

The Higgs Machine Learning Challenge

C Adam-Bourdarios¹, G Cowan², C Germain-Renaud³, I Guyon⁴,
B Kégl¹, D Rousseau¹

¹ Laboratoire de l'Accélérateur Linéaire, Orsay, France

² Department of Physics, Royal Holloway, University of London, UK

³ Laboratoire de Recherche en Informatique, Orsay, France

⁴ ChaLearn, California, USA

E-mail: `g.cowan@rhul.ac.uk`

Abstract. The Higgs Machine Learning Challenge was an open data analysis competition that took place between May and September 2014. Samples of simulated data from the ATLAS Experiment at the LHC corresponding to signal events with Higgs bosons decaying to $\tau^+\tau^-$ together with background events were made available to the public through the website of the data science organization Kaggle (kaggle.com). Participants attempted to identify the search region in a space of 30 kinematic variables that would maximize the expected discovery significance of the signal process. One of the primary goals of the Challenge was to promote communication of new ideas between the Machine Learning (ML) and HEP communities. In this regard it was a resounding success, with almost 2,000 participants from HEP, ML and other areas. The process of understanding and integrating the new ideas, particularly from ML into HEP, is currently underway.

1. Introduction

Multivariate methods have become an increasingly important tool in HEP for distinguishing between event types and searching for new signal processes. The methods developed in the closely related field of Machine Learning (ML) have provided rapid advances in this area in recent years. To help move these advances into HEP, a public competition — the Higgs Machine Learning Challenge — was held between May and September 2014. This highly popular activity has succeeded in developing sophisticated algorithms and establishing new links between the ML and HEP communities. The detailed documentation for the Challenge is available in Refs. [1, 2].

To describe the motivation for the competition, we first discuss the general use of multivariate methods in HEP in Sec. 2. Section 3 describes the Challenge itself and some of the important outcomes and next steps are discussed in Secs. 4 and 5.

2. Multivariate Methods in HEP

In a particle physics experiment, suppose one measures n quantities for each event, written here as the vector $\mathbf{x} = (x_1, \dots, x_n)$. Here, x_1 might represent the momentum of a lepton or jet, x_2 could be the missing transverse energy, and so forth. The vector \mathbf{x} will follow a joint probability density function (pdf) that depends on the type of event in question, say, $f(\mathbf{x}|s)$ for signal events and $f(\mathbf{x}|b)$ for background events.

This type of situation is illustrated in Fig. 1(a), which shows symbolically a two-dimensional subspace of the data vector \mathbf{x} . The densities for signal and background events are indicated by

the red dots and blue triangles, respectively. To select events or to define a search region for the signal process, a common practice in HEP is to make cuts on the variables, as indicated in Fig. 1(a). This is a simple procedure and allows one to use some physical intuition in determining the value of the cut.

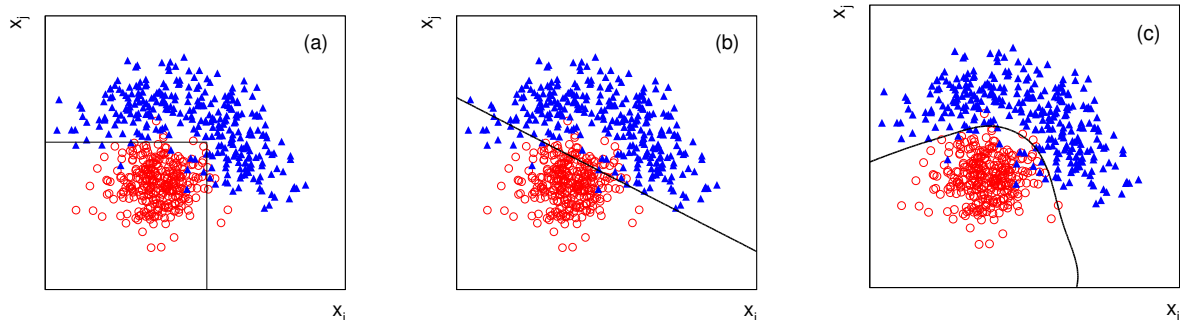


Figure 1. Scatter plots of two variables corresponding to two hypotheses: signal and background. Event selection could be based, e.g., on (a) cuts, (b) a linear boundary, (c) a nonlinear boundary.

Another relatively simple procedure is to classify events using a linear boundary as shown in Fig. 1(b), but to exploit the potentially complicated shapes of the pdfs $f(\mathbf{x}|s)$ and $f(\mathbf{x}|b)$ one should use a nonlinear boundary as in Fig. 1(c). In fact, the Neyman-Pearson lemma [3] guarantees that an optimal boundary for event classification is obtained using contours of constant likelihood ratio,

$$\lambda(\mathbf{x}) = \frac{f(\mathbf{x}|s)}{f(\mathbf{x}|b)}. \quad (1)$$

As the pdfs $f(\mathbf{x}|s)$ and $f(\mathbf{x}|b)$ are in general not known in closed form, one is unable to evaluate $\lambda(\mathbf{x})$ at an arbitrary point \mathbf{x} , and thus the likelihood ratio is not directly usable in practice. Instead, Monte Carlo models for the signal and background processes are used to generate samples of training data. Using these, Machine Learning algorithms allow one to find a function $y(\mathbf{x})$ such that the classification boundary is a surface of constant $y(\mathbf{x})$. Ideally one would like to find the $y(\mathbf{x})$ that best approximates the likelihood ratio or a monotonic function thereof.

If both the signal and background processes are known to exist, then the problem is in effect one of event classification. For example, one may want to select a sample enriched in signal events for further study. For given training samples of signal and background events, this is the usual problem of supervised learning. In many HEP problems, however, the existence of the signal process is not established and it is the goal of the analysis to see whether it is present at all. This was the case with the Higgs boson before its discovery in 2012 [4, 5] and it remains relevant for processes involving Higgs such as the decay $H \rightarrow \tau^+ \tau^-$, for which the significance of current observations has not yet reached the level of five standard deviations [6, 7].

Multivariate methods such as Fisher discriminants and neural networks have been used for many years in HEP for the types of problems described above, and played an important role in the analyses carried out at the Tevatron at Fermilab and the LEP Collider at CERN in the 1990s. More recently, advances in Machine Learning have led to new tools such as the Boosted Decision Tree (BDT) and Support Vector Machine (SVM). Many of these developments are described in standard texts such as those by Hastie et al. [8] and Bishop [9]. Although BDTs

are now widely used in HEP, rapid progress in Machine Learning has resulted in many advances that are only slowly percolating into Particle Physics.

3. The Higgs ML Challenge

In order to promote the knowledge transfer between Machine Learning and Particle Physics, a public competition — the Higgs Machine-Learning Challenge — was held in the summer of 2014. Challenges have become a widely used activity in the Machine Learning community to test and develop new algorithms. They give computer scientists the opportunity to test their ideas on real-life problems, which provides added value and interest to the results. Several activities of this type have recently drawn large numbers of participants, including a competition by the movie-lending company Netflix to improve the accuracy of its film-recommendation algorithm [10] and a crowd-sourced investigation of dark matter by NASA and the Royal Astronomical Society [11].

From the standpoint of those creating the Challenge, such competitions provide an opportunity to outsource difficult problems to a wider community and thus to take advantage of a broad range of expertise. Those outside the narrow circle of experts may not have the background knowledge to understand fully all of the subtleties of the problem, but because they come from a wide range of backgrounds they are more likely to “think outside the box” and bring in useful ideas that might be overlooked by the experts.

The competition also serves an important role as an outreach activity. In the case of the present competition it helped publicize the science of the LHC and the Higgs boson to a large number of scientifically interested people.

The Challenge was hosted on the website of the data science organization Kaggle [12]. Kaggle (kaggle.com), founded in 2010, provides a platform for data-science competitions and has at any given time roughly 20 running in different areas, ranging from the physical and life sciences to economic and marketing problems, pattern recognition, weather forecasting, etc. The Kaggle website provides a repository for the data sets needed for the Challenge as well as a framework for publicizing the competition and offering monetary prizes, an online discussion forum and a continually updated public leaderboard.

The Challenge was organised by a team containing physicists from the ATLAS Collaboration and computer scientists from the Paris Saclay Center for Data Science and from the data science organization ChaLearn. It received further support from CERN, Google and INRIA.

3.1. The Mathematical Problem

The data for the Challenge consisted of 800,000 fully simulated events provided by the ATLAS Collaboration corresponding to the signal process with Higgs decaying to $\tau^+\tau^-$ and background events from top-antitop and $Z \rightarrow \tau^+\tau^-$. Details of the ATLAS Experiment can be found in Ref. [13]. For each event, 30 numbers were recorded, including “primitive” quantities such as the jet and lepton momenta as well as derived quantities like the missing transverse energy and visible mass.

A subsample of 250,000 of these events were labeled accordingly as signal or background. This training sample was used by participants to design an algorithm that could be used to search for the signal process. The remaining 550,000 events were used for testing the performance of the algorithm.

In many competitions in the ML community, the problem is only to classify different types of events, and the algorithm with the smallest error rate on a statistically independent test sample is the winner. The Higgs ML Challenge was somewhat unusual in that the goal of the analysis is not simply to select Higgs events, but rather to try to establish whether the signal process $H \rightarrow \tau^+\tau^-$ exists.

To test for the existence of signal, a simple and widely used procedure is to designate a “search region”, i.e., a region of the 30-dimensional space of input variables, in which one expects signal events to be present. If the number of events actually observed there is significantly greater than the expected number of background events, then one may reject the background-only hypothesis. In HEP, a statistical significance of five standard deviations for this test is often regarded as reaching the standard of a “discovery”.

This type of statistical analysis has been widely studied in HEP (see, e.g., Ref. [14]). The number of events n one would observe in the search region is modeled as following a Poisson distribution with a mean of $s + b$, where s and b are the expected numbers of events from signal and background processes, respectively. If the signal process is indeed present, then observed statistical significance with which one rejects the background-only hypothesis can be approximated by the “AMS” (approximate median significance, see Ref. [14]),

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

In the limit where $s \ll b$, this quantity reduces to s/\sqrt{b} , which has often been used in HEP as a measure of experimental sensitivity.

The typical approach to the competition involved two steps. First, a function $y(\mathbf{x})$ is defined as a basis of classifying the events as signal or background such that, e.g., $y(\mathbf{x}) \geq C$ corresponds to the search region where the excess of signal events is expected. Second, a value for the threshold C is chosen such so as to maximize the AMS. For a given function $y(\mathbf{x})$ and threshold C , the search region is fully defined and thus the values of s and b could, given enough simulated data, be estimated to arbitrary precision. In practice, of course, the data samples are finite and the estimates of s and b are themselves subject to some statistical uncertainty.

For the purposes of the Challenge, a particular problem arose in this regard. One could imagine a participant designating a very small search region, i.e., the true values of both b and s would be small. Because of the random nature of the test data, the estimate of b might fluctuate very low, even down to zero. The estimated experimental sensitivity would then be large (or infinite). This apparently high performance would, however, be an illusion, as the the median significance could well be wildly overestimated.

To avoid allowing such an entry from winning the competition, the formula for the AMS (2) was regularized by adding to the estimate of b an offset b_{reg} . This step was taken solely for purposes of ensuring a fair and interesting competition. In a more detailed analysis one would try not only to obtain a high AMS but also ensure that it has a small statistical error.

For the Challenge, the estimates of s and b were based on the 550,000 events designated as the test sample, in two different settings. First, 100,000 of these events were used to evaluate the AMS and the result was posted on the public leaderboard. This was an important factor for making an interesting and exciting competition. If this were to be the sole judge of quality, however, participants would be able to “overtrain” their classifier on this data sample and the resulting algorithm would not reflect such high performance on unseen data. Therefore the remaining 450,000 events were used to determine the actual winners.

3.2. Running the Challenge

The Challenge ran from 12 May to 15 September 2014. 6,517 people downloaded the data, and 1,785 teams (1,942 people) uploaded 35,772 solutions, making the Higgs ML Challenge the most popular competition hosted by Kaggle to date. Throughout the competition an online forum collected 1,100 posts, allowing participants and organizers to discuss a wide range of issues.

The uploaded solutions were evaluated using the public test sample (100,000 events) and the ranked results shown on the public leaderboard. To prevent participants from excessive

training of the algorithm using the test data sample, uploaded solutions were limited to five per day. Immediately after the start of the Challenge, the leading AMS value grew to around 3.5. Although statistically significant gains became more difficult to achieve after the first few weeks, some additional progress continued, as shown in Fig. 2

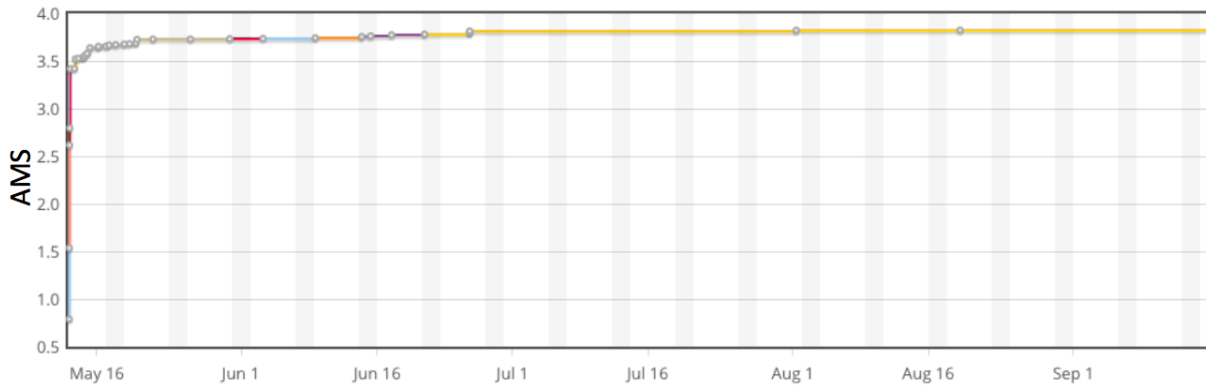


Figure 2. The leading value of AMS versus time over the course of the Higgs Machine Learning Challenge [12].

4. Results of the Challenge

The highest score at the end of the competition was by Gábor Melis from Hungary, with an AMS value of 3.80581. He was followed closely by Tim Salimans from the Netherlands with 3.78913 and Pierre Courtiol from France with 3.78682. They received prizes of \$7,000, \$4,000 and \$2,000, respectively.

A special award of a trip to CERN was given for the solution judged to have the greatest potential benefit for HEP. This “HEP meets ML” award was won by postgraduate students Tianqi Chen from University of Washington, Seattle, and Tong He from Simon Fraser University, Canada. Their XGBoost algorithm, a parallelised procedure for training boosted decision trees, was made public early on in the competition [15] and was adopted by a number of other participants.

The winning entry by Melis was a “bagged” combination of 70 neural networks. The most difficult problem in constructing a winning entry turned out to stem from the noisy nature of the AMS when plotted as a function of the threshold used to define the search region. This can be seen in the AMS shown versus the percent rejected (equivalent to the threshold) in Fig. 3. A key feature of Melis’s entry was a careful cross validation to obtain a robust estimate of the AMS based on the relatively small training sample of 250,000 events.

5. Next Steps

To allow people to continue developing their algorithms, the data from the Challenge will remain available on the CERN Open Data Portal [16] with a citeable d.o.i.. The signal/background labels and weights for all 800,000 simulated events are now revealed.

Digesting the many ideas developed in the Challenge and importing them into HEP is an ongoing process expected to take some time. An important first step was the highly successful satellite workshop on the Challenge that took place during the NIPS Conference in Montreal in December 2014 [17]. The contributions to this meeting will be published in the Proceedings of Machine Learning Research 42. As a further important step, a workshop at CERN (including HEP meets ML winners Chen and He along with overall winner Melis) is to take place 19 May 2015 at CERN, with the materials available through Ref. [18].

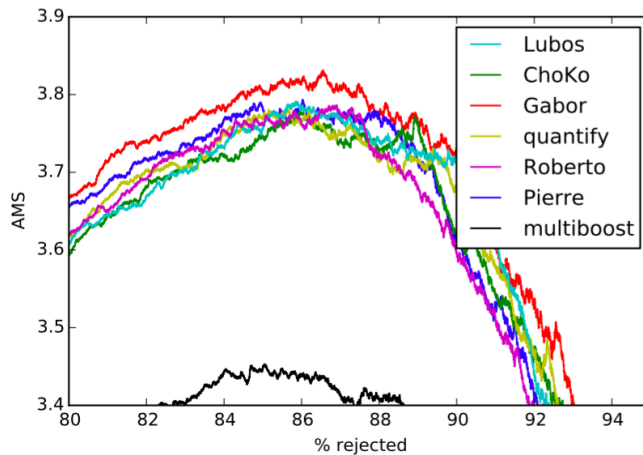


Figure 3. The AMS values of some leading competitors versus fraction of events rejected, which determines the threshold used to define the search region.

The Higgs Boson Machine Learning Challenge has been in itself an interesting piece of experimental science, testing whether it is possible through this type of activity to generate new ideas and foster long-term collaborations between the HEP and ML communities. The initial outcomes of this experiment have been highly positive.

Acknowledgements

The authors would like to thank the ATLAS Collaboration, CERN, Google, INRIA, and the Paris Saclay Center for Data Science for their support of the Higgs ML Challenge. Special thanks are due to the CHEP organizers and local hosts at OIST for a highly stimulating meeting in a spectacular environment.

References

- [1] LAL website of the Higgs Boson Machine Learning Challenge <https://higgsml.lal.in2p3.fr>
- [2] Adam-Bourdarios C, Cowan G, Germain C, Guyon I, Kégl B and Rousseau D CERN Open Data Portal DOI:10.7483/OPENDATA.ATLAS.MQ5J.GHXA, <http://opendata.cern.ch/record/329>
- [3] Kendall M G, Stuart A and Ord J K (eds) 1987 *Kendall's Advanced Theory of Statistics* (New York, NY, USA: Oxford University Press, Inc.) ISBN 0-195-20561-8
- [4] ATLAS Collaboration 2012 *Phys.Lett.* **B716** 1–29 (*Preprint* 1207.7214)
- [5] CMS Collaboration 2012 *Phys.Lett.* **B716** 30–61 (*Preprint* 1207.7235)
- [6] ATLAS Collaboration 2015 *JHEP* **1504** 117 (*Preprint* 1501.04943)
- [7] CMS Collaboration 2014 *JHEP* **1405** 104 (*Preprint* 1401.5041)
- [8] Hastie T, Tibshirani R and Friedman J 2009 *The Elements of Statistical Learning, Data Mining, Interference, and Prediction* (Springer)
- [9] Bishop C M 2006 *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag New York, Inc.) ISBN 0387310738
- [10] Website of the Netflix prize www.netflixprize.com
- [11] Website of Mapping Dark Matter <https://www.kaggle.com/c/mdm>
- [12] Kaggle website of the Higgs Boson Machine Learning Challenge <https://www.kaggle.com/c/higgs-boson>
- [13] ATLAS Collaboration 2008 *JINST* **3** S08003
- [14] Cowan G, Cranmer K, Gross E and Vitells O 2011 *Eur. Phys. J. C* **71** 1554 (*Preprint* arXiv:1007.1727)
- [15] Chen T *et al.*, XGBoost: eXtreme Gradient Boosting github.com/dmlc/xgboost
- [16] ATLAS Collaboration, Data set from the Higgs Machine Learning Challenge, CERN Open Data Portal opendata.cern.ch/collection/ATLAS-Higgs-Challenge-2014
- [17] HEPML workshop at NIPS14 <https://indico.lal.in2p3.fr/event/2632/>
- [18] Workshop: Higgs Machine Learning Challenge visits CERN cern.ch/higgsml-visit