How Machine Learning won the Higgs Boson Challenge

Claire Adam-Bourdarios, G. Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, David Rousseau

How Machine Learning won the Higgs Boson Challenge

Claire Adam-Bourdarios[1,2], Glen Cowan[4], Cécile Germain[1,3]
Isabelle Guyon[1,5], Balàzs Kégl[1,2,3], and David Rousseau[1,2]

1-LRI, UPSud, Université Paris-Saclay, France. 2-LAL, IN2P3/CNRS, France.
3-CNRS/INRIA, France. 4-Royal Holloway, London, UK. 5-ChaLearn, USA.

Abstract.
In 2014 we ran a very successful machine learning challenge in High Ernergy physics attracting 1785 teams, which exposed the machine learning community for the first time to the problem of "learning to discover" (www.kaggle.com/c/higgs-boson). While physicists had the opportunity to improve on the state-of-the-art using "feature engineering" based on physics principles, this was not the determining factor in winning the challenge. Rather, the challenge revealed that the central difficulty of the problem is to develop a strategy to optimize directly the Approximate Median Significance (AMS) objective function, which is a particularly challenging and novel problem. This objective function aims at increasing the power of a statistical test. The top ranking learning machines span a variety of techniques including deep learning and gradient tree boosting. This paper presents the problem setting and analyzes the results.

Increasingly, machine learning scientists are getting away from canonical problems of classification and regression and are venturing into new domains by formulating new tasks as learning problems. In the recent years we have seen, for example, the task of *learning to rank* [1] and *learning to recommend* [2], which have become pervasive in applications. In this paper, we tackle a new machine learning task addressing the problem of evaluating the significance of a sientific discovery: ***learning to discover***. Superficially, this problem is a two-class classification problem separating events of interest (called *signals*) that have not been encountered or characterized before in nature (but were predicted by a theory) from events produced by already observed processes (called *backgrounds*). However, the problem setting differs from regular classification problems in two respects: (1) **Discovery**: Since signals were never observed before in real data, no labeled training example from real data are available. Rather, simulated data (from a simulator implementing theoretical predictions) can be produced to generate training data. The learning machine can then address the "inverse problem" of predicting which events are signals in real data. (2) **Evaluation**: Typically the number of signals is expected to be orders of magnitude smaller that the number of backgrounds. Hence, the statistical significance of the detection of a number of signal events must be assessed to claim a discovery. For this reason, the evaluation function of the classifier is a metric of a statistical test.

The Discovery-Evaluation criteria define a wide scope that has previously been addressed mainly by Neyman-Pearson learning [3]. In this paper, we examine a new case of the learning to discover problem, the discovery of new

particles, that leads to a different framework. The problem has been used recently in a comparison of deep *vs.* shallow representation learning [4]. With the HiggsML Machine Learning challenge organized in 2014[1] [5], it becomes a benchmark problem. The data were released at http://opendata.cern.ch/ after the end of the challenge. This unprecedented disclosure of precious data belonging to the ATLAS collaboration highlights the importance of the learning to discover task. The dataset, and the subject of the challenge correspond to particular physical process, the $H \to \tau^+\tau^-$channel. However, the methodology is fully generic to the discovery of a new particle, and could generalize to other discovery settings.

From the algorithmic point of view, the novelty of the problem mainly comes from its exotic objective function, the Approximate Median Significance (AMS). The AMS presents several undesirable features to train a learning machine: it is **discontinuous** (discontinuity arise for each sample); it is **non differentiable**; it is **non additive** (the overall AMS is not the sum of individual contributions of the samples); it uses **sample weights** available only for training. This seems to have drawn a lot of interest because off-the-shelf packages do not directly support the optimization of non-standard objective functions and the participants saw an opportunity to make novel contributions and distinguish themselves from the rest of the crowd. The top ranking challenge participants greatly improved the performance over baseline methods.

1 The physics problem

The ATLAS and the CMS experiments jointly claimed the discovery of the Higgs boson [6, 7] in 2012. The Higgs boson has many different processes (called *channels* by physicsts) through which it can *decay*, that is produce other particles. 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. 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 first evidence of the H to tau tau channel was recently reported by the ATLAS experiment [8]. The aim of the challenge is to increase the statistical significance of this discovery.

Discovery and characterization rely on experiments that run at the Large Hadron Collider (LHC) at CERN. Hundreds of millions of proton-proton collisions per second are produced. The particles resulting from each bunch crossing are detected by sensors and filtered in real-time. For each collision, the raw data produced by the sensors are ultimately digested into a vector of features containing up some tens of real-valued variables. This vector is called an *event*.

The vast majority of events represent known (background) processes: they are mostly produced by processes which are exotic in everyday terms, but known, having been discovered in previous generations of experiments. The learning

[1] www.kaggle.com/c/higgs-boson and https://higgsml.lal.in2p3.fr

problem is to find a region in feature space with a significant excess of signal events. Discovery of a new particle ultimately boils down to classical statistical testing. The null hypothesis is that the experiment produces only background events, and the alternative hypothesis is that it produces some signal events.

2 The AMS objective function

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[2] .

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 $p(\mathbf{x}, y)$.[3] First, as many events of the signal class as needed can be simulated (within a computational budget), so the proportion $n_\mathrm{s}/n_\mathrm{b}$ of the number of points in the two classes does not have to reflect the proportion of the prior class probabilities $P(y = \mathrm{s})/P(y = \mathrm{b})$. This is actually a good thing: since $P(y = \mathrm{s}) \ll P(y = \mathrm{b})$, the training sample would be very unbalanced if the numbers of signal and background events, n_s and n_b, were proportional to the prior class probabilities $P(y = \mathrm{s})$ and $P(y = \mathrm{b})$. Second, the simulator produces importance-weighted events. Since the objective function (3) 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 be kept fixed, that is,

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

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

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

where $p_\mathrm{s}(\mathbf{x}_i) = p(\mathbf{x}_i | y = \mathrm{s})$ and $p_\mathrm{b}(\mathbf{x}_i) = p(\mathbf{x}_i | y = \mathrm{b})$ 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 used by the simulator.

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

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

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 $s = \sum_{i \in \mathcal{S} \cap \widehat{\mathcal{G}}} w_i$ is an unbiased estimator of the expected number of signal events selected by g, and, $b = \sum_{i \in \mathcal{B} \cap \widehat{\mathcal{G}}} w_i$ is an unbiased estimator of the expected number of background events selected by g, **In machine learning terminology, s and b are are true and false positive rates.** Given a classifier g, the AMS objective function used for the challenge is defined by

$$\mathrm{AMS} = \sqrt{2\left((s + b + b_{\mathrm{reg}}) \ln\left(1 + \frac{s}{b + b_{\mathrm{reg}}}\right) - s\right)} \tag{3}$$

where b_{reg} is a regularization term (e.g. $b_{\mathrm{reg}} = 10$). The derivation of this formula is given in [9]. Quantitatively, a fluctuation is considered anomalous by physicists (the evidence of a discovery) if it exceeds 5σ, that is if AMS > 5, which corresponds to a p-value of the one-sided Z-test of 3×10^{-7}.

3 Results of the challenge

Organizing a challenge can be thought of as designing a large scale numerical experiment in which research is "crowd-sourced". Given a well-posed scientific question, challenges are very effective in obtaining an answer. This has been the case for the Higgs-Boson challenge. The challenge helped making both qualitative and quantitative advances to the problem of optimizing the AMS.

First, from the quantitative point of view, the AMS of the top ten participants ranged in [3.76, 3.80], while the benchmark (based on a software widely used in High Energy physics called TMVA[4], ranked only 782 with an AMS of ∼3.50. This brings us closer (albeit still far) from the target set forth by physicists of AMS=5. The winning solution of Gabor Melis [10] uses a deep learning method (an ensemble of 70 3-layers neural networks, fully interconnected, with 600 hidden units per layer), confirming that deep learning methods can be very competitive. But, most top ranking participants used ensembles of decision trees, and particularly the XGBoost software[5]. The ease of interpretability of the model helped the crowwork team in feature construction and hyper-parameter tuning to attain an AMS of 3.76 [11]. The major advantage of the XGBoost over straightforward gradient boosting relies on explicit regularization, reducing the need of manual hyper-parameter tuning. However, the solution of Melis is clearly more robust as indicated by the fact that it dominated all other solutions, regardless of the choice on the threshold monitoring the tradeoff between true positive rate and false positive rate. This robustness may be attributable to the use of a voting ensemble and the use of "drop-out", a method consisting in

[4] http://tmva.sourceforge.net
[5] https://github.com/dmlc/xgboost

silencing neurons at random during training. The price to be payed for such a deep learning solution is months of architecture and hyper-parameter tuning, using a fast GPU-powered computer. In contrast, the boosted decision tree solutions are obtained relatively fast (in minutes) with little need for hyper-parameter tuning: the completely untuned XGBoost achieved ∼3.64 on the test set, ranking in the 200th.

Second, several important qualitative (conceptual) advances were made. Most of the participants (including the winner), followed a bi-level optimization approach, optimizing a surrogate cost function (logistic loss or AUC) and adjusting the cut-off classification threshold to optimize the AMS by cross-validation. For the logistic loss, [12] provides a theoretical justification for the default approach to post-fit the cut-off, proving asymptotic consistency and bounding the rate of convergence. However some effort was put into finding other ways of optimizing the AMS. For example, the 9th solution [13] uses a weighted version of the AUC as a surrogate cost function, thus approximating the AMS with an additive loss. Another principled approach to exploit all classical additive cost functions with a simple sample reweighing has been proposed by [14] who applies a variational method to optimize iteratively a set of linearized versions of the AMS.

We also had an interesting "negative" result. The promise that deep learning methods can save on human effort by learning internal representation in place of feature engineering was not held in the challenge. In the winning solution, the AMS drops by 13% if the derived (human engineered) features are not exploited. A recent benchmark of deep and shallow networks also show disappointing results for deep learning for the Higgs in tau tau task [4], with a 6% performance drop. Although the results for learning internal representations are more encouraging for other high energy physics tasks, the ratio gain to sample size indicates that learning representations is extremely data demanding.

Another noteworthy development of the challenge concerns cross-validation. All top ranking participant carefully avoided to rely on the performances shown on the public leaderboard (computed on a too small data sample to provide reliable performance evaluation). Rather they repeated multiple times 10-fold cross-validation and averaged the results. For computation efficiency reasons, they used the learning machines thus trained as part of a voting ensemble. A interesting additional twist helped Melis win the challenge: Since the AMS is not an additive loss, it is different (i) to compute the AMS on the (very small) held out sets, then averaging the results or (ii) to collect statistics about false positive and false negative on the the held out sets, average them, then compute the AMS. The latter provides much smoother results.

In conlusion, the challenge revealed that participants who are skilled data scientitists with only elementary knowledge of physics could contribute to significantly improve the power of the AMS using machine learning. Deep Learning is a favorite technique but boosted decision trees are a strong contender, with several practical advantages. Bi-level optimization using a surrogate cost function beats direct optimization of the AMS, but variational methods, which tackle the problem head front, provide a new avenue of research worth exploring further.

Bibliography

[1] O. Chapelle and Y. Chang. Yahoo! learning to rank challenge overview. In *Procs of the Yahoo! Learning to Rank Challenge, held at ICML*, JMLR Proceedings, pages 1–24, 2011.

[2] Prem Melville and Vikas Sindhwani. Recommender systems. In C. Sammut and G. I. Webb, editors, *Encyclopedia of Machine Learning*, pages 829–838. Springer, 2010.

[3] Clayton Scott and Robert Nowak. A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806 – 3819, 2005.

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

[5] G. Cowan, C. Germain, I. Guyon, B. Kegl, and D. Rousseau. The Higgs boson machine learning challenge. volume 42 of *Procs of Machine Learning Research*, 2015. http://jmlr.csail.mit.edu/proceedings/papers/v42/.

[6] 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, 2012.

[7] 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.

[8] 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.

[9] 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.

[10] G. Melis. Dissecting the winning solution of the higgsml challenge. volume 42 of *Procs. of Machine Learning Research*, 2015.

[11] T. Chen and T. He. Higgs boson discovery with boosted trees. volume 42 of *Procs. of Machine Learning Research*, 2015.

[12] W. Kotlowski. Consistent optimization of AMS by logistic loss minimization. volume 42 of *Procs of Machine Learning Research*, 2015.

[13] R. Diaz-Morales and A. Navia-Vazquez. Ensemble of maximized weighted auc models for the maximization of the median discovery significance. volume 42 of *Procs of Machine Learning Research*, 2015.

[14] L. Mackey, J. Bryan, and Y M Mo. Weighted classification cascades for optimizing discovery significance in the higgsml challenge. volume 42 of *Procs of Machine Learning Research*, 2015.