from May 12 to September 15, 2014). The participants received immediate feedback on the validation data through the public leaderboard. Their ranking was decided on the test data (the private leaderboard) and remained hidden until the end of the challenge. Table 2 shows some scores from the private leaderboard. The calculated metric is the best AMS_c (7) among two solutions formally submitted by each participant. The results of the top ranking participants who submitted their code were checked by the organizers who successfully reproduced their results.

Three cash prizes awarded to the top three submissions. The *HEP meet ML* special award was for the team that, as judged by the ATLAS collaboration members on the organizing committee, creates a model that is most useful for the ATLAS experiment: optimized AMS, simplicity/straightforwardness of approach, performance requirements (CPU and memory demands), and robustness with respect to lack of training statistics.

6. Challenge analysis

6.1. Facts and figures

Quantitatively, the challenge was a big success. 1785 teams (1942 people) participated (submitted of at least one solution) and 6517 people downloaded the data. It is one of the most popular challenge on the Kaggle platform so far, to be compared with the Amazon.com employee access challenge (1687 teams) or the Allstate Purchase Prediction Challenge (1567 teams). Overall, 35772 solutions were uploaded. The untuned TMVA benchmark (a software widely used in high energy physics[14]) was beaten the first day, and the MultiBoost benchmark[15] supplied at the Challenge opening (last line of table 2) was outperformed in a few days. Moreover, the community involvement was intense, with 1100 posts in the forum organized in 136 topics.

[14] http://tmva.sourceforge.net

[15] http://higgsml.lal.in2p3.fr/software/multiboost/

Table 2: Excerpts of the private leaderboard. The relevance of the six-digits scores is discussed in Section 6.2. See details on the methods in the text.

Rank	Team	Score	Entries	Method
1	Gábor Melis	3.80581	100	DNN
2	Tim Salimans	3.78913	57	RGF and meta ensemble
3	nhlx5haze	3.78682	254	Ensemble of neural networks
8	Luboš Motl's team	3.76050	589	XGboost and Intensive feature engineering
31	Mymo	3.72594	73	Ensemble of cascades and non-cascaded models
45	Crowwork	3.71885	94	XGBoost Tuned
782	Eckhard	3.49945	29	TMVA Tuned
902	Benchmark	3.40488	NA	MultiBoost
991	Benchmark	3.19956	NA	TMVA

With respect to the goal of demonstrating the potential of machine learning methods to the HEP community, the challenge was a big success too, as the tuned version of TMVA was greatly outperformed (see table 2 for details).

6.2. Evaluation

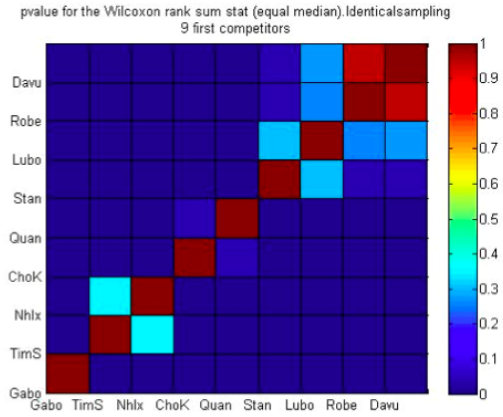
Evaluating the results of the challenge raises two questions. With the terminology of (Dietterich, 1998), the questions are

- Is the ranking of the *classifiers* statistically significant?
- What information does the ranking provide with respect to the *learning algorithms*?

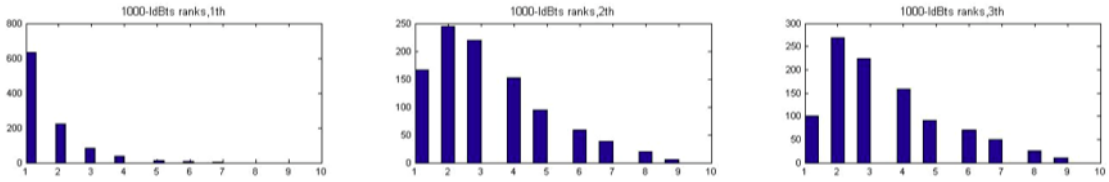
A classifier is a function that, given an input example, assigns that example to a class. A learning algorithm is a function that, given a set of labeled examples, constructs a classifier[16]. In the Challenge, the only information used for ranking was the score on the test set. Thus, what is evaluated is the performance of the classifiers without the possibility of retraining. How random the ranking is? Immediately after the challenge, that was a hot topic in the forum, with approximate computations of the "pure luck" effects, and numerous claims of a solution with a winning score, which was (unfortunately) not the one finally selected.

In order to more precisely quantify the likelihood of a different ranking, we resorted to bootstrap. Each solution (classified private test set) is bootstrapped 1000 times, with identical sampling across submissions, resulting in 1000 AMSes. The distribution of the ranks (Figure 8(*b*)) is consistent with the first position of the winner, indicating that the second and the third teams are close to each other, and well separated from the fourth

[16] As explained before, we are not considering classifiers but selectors, but the distinction does not matter here.



(a) p-values of the pairwise Wilcoxon rank sum test



(b) Rank distribution

Figure 8: Statistical significance of the ranking

one. The classical non-parametric test to quantify this intuition is the pairwise Wilcoxon rank-sum test. The null hypothesis is equal medians, against the alternative that they are not. The test assumes that the two samples are independent, which is reasonable here, as the underlying methods are different. For completeness, we recall that the Wilcoxon rank sum statistic is $W = U + n(n+1)/2$, where U is the Mann-Whitney U-test statistic, which is the number of times a y precedes an x in an ordered arrangement of the elements in the two independent samples X and Y, and n is the sample size. Figure 8(a) shows the p-value of the statistic for the top entries; the larger values indicate that the null hypothesis cannot be rejected. At 95% confidence level, the first solution is indeed better than the rest, whereas the second and third are indistinguishable, but better than the following ones and so on.

Comparing learning algorithms is of course more important in the long run. What information does the ranking provide to the end-users, here the ATLAS collaboration, about the comparison of the methods performance-wise only (not taking into account other factors such as usability)? The question has no possible answer with the ranking alone. However, the evaluation methodologies used by the challengers can help: qualitatively, if the selection of the solution is less dependent on the public score than on internal validation, the results can be expected to be robust (see Section 6.3.3 for details).

6.3. Methods

6.3.1. The learners

As usual in general classification problems, combining multiple models was recognized to be beneficial. However, the challenge does not show a clear advantage towards a particular model class or combination strategy. On the one hand, the winning solution described in a companion paper (Melis, 2015) is an ensemble of moderately deep neural networks (DNN) with identical hyper parameters that only differ in their initializations and training sets. The combination is simple bagging, advocated by the alleged noisiness of the AMS that would preclude accurate performance prediction. The approach of the third participant[17] is also based on an ensemble of neural networks, but with different hyper parameters, and a specific training.

On the other hand, gradient boosting was very popular. Many successful solutions are based on the XGBoost implementation of boosted decision trees. The creators of XGBoost competed as the Crowwork team, which received the HEP meet ML award. Beyond the practical advantages of the software (in particular computational efficiency and modularity), the untuned XGBoost provided an effective baseline (∼3.64 on the test set, ranking in the 200th) (Chen and He, 2015). Some physicists strongly involved in the competition decided in the course of the challenge to switch from their original models to XGBoost. The ease of interpretability of the model probably helped both the Crowwork team and these physicists in feature construction, which, together with tuning, improved their results in the range of 3.71-3.76. The major advantage of the XGBoost with the challenge over straightforward gradient boosting relies on explicit regularization instead of manual hyper parameter (shrinkage, tree size and number of trees) tuning to counterbalance the overfitting effect of the greedy model search.

The solution of the second winner,Tim Salimans, pushes decision trees further away from greediness. It is based on the Regularized Greedy Forest (RGF) algorithm (Johnson and Zhang, 2014), a variation on gradient boosting that decouples structure search and optimization. The final solution combines multiple RGF through stacking, with a linear regression as the learning model.

The ChoKo team, which ranked 4th, used a meta-ensemble approach including GBT (XGboost), DNN and RGF models, with heterogeneity also within each model class. Meta-ensembling was based on a genetic algorithm developed during the last days of the challenge for merging linearly the logistic output of the classifiers.

The overwhelming majority of the participants adopted the classical two-step procedure described in Section 4.2 of optimizing first the classification error and considering the AMS only for the selection threshold. It has been argued that this was to be preferred because of the instability (noisiness) of the AMS. But the results of the Mymo team show that combining direct optimization of the AMS and regularization is promising. Moreover, the cascade method described in a companion paper (Mackey et al., 2015) has the important advantage of allowing to integrate off-the-shelf classifiers.

[17] https://www.kaggle.com/c/higgs-boson/forums/t/10481/third-place-model-documentation/55390

6.3.2. Feature engineering

Feature construction was intensively explored. Automatic discovery was repeatedly reported unsuccessful. Several physicists teams (*e.g.,* . Phunter, C.A.K.E. and Luboš Motl) created theory-rooted features that significantly increased the discriminative power of XGBoost. The top non-physicists palliated to their lack of domain understanding by systematically looking for synthetic features and discriminative transformations of the feature space. Overall, the physics-agnostic features were reported to be beneficial, but only moderately, and the physics-based features were less useful than the specific statistical/Machine Learning know-how about validation procedures in a noisy environment[18]. The effect of the combination of both expertise has not been fully tested in the challenge.

The effect of the "CAKE features" is particularly important to analyze to this respect. Two new derived features, *CakeA* and *CakeB*, were proposed by a competing team related to ATLAS[19]. These variables rely on the calculation of the likelihood from first principles for the event to be a signal rather than a Z to tau tau (the main background). Adding this feature had mixed results, being moderately beneficial for some methods but increasing model instability. This instability can easily be understood when ensembling involves optimization: as the most hard to classify events are precisely the Z-boson ones, the feature is a model in itself, and should be promoted to this level. The fact that these features were provided only a few days before the challenge closure caused a storm of discussion about the fairness of that action on the forum.

The traditional way to deal with missing features is imputation, where some reasonable values replace the missing ones, possibly with adding supplementary features (flags) for denoting the absence in a given instance. For structurally missing features, it is well known that imputation may produce meaningless completions, and hamper classification performance (Chechik et al., 2008). Despite the high frequency of missing values, there was no consensus on their management, from basic mean imputation (Gábor Melis) to a specialized tree-splitting algorithm (Crowwork).

6.3.3. Validation

A crucial ingredient for model selection was finding a reliable way to measure generalization performance. Some expert participants in the forum remarked early that the score on the public leaderboard was to be largely ignored, and that intensive cross-validation was required. However, the problem might be subtle: the scores on the public and private leaderboard were consistently close for the top participants (Figure 9). Figure 10 gives some insight. The public AMS curve is necessarily noisier than the private one, due to the difference in size (the dispersion being proportional to $1/\sqrt{n}$). However, the public AMS curve is more peaked, and in particular drops much faster.

As has been pointed in to forum, the public test set might have been harder to classify than the private test set *i.e.,* it over-represented hard-to-classify events at the border between signal and background. More precisely, Figure 10(*d*) shows that the private AMS curves are close, except for the winner's one, and that the difference in score mostly resulted from the selection of the cutoff, for which the public score was a poor indicator.

[18] https://www.kaggle.com/c/higgs-boson/forums/t/10350/how-physicists-fared

[19] The associated code was released on https://bitbucket.org/tpgillam/lesterhome

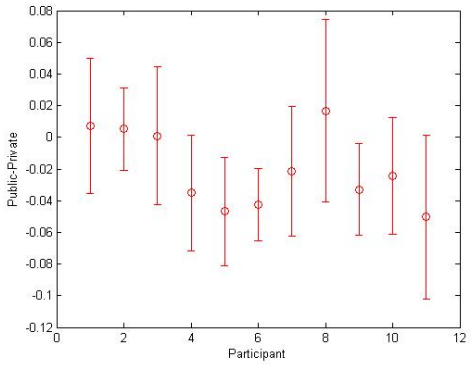


Figure 9: Summary statistics over all submissions for the top ten participants in rank order and Crowwork team.

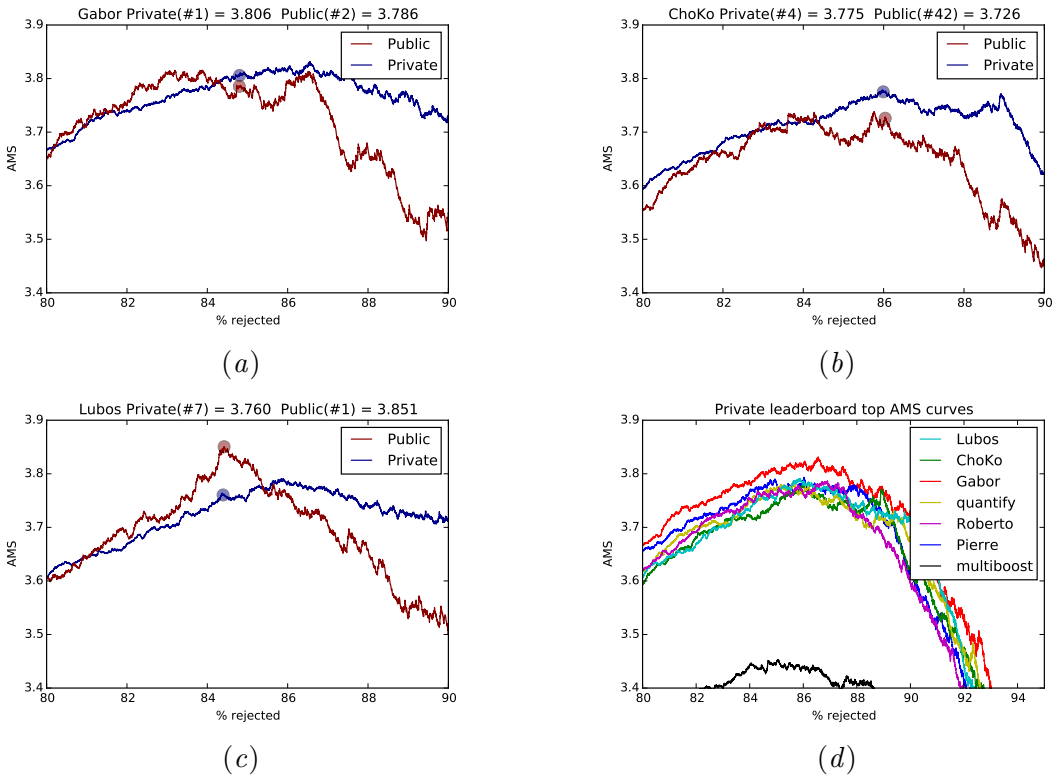


Figure 10: Comparison of private and public score for the final submissions. The horizontal axis is the weighted proportion of selected backgrounds

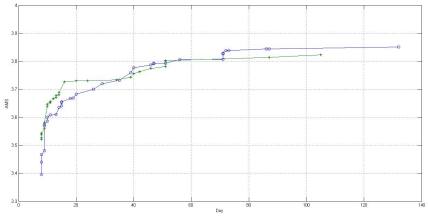


Figure 11: Maximum scores over time.

Figure 11 illustrates the dynamics of the competition through the timeline of the public and private AMS scores. Besides the natural asymptotic aspect of the curve, a curious observation is the inversion of the public/private ordering: during the first month, the public scores are well below the private ones, whereas they are much closer afterwards. A possible explanation could be that in the first month, the participants were more concerned by experimenting strategies, and went to detailed optimization of their preferred method in the remaining time, with more risk of overfitting.

7. Discussion

Beyond its immediate goals, the Challenge served to formalize new questions, which remain largely open and may raise some interest in both the statistics and machine learning communities.

7.1. More on the Challenge objective function

7.1.1. The systematics

When defining the likelihood ratio (11), we assumed that the background expectation μ_b in the selection region $\mathcal{G}$ is known. In practice, it is usually not the case: μ_b is known only within $\pm\sigma_b$ either because it is estimated on a random set of simulations or because the model used to generate the background sample has quantifiable *systematic uncertainties* ("known unknowns").

In the physics analysis of (The ATLAS Collaboration, 2013) the final selection region is much smaller than the one we find when maximizing the Challenge AMS (7). Our AdaBoost significance of ~ 3.7 sigma is obtained in a region with about $b = 3700$ background and $s = 230$ signal points. The problem is that the systematic uncertainty of the background is *relative*: it is about 10% of b itself which means that the excess of $s = 230$ points can be easily explained by a systematic misestimation of the background. For this reason, the real analysis can only accept much smaller regions, in which $\sqrt{b} > 0.1b \Rightarrow b < 100$, typically containing only some tens of background and signal points.

We can model this uncertainty using a Gaussian distribution with mean μ_b and variance ${\sigma_b}^2$, and rewrite the likelihood as

$$L(\mu_s, \mu_b) = P(n, b|\mu_s, \mu_b, \sigma_b) = \frac{(\mu_s + \mu_b)^n}{n!} e^{-(\mu_s+\mu_b)} \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(b-\mu_b)^2/2{\sigma_b}^2}. \quad (18)$$

Since μ_b is now a nuisance parameter, we construct a statistical test of the background-only ($\mu_\mathrm{s} = 0$) hypothesis using the *profile* likelihood ratio

$$\lambda(0) = \frac{L(0, \hat{\hat{\mu}}_\mathrm{b})}{L(\hat{\mu}_\mathrm{s}, \hat{\mu}_\mathrm{b})}, \tag{19}$$

where $\hat{\mu}_\mathrm{s}$ and $\hat{\mu}_\mathrm{b}$ are the values of μ_s and μ_b that maximize $L(\mu_\mathrm{s}, \mu_\mathrm{b})$ and $\hat{\hat{\mu}}_\mathrm{b}$, called the profiled value of μ_b, is the value which maximizes $L(0, \mu_\mathrm{b})$, that is, $L(\mu_\mathrm{s}, \mu_\mathrm{b})$ with μ_s fixed to zero. From Eq. (19) we can then derive a refined approximate median significance

$$\mathrm{AMS}_1 = \sqrt{2\left((s+b)\ln\frac{s+b}{b_0} - s - b + b_0\right) + \frac{(b-b_0)^2}{{\sigma_\mathrm{b}}^2}}, \tag{20}$$

where s and b are defined in Eqs. (3) and (5), and b_0 (which is $\hat{\hat{\mu}}_\mathrm{b}$ with n replaced by b) is given by

$$b_0 = \frac{1}{2}\left(b - {\sigma_\mathrm{b}}^2 + \sqrt{(b-{\sigma_\mathrm{b}}^2)^2 + 4(s+b){\sigma_\mathrm{b}}^2}\right). \tag{21}$$

The refined AMS_1 (20) captures the sytematics: when σ_b is set to $0.1b$, the AMS is maximized in a region similar in size to the region used by (The ATLAS Collaboration, 2013) Unfortunately, small region (and number of *simulated* points) means high variance; for this reason, the refined AMS_1 is very difficult to use for comparing different classifiers.

7.1.2. Testing and learning

Assume that we are given a real-valued discriminant function $f(\mathbf{x})$ that maps data points x_i to $z_i = f(\mathbf{x}_i)$. Then, instead of a counting test in the selection region $\mathcal{G} = \{\mathbf{x} : f(\mathbf{x}) > \theta\}$, on could consider a classical likelihood ratio test

$$\prod_{i=1}^{n} \frac{q_\mathrm{b}(z_i)}{q_\mathrm{s+b}(z_i)}, \tag{22}$$

where $q_\mathrm{b}(z)$ is the probability density of observations $z = f(\mathbf{x})$ under the *only* background hypothesis, and $q_\mathrm{s+b}(z)$ is the probability density of observations $z = f(\mathbf{x})$ under the hypothesis in which *both* the background and the signal processes are present. The well-known result of (Dempster and Schatzoff, 1965) is that the *expected* significance of this test is the probability that $X < Y$ where X and Y are independent random variables drawn from q_b and $q_\mathrm{s+b}$, respectively. Maximizing the expected significance with respect to f thus boils down to maximizing the AUC of f on a training sample in which negative points come from the background distribution p_b and positive points come from the mixture of the signal and background distributions

$$p_\mathrm{s+b}(\mathbf{x}) = P(y = \mathrm{s})p_\mathrm{s}(\mathbf{x}) + P(y = \mathrm{b})p_\mathrm{b}(\mathbf{x}).$$

The question of replacing the counting test by more powerful statistical tests is also very much related to the systematics. We may even conjecture that when the background density $p_\mathrm{b}(\mathbf{x})$ or the projected density $q_\mathrm{b}\big(f(\mathbf{x})\big)$ is only known within an uncertainty relative to the density, the optimal test includes (hard) cutting away observations at which $q_\mathrm{b+s}\big(f(\mathbf{x})\big)$ does not exceed significantly $q_\mathrm{b}\big(f(\mathbf{x})\big)$. This would mean that learning to discover is more closely related to classification than to rank (AUC) optimization, suggested by the classical likelihood ratio test (22).

8. Conclusion

The Challenge was a great success, both by the sheer number of participants and by its considerable visibility. Despite the relatively exotic problem, the results are largely in line with the general trends: ensemble methods, deep neural networks and advanced tree-based algorithms heavily dominate. This is good news, creating numerous opportunities for the transfer of the know-how from machine learning to high energy physics The ATLAS experiment has already started re-importing into HEP some the challenge developments. The other way is also exciting: the Challenge highlights new questions related to the learning for discovery model

We have touched upon some of the open questions in Section 2. A classical machine learning question is to design algorithms to optimize the concrete AMS objective measure, with the usual tools (e.g, regularization, surrogate design, etc.) It is clear that the two-step procedure of optimizing the classification error then the selection threshold is suboptimal, and it is likely that the learning algorithm can be improved by optimizing the AMS or one of its surrogates directly. This Challenge was also designed for answering to this question, and first results were reached.

At the statistical side, the most interesting open questions are 1) the design of more powerful tests to replace the simple counting test, and 2) the design of approximate but analytical measures to be optimized, which then can be fed into learning algorithms. The key concept that has to be understood and modeled to answer these questions is the systematic uncertainty, mentioned in Section 7.1. The main question here is whether we can design a regularized or smoothed version of AMS_1 which can model the systematic error and, at the same time, has a low variance. The question of replacing the counting test by more powerful statistical tests is also very much related to the systematics.

The final, and arguably the most futuristic, question is whether deep learning can be used to improve the discovery significance. There are hints that it could (Baldi et al., 2014), especially when looking for exotic particles. The idea would be to go closer to the raw output of the detector, and let some automatic techniques discover the representation of the events that can maximize the discovery significance.

Some initiatives are taken to advance these questions, and to explore other interaction areas concerning both high energy physics and Data Science, in particular machine learning. They will be announced on the mailing list HEP-data-science@googlegroups.com.

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 (funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02), 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. Balaázs Kégl was supported by the ANR-2010-COSI-002 grant of the French National Research Agency.

References

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.*, B716:1–29, 2012a. doi: 10.1016/j.physletb.2012.08.020.

G. Aad et al. A Particle Consistent with the Higgs Boson Observed with the ATLAS Detector at the Large Hadron Collider. *Science*, 338:1576–1582, 2012b.

G. Aad et al. Evidence for higgs-boson yukawa couplings in the $h \to \tau\tau$ decay mode with the atlas detector. *JHEP* 1504(2015)117 AarXiv:1501.04943.

Aaltonen, T. et. al. Observation of electroweak single top-quark production. *Phys. Rev. Lett.*, 103:092002, Aug 2009. doi: 10.1103/PhysRevLett.103.092002. URL http://arxiv.org/abs/0903.0885.

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. doi: 10.7483/OPENDATA.ATLAS.MQ5J.GHXA. http://dx.doi.org/10.7483/OPENDATA.ATLAS.MQ5J.GHXA.

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

P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nat Commun*, 5, 07 2014.

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

S. Chatrchyan et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys.Lett.*, B716:30–61, 2012. doi: 10.1016/j.physletb.2012.08.021.

G. Chechik, G. Heitz, G. Elidan, and D. Abbeel, P.and Koller. Max-margin classification of data with absent features. *Journal of Machine Learning Research*, 9:1–21, 2008.

T. Chen and T. He. Higgs boson discovery with boosted trees. In Cowan et al., editor, *JMLR: Workshop and Conference Proceedings*, number 42, 2015.

S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In *Proceedings of the 18th Annual Conference on Learning Theory*, COLT'05, pages 1–15. Springer-Verlag, 2005.

G. Cowan, K. Cranmer, E. Gross, and O. Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71:1554–1573, 2011. ISSN 1434-6044. doi: 10.1140/epjc/s10052-011-1554-0. URL http://arxiv.org/abs/1007.1727.

A. P. Dempster and M. Schatzoff. Expected Significance Level as a Sensitivity Index for Test Statistics. *Journal of the American Statistical Association*, 60(310):420–436, 1965.

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

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.*, 10(7):1895–1923, 1998.

R. Johnson and T. Zhang. Learning nonlinear functions using regularized greedy forest. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):942–954, 2014.

Brian L. Joiner. The median significance level and other small sample measures of test efficacy. *Journal of the American Statistical Association*, 64(327):971–985, 1969.

L. Mackey, J. Bryan, and Y M Mo. Weighted classification cascades for optimizing discovery significance in the higgsml challenge. In Cowan. et al., editor, *JMLR: Workshop and Conference Proceedings*, number 42, 2015.

G. Melis. Dissecting the winning solution of the higgsml challenge. In Cowan et al., editor, *JMLR: Workshop and Conference Proceedings*, number 42, 2015.

S. Rosset, C. Perlich, G. Swirszcz, P. Melville, and Y. Liu. Medical data mining: insights from winning two competitions. *Data Mining and Knowledge Discovery*, 20(3):439–468, 2010.

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

V. M. Abazov et al. Observation of single top-quark production. *Physical Review Letters*, 103(9), 2009. doi: 10.1103/PhysRevLett.103.092001.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, 9:60–62, 1938.

Appendix 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,

$$E^2 = p^2c^2 + m^2c^4, \tag{23}$$

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. (23) simplifies to

$$E^2 = p^2 + m^2. \tag{24}$$

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 $\boldsymbol{v}$ and $\boldsymbol{p}$, that is, $\boldsymbol{p} = \gamma m\boldsymbol{v}$, where, by convention, $p = |\boldsymbol{p}|$ and $v = |\boldsymbol{v}|$. 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[20], so its energy can be computed from the momentum and mass using Eq. (24). Conversely, the mass of a particle with known momentum and energy can be obtained from

$$m = \sqrt{E^2 - p^2}. \tag{25}$$

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

[20] neglecting the particle width

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 $(p_x, p_y, p_z, E) = (\boldsymbol{p}, E)$ 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. By conservation of the energy and momentum (which are fundamental laws of nature), we can write $E_H = E_A + E_B$ and $\boldsymbol{p}_H = \boldsymbol{p}_A + \boldsymbol{p}_B$. Since the energies and momenta of A and B are measured in the detector, we can compute E_H and $p_H = |\boldsymbol{p}_H|$ 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

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

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

$$p_{\mathrm{T}} \times \cosh\eta, \tag{26}$$

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

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

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

where ϕ_{T} is 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),

$$m_{\mathrm{inv}}(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\left(\sqrt{a_x^2 + a_y^2 + a_z^2} + \sqrt{b_x^2 + b_y^2 + b_z^2}\right)^2 - (a_x + b_x)^2 - (a_y + b_y)^2 - (a_z + b_z)^2}. \tag{27}$$

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),

$$m_{\mathrm{tr}}(\boldsymbol{a},\boldsymbol{b}) = \sqrt{\left(\sqrt{a_x^2+a_y^2}+\sqrt{b_x^2+b_y^2}\right)^2-(a_x+b_x)^2-(a_y+b_y)^2}. \tag{28}$$

The pseudorapidity separation between two particles A and B is

$$|\eta_A-\eta_B|. \tag{29}$$

The R separation between two particles A and B is

$$\sqrt{(\eta_A-\eta_B)^2+(\phi_A-\phi_B)^2}, \tag{30}$$

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

Appendix 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 (The ATLAS Collaboration, 2013) 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 (28) between the missing transverse energy and the lepton.

`DER_mass_vis` The invariant mass (27) of the hadronic tau and the lepton.

`DER_pt_h` The modulus (26) 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 (29) between the two jets (undefined if `PRI_jet_num` ≤ 1).

`DER_mass_jet_jet` The invariant mass (27) 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 (30) between the hadronic tau and the lepton.

`DER_pt_tot` The modulus (26) 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` $\geq$ 1) and the subleading jet (if `PRI_jet_num` $= 2$) (but not of any additional jets).

`DER_sum_pt` The sum of the moduli (26) 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 $\boldsymbol{E}_{\text{T}}^{\text{miss}}$ is on the bisector of the transverse momenta of the lepton and the hadronic tau. It decreases to 1 if $\boldsymbol{E}_{\text{T}}^{\text{miss}}$ is collinear with one of these vectors and it decreases further to $-\sqrt{2}$ when $\boldsymbol{E}_{\text{T}}^{\text{miss}}$ 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 $\boldsymbol{E}_{\text{T}}^{\text{miss}}$.

`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 5.1. Not to be used as a feature. Not available in the Kaggle test sample, but available for all events in the `opendata.cern.ch` dataset

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

`KaggleSet` Specific to the `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 the `opendata.cern.ch` dataset: weight normalized within each Kaggle data set according to Eq. (31).

The events (instances) and the features were used for the training and optimization of the reference ATLAS analysis (The ATLAS Collaboration, 2013). 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 (The ATLAS Collaboration, 2013) have not been applied.
- In the reference analysis (The ATLAS Collaboration, 2013), 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.

Appendix C. The opendata.cern.ch dataset

The data set available to Challenge participants on the Kaggle platform (https://www.kaggle.com/c/higgs-boson) during the Challenge is now permanently available on the opendata.cern.ch platform (http://opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014). The following minimal modifications of the data set were necessary.

- The training set (250K events), the validation set and the test set (100K public leaderboard events and 450K private leaderboard events) were merged and added to a small data set withheld by the organizers during the challenge, making a full set of a total of 818238 events. It is now the responsibility of the user of this data set to adopt the cross validation method of his choice to avoid overtraining.

- For all events, the weights and labels are available. Weights are normalized such that the whole data set corresponds to LHC 2012 running, that is, the sums of signal and background weights *of the total set* are $N_{\rm s}$ and $N_{\rm b}$, respectively (see Eq. (1)). Hence, if a subset S' is defined, for example for testing, the weights should be renormalized by Eq. (1). More precisely, the weight w_j should be set to

$$w'_j = w_j \frac{\sum_i w_i \mathbb{1}\{y_i = y_j\}}{\sum_{i \in S'} w_i \mathbb{1}\{y_i = y_j\}}, \tag{31}$$

  where y_i is the label (s or b) of the ith event and $\mathbb{1}$ is the indicator function (*i.e.*, $\mathbb{1}\{y_i = s\} = 1$ 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). In other words, the weights of signal (background) events in the subset have to be scaled by the fraction of the sum of weights of signal (background) events in the complete data set.

- Two additional variables were made available, `KaggleSet` and `KaggleWeight`, which allow to recover the original Kaggle training, public and private data sets (see Appendix B for details), and to recompute the original exact public and private Kaggle leaderboard scores for any submission (without needing to renormalize the weights.)