Direct optimization of the discovery significance in machine learning for new physics searches in particle colliders

A Elwood[1], D Krücker[1], M Shchedrolosiev[2]

[1]Deutsches Elektronen-Synchrotron, Notkestr. 85, 22607 Hamburg, Germany
[2]Taras Shevchenko National University of Kyiv, 03127, Kyiv, Ukraine

E-mail: dirk.kruecker@desy.de

Abstract. We introduce two new loss functions designed to directly optimize the statistical significance of the expected number of signal events when training neural networks and decision trees to classify events as signal or background. The loss functions are designed to directly maximize commonly used estimates of the statistical significance, $s/\sqrt{s+b}$, and the so-called Asimov estimate, Z_A. We consider their use in a toy search for Supersymmetric particles with 30 fb^{-1} of 14 TeV data collected at the LHC. In the case that the search for this model is dominated by systematic uncertainties, it is found that the loss function based on Z_A can outperform the binary cross entropy in defining an optimal search region. The same approach is applied to a boosted decision tree by modifying the objective function used in gradient tree boosting.

1. Introduction

The optimal way to design a search for new particles is a typical problem in particle physics. With the widespread use of machine learning techniques, it is necessary to reconsider the common approaches to this problem. This is directly related to the statistical question how to claim a discovery which has been discussed in a seminal paper on likelihood-based tests of new physics [1]. Within this approach, it is possible to define the median experimental sensitivity of a search. Following the wording of the authors in [1], we call this the *Asimov* discovery significance.

The idea to use the *Asimov* discovery significance as a metric in machine learning (ML) is not new. For example, it has been used for the *Higgs Boson Machine Learning Challenge* [2], in this context named AMS. An important feature of this kind of significance measure is that it allows the inclusion of systematic uncertainties. The main difference of the approach presented here is the use of the Asimov significance as the loss function in the training of a classifier, in contrast to being used only as a metric. Starting with the concept that the trained classifier will cut out an area in phase space, where a certain amount of signal and background events are observed, we aim to find the best region such that the discovery significance for Poisson distributed events becomes maximal. For the case of Poisson distributed background (b) and signal (s) events with background uncertainty σ_b, the approximated median discovery significance becomes:

$$Z_A = \left[2 \left((s+b) \ln \left[\frac{(s+b)(b+\sigma_b^2)}{b^2 + (s+b)\,\sigma_b^2} \right] - \frac{b^2}{\sigma_b^2} \ln \left[1 + \frac{\sigma_b^2 s}{b(b+\sigma_b^2)} \right] \right) \right]^{1/2}. \tag{1}$$

In the case where the background is known exactly ($\sigma_b = 0$) this simplifies to:

$$Z_A(\sigma_b = 0) = \sqrt{2\left((s+b)\ln(1+s/b) - s\right)}, \tag{2}$$

and for the case where s, $\sigma_b^2 \ll b$ we are left with $s/\sqrt{b+\sigma_b^2}$. The systematic uncertainty σ_b^2 is assumed to be proportional to b and is given as a percentage on the background for the results presented here. This is in contrast to the AMS, where a constant term $b_r = 10$ is added. We note, that such a term would prevent us from finding a purely background free region. In addition, we use $s/\sqrt{s+b}$ in the following, which can be understood as the exclusion significance in the large sample limit. To calculate the uncertainties on the significance, we propagate the uncertainties in the significance approximations and assume Poisson uncertainties for the number of selected events before applying physical event weights.

We would like to note that our approach can be related to other attempts to use the Neyman-Pearson Lemma [3] as starting point for a statistic aware classifier training [4] and [5].

2. Toy SUSY model

Table 1. Basic parameters of the two SUSY scenarios (left). Summary of all low level and high level variables used in this analysis (right).

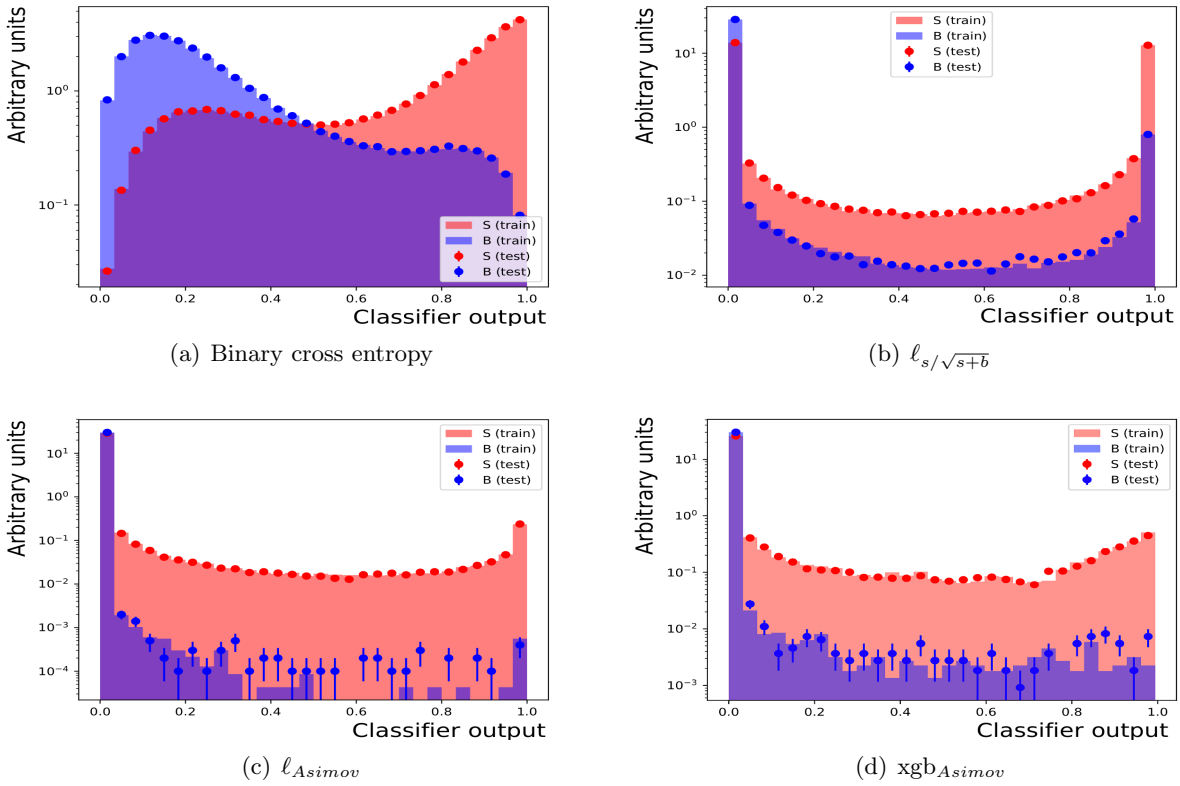
model	m_{stop}	m_{LSP}	σ
uncomp.	900 GeV	100 GeV	228 fb
comp.	600 GeV	400 GeV	17.6 fb

variable	names
low level	p_{l}, $p_{jet(1,2,3)}$, n_{jet}, $n_{\mathrm{b}\,jet}$
high level	$\not{E}_{\mathrm{T}}$, H_{T} m_{T}, $m_{\mathrm{T2}}^{\mathrm{W}}$

As a physics example, we consider a toy search for supersymmetric top quarks (stops) at the LHC with a single lepton in the final state. The search is designed for the case of direct stop pair production with subsequent decays of each squark into a top quark and a neutralino, the lightest supersymmetric particle (LSP) in this model. We assume that the top squark and the LSP are the only SUSY particles that have low enough masses to be accessible at 14 TeV and consider two model scenarios depending on the mass difference between stop and LSP. In the case that the stop is much heavier than the LSP, the SM decay products are produced with significant energy, making them easy to observe in the detector. These models are known as *uncompressed*. If the mass difference between the stop and LSP is small, the SM decay products are produced close to rest, making the signature difficult to distinguish from the background. These models are known as *compressed*. For our toy study, we assume two sets of mass parameters (table 1) that are expected to be on the border of discovery with about 30 fb^{-1} of 14 TeV LHC data.

We consider only the dominant background in the one-lepton channel, top-antitop ($\mathrm{t\bar{t}}$) production. A leading-order Pythia8 [6, 7] simulation for signal and background is sufficient for our toy studies with next-to-leading order results for the total cross sections (844 fb for $\mathrm{t\bar{t}}$, details and references are given in [8, 9]). Delphes3 [10] is used to model the detector response using the CMS detector model. Jets are clustered with the anti-k_T algorithm [11] with a cone parameter of $R = 0.4$. For simplicity, we do not consider tau leptons and use *lepton* as a generic term for electrons and muons.

Signal and background events are pre-selected to reduce training times. We require at least one lepton with transverse momentum $p_{\mathrm{T}} > 30$ GeV within a pseudorapidity of $|\eta| < 2.4$. Each event must contain at least four jets with $p_{\mathrm{T}} > 40$ GeV, where the highest-p_{T} (leading) jet is required to have $p_{\mathrm{T}} > 80$ GeV, and the sub-leading jet $p_{\mathrm{T}} > 60$ GeV. At least one of the jets must be tagged as originating from a bottom quark. The missing energy perpendicular to the



(a) Binary cross entropy

(b) $\ell_{s/\sqrt{s+b}}$

(c) ℓ_{Asimov}

(d) xgb_{Asimov}

Figure 1. Histograms of the classifier outputs for signal (red) and background (blue) for the testing (dots) and training (solid) datasets. (a) Neural network with cross entropy training. (b) Neural network with significance-loss training. (c) Neural network with Asimov-significance loss training. (d) Gradient boosted decision tree with Asimov-significance loss.

beam direction ($\not{E}_{\mathrm{T}}$) is required to be above 200 GeV, and the scalar sum over the transverse momenta of all preselected jets (H_{T}) to be above 300 GeV. The selected sample, which is used for training and testing, consists of 1.4×10^6 events with 50% signal and 50% background events. An independent sample of 6×10^5 events is used in the final evaluation step.

We include both *low level* and *high level* variables when training the classifiers. The low level variables consist of the multiplicities of jets (n_{jet}) and b-quark jets ($n_{\mathrm{b-jet}}$) and the Lorentz vector components (E, p_{T}, ϕ and η) of the reconstructed physics objects. A summary of all used variables are listed in table 1. More details about this stop toy search can be found in [8, 9].

3. Neural network implementation

We consider two commonly used approximations of statistical significance, $s/\sqrt{s+b}$ and the Asimov estimate, introduced in section 1. The two estimates are defined for a training batch, where s is the number of correctly classified signal events and b is the number of background events classified as signal. When used as a metric, a certain cut is applied to the classifier probability and the discrete number of events above this cut is counted. When used as a loss function, differentiability must be ensured. The new idea introduced here is to define the values

Table 2. SUSY toy model results for the neural network. All numbers are derived from an independent evaluation sample. The Asimov significance Z_A is used as metric in all cases with a cut on the classifier output that maximizes Z_A.

Uncomp. mod.	s	b	$Z_A(10\%)$	s	b	$Z_A(30\%)$	s	b	$Z_A(50\%)$
Loss:									
cross entropy	8.7	1.7	4.5 ± 0.3	7.7	1.2	4.0 ± 0.3	7.7	1.2	3.5 ± 0.3
ℓ_{Asimov}	6.6	1.6	3.6 ± 0.2	3.5	0.1	4.0 ± 0.5	3.3	0.1	4.2 ± 0.7
Comp. mod.	s	b	$Z_A(10\%)$	s	b	$Z_A(30\%)$	s	b	$Z_A(50\%)$
Loss:									
cross entropy	74.4	18.2	10.7 ± 0.3	44.0	7.7	6.8 ± 0.3	40.5	6.8	4.8 ± 0.3
ℓ_{Asimov}	78.4	19.4	10.8 ± 0.3	25.9	3.2	6.8 ± 0.4	11.9	0.5	6.2 ± 0.6

of s and b as smooth functions of y_i^{pred}, the classifier output modelled by a single sigmoid.

$$s = W_s \sum_i^{N_{batch}} y_i^{pred} \times y_i^{true}, \qquad b = W_b \sum_i^{N_{batch}} y_i^{pred} \times (1 - y_i^{true}), \tag{3}$$

where N_{batch} is the number of events in each batch and y_i^{true} is the true value, 1 for signal and 0 for background. W_s and W_b are the physical weights of the signal and background samples. They scale up the number of signal and background events per batch to the numbers expected to be observed in the experiment given a certain luminosity. The above definitions ensure that s and b are continuous w.r.t. the network parameters, as it is needed for backpropagation. It can be seen that in the limit of an absolute certainty in the value of y^{pred}, either 1 or 0, s and b converge to a discrete definition. With this definition of s and b, two loss functions are defined as the squared inverse of the two approximations of significance that are considered here:

$$\ell_{s/\sqrt{s+b}} := (s+b)/s^2, \qquad \ell_{Asimov} := 1/Z_A, \tag{4}$$

where Z_A is defined in (1). Minimizing one of these loss function corresponds to maximizing the corresponding significance.

This approach had been implemented with the Keras python library [12]. To avoid strong statistical fluctuations of the estimated significance, a sufficient number of events in a single batch is necessary. It was found empirically that a batch size of 4096 works well. The results shown here are evaluated from a network with a single hidden layer of 23 neurons, as it had the least sensitivity to overtraining and adequate performance for the simple toy model.

4. Implementation in gradient boosting

The Asimov-loss approach can also be used to train a Boosted Decision Tree (BDT). XGBoost [13] uses a Taylor expansion on an event-wise objective function during tree building. The default loss function, cross-entropy, can be replaced with the Asimov loss and the "smoothing" trick (3). The Taylor expansion of $1/Z_A$ forms a sum over events in this case which can be implemented to first order, replacing the gradient in XGBoost by:

$$g_i^{Asimov} = -Z_A^{-2} \left(\frac{\partial Z_A}{\partial s} W_s y_i + \frac{\partial Z_A}{\partial b} W_b \left(1 - y_i\right) \right). \tag{5}$$

The gradient has been implemented with Autograd [14]. We note that this approach does not allow the implementation of the Hessian, which is therefore assumed to be a constant.

5. Results and conclusions

The specific behaviour of the new Asimov loss functions can be clearly seen in the classifier output distributions in figure 1: the default cross-entropy training (a) aims to minimize the classification error for signal and background equally, while the Asimov-loss training (c) maintains the purity of the signal classification at the expense of signal events which are misclassified as background. Significance is gained at the cost of reduced efficiency. The Asimov-loss training tries to find an almost background free region (classifier output > 0.5) in both implementations (c) and (d). It is important to note that the simulation statistics must be large enough to prevent the algorithm from just finding random holes in the trainings data due to low statisics. We account for this by large sample sizes, proper error calculation and independent evaluation samples. For $s/\sqrt{s+b}$, (b) in figure 1, the advantage to finding a background free region lower. This loss function finds a different optimum, which is also not identical to the optimum with the highest classification accuracy. Optimizing for discovery usually means finding an area in phase space with high signal-to-background ratio, even at the cost of efficiency, while optimizing for exclusion usually requires a better signal acceptance even with more background. Our algorithm supports both strategies.

Numerical results for the SUSY toy model with the neural network implementation are given in table 2. More results can be found in [8]. The results for the cross-entropy training are also optimized for highest Asimov significance in the conventional way of applying a cut to the classifier output. This approach forces the numerical results to be similar in most cases. Clear differences appear at high systematic uncertainties. For example, at 50% systematic uncertainty in the compressed case the Asimov-loss training performs better. This shows that it can be important to include systematic uncertainties into the training.

The validity of the approximations used in (1) for small s and b had been confirmed numerically in [1]. We intend to invest further work in the low s and b region ($s, b < 2 - 3$), to evaluate the performance of the classifier in such cases. Also the comparison with a full limit setting will be considered.

5.1. Acknowledgments

We would like to thank our DESY colleague Oleksii Turkot for his support during the BDT studies.

References

[1] Cowan G, Cranmer K, Gross E and Vitells O 2011 *Eur. Phys. J.* C **71** 1554
[2] Cowan G, Germain C, Guyon I, Kégl B and Rousseau D 2015 *Proc. of the NIPS 2014 Workshop on High-energy Physics and Machine Learning,* J. Mach. Learn. Res.: Workshop and Conf. Proc. **42** 19
[3] Neyman J and Pearson E S 1933 *On the problem of the most efficient tests of statistical hypotheses*, Phil. Trans. R. Soc. Lond. A. **231** 289
[4] Brehmer J, Cranmer K, Louppe G and Pavez J 2018 *Phys. Rev.* D **98** 052004 and ref. therein
[5] De Castro P and Tommaso D 2018 *INFERNO: Inference-Aware Neural Optimisation*, arXiv:1806.04743 [stat.ML]
[6] Sjöstrand T et al. 2015 *Comput. Phys. Commun.* **191** 159
[7] Sjöstrand T Mrenna A and Skands P Z 2006 *JHEP* **0605** 026
[8] Elwood A and Krücker D 2018 *Direct optimisation of the discovery significance when training neural networks to search for new physics in particle colliders*, arXiv:1806.00322 [hep-ex]
[9] Sahin O M, Krücker D and Melzer-Pellmann I A 2016 *Nucl. Instrum. Meth.* A**838** 137
[10] de Favereau J et al. 2014 *JHEP* **02** 057
[11] Cacciari M, Salam G P and Soyez G 2008 *JHEP* **04** 063
[12] Chollet F et al. 2015, https://keras.io
[13] Chen T and Guestrin C 2016 *Proc. 22nd ACM SIGKDD Int. Conf. on KDD*, p. 785
[14] Maclaurin D 2016 *Modeling, Inference and Optimization with Composable Differentiable Procedures*, PhD thesis; https://github.com/HIPS/autograd