ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Systematics aware learning: a case study in High Energy Physics

Victor Estrade[1] and Cécile Germain[1] and Isabelle Guyon[1] and David Rousseau[2]

1. LRI, TAO, UPSud. 2. IN2P3, LAL.
Université Paris Saclay, France

Abstract. Experimental science often has to cope with systematic errors that coherently bias data. We analyze this issue on the analysis of data produced by experiments of the Large Hadron Collider at CERN as a case of supervised domain adaptation. Systematics-aware learning should create an efficient representation that is insensitive to perturbations induced by the systematic effects. We present an experimental comparison of the adversarial knowledge-free approach and a less data-intensive alternative.

1 Introduction

The 2010s extolled the so-called 4th paradigm of science: data-intensive scientific discovery. A subsequent trend is the usage of Machine Learning techniques to actually make use of collected data. On the other hand, since its inception, experimental science has concerned itself with coping with uncertainties. We explore how these two paradigms interact in a specific setting, which is *simulation based* experimental science, and specifically on the impact of data bias.

In order to be concrete, we focus on a demanding example: the analysis of the data produced by the experiments of the LHC (Large Hadron Collider) at CERN, e.g. the measurement of the characteristics of new particles, such as the Higgs boson. An essential component of this analysis is a procedure for selecting a region of interest in the space of measured features. Multivariate classification has become the standard tool to optimize the selection region. As such, the classifier is an integral part of the measurement apparatus. For discovery and measurement of a new particle such as the Higgs boson, by definition no labeled real data are available. The classifier has to be trained on simulated data [1].

An experiment includes two sort of uncertainties: the uncertainty resulting from unknown random effects that we shall call **"statistical uncertainty"**, and that resulting from known nuisance factors that we shall call **"systematic uncertainty"** . A systematics-aware learning procedure should optimize the trade-off between both, by learning an efficient representation, which is insensitive to perturbations induced by systematic effects. Two strategies are considered in this paper, stemming from the related field of "domain adaptation" [2]: (1) a knowledge-free setting, where the invariant representation is discovered from the data (adversarial supervised learning); or (2) the integration of prior knowledge. The main goal of this paper is to evaluate the two approaches.

2 Systematics in simulation-based analysis

We briefly describe the problem setting we are interested in as a machine learning problem, without dwelling into the specifics of the underlying physics.

Classification. We address a two-class problem: *signal* (S) vs. *background* (B). The S class in the "positive" class. To cope with class imbalance (elements of class S are very rare), the simulator produces an even number of examples of the two classes with importance-weights. This allows us to train the classifier on a balanced dataset while taking weights into account for performance evaluation. The quantities of interest are the weighted true and false positive counts where the w_i are the weights and t is the classification threshold of the discriminant value $score_i$:

$$s = \sum_{S, score_i > t} w_i \qquad \text{and} \qquad b = \sum_{B, score_i > t} w_i .$$

In the simplest machine learning settings, data are drawn *i.i.d.* from an unknown, but fixed data distribution. The generalization error in that case combines the modeling error (due to model bias/model finite capacity and finite sample size) and the intrinsic error (lowest achievable by the Bayes optimal classifier). This is termed *statistical error*, although the capacity limitation truly is a systematic error. The type of *systematic error* studied in this paper comes from departures of the data from the classical *i.i.d.* assumption in the following way: test data may be differently distributed from training data due to known nuisance factors (noted Z), the effect of which is bounded by known values. This coincides with the well known "domain adaptation" problem in machine learning, under the covariate shift paradigm [3], as it coherently biases data.

Figure of merit. Without fully justifying it, we provide in this section the figure of merit used by particle physicists, which is not the classification accuracy, but a non-linear function of the number of true and false positives. We call "signal region" the region of input space classified as S by the classifier. Particle detection in physics boils down to counting the number of events (data samples) falling into the signal region, called μ and determining its statistical significance. To that end, it must be compared to the uncertainty. Let s_z and b_z be the weighted counts of true and false positives with systematics at $Z = z$. The figure of merit is the relative error:

$$\frac{\sigma_\mu}{\mu} = \sqrt{\sigma_{\text{sta}}^2 + \sigma_{\text{sys}}^2}, \tag{1}$$

where $\sigma_{\text{sta}} = s_0^{-1}\sqrt{s_0 + b_0}$ is the relative statistical error and $\sigma_{\text{sys}} = s_0^{-1}(s_z + b_z - s_0 - b_0)$ the relative systematic error. The nuisance parameter Z is 0 in the nominal case.

This expression can be formally derived with the profile likelihood ratio method [4]. An intuitive explanation works as follows. Let N be the selected events, *i.e.* $N = s + b$. N is assumed to follow a Poisson distribution. The measurement μ is proportional to number of signal events, and normalized to the expected one from the standard model, that is $\mu_0 = 1$. The systematic error is by definition $\mu_z - \mu_0$. Because the value of the Nuisance Parameter is unknown, the best estimate of the number of signal events at $Z = z$ is $N_z - b_0$, obtained by subtracting the nominal number of background events, yieding:

$$\sigma_{\text{sys}} = \frac{s_z + b_z - s_0 - b_0}{s_0}$$

The statistical error measures the impact of the intrinsic deficiencies of the selection procedure. In the real experiment, all we have is N, thus the relative statistical error is the ratio of the Poisson variance to the nominal estimate of the number of signals:

$$\sigma_{\text{sta}} = \frac{\sqrt{s_0 + b_0}}{s_0}.$$

The dataset. In the experiments, we use the Enhanced Higgs Boson to $\tau^+\tau^-$ dataset [5] [1] which correspond to nominal data. The software to compute the impact of the Nuisance parameter on the simulation is described in [6]. In the following this will be called *skewing* the data.

3 Adversarial learning and Systematics

Learning with systematics is related to domain adaptation [2].However, classical domain adaptation usually addresses the semi-supervised setting [7] while learning with systematics is fully supervised: with simulations, at training time we have all the labels, and even the values of the nuisance parameter. We compare several approaches, from the “naive” *data perturbation* method to elaborate supervised learning procedures learning representations that do not contain information that is domain-specific (or nuisance parameter-dependent):

Data Perturbation or Data Augmentation. Training on a mix of data generated by varying the nuisance parameter in an adequate range. With enough training data and a classifier of sufficient capacity, the algorithm should discover the invariant manifold in data space.

Adversarial Learning. Learning to classify S *vs.* B while “unlearning” to predict Z. This purely data driven approach is based on the Generative Adversarial Network (GAN) framework [8].The Pivot Adversarial Network [9] exemplifies this approach for the HEP (High Energy Physics) case we are interested in. The Pivot Network is a GAN where the generated data is the distribution of classification score.It enjoys the same theoretical optimality results in an equal opportunity [10] model. The adversary tries to predict Z from the classifier output. This method requires a large training set that is representative of the full range of the data distribution, from nominal to perturbed.

Tangent Propagation. Regularizing with the partial derivative of the classifier score *w.r.t.* the nuisance parameter [11]: the smaller the derivative, the less sensitive the classifier. This is a method incorporating *a priori* knowledge of the impact of systematics as a coherent geometric transform in feature space. The systematics must be a differentiable transform $f(x, Z)$ of the inputs. This approach requires much less data than data perturbation/augmentation or adversarial learning.

[1] available on UCI at `http://mlphysics.ics.uci.edu/data/htautau/`

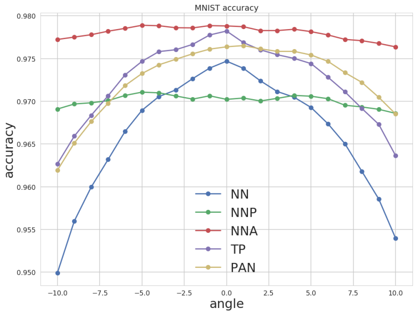


Figure 1: Accuracy on the MNIST dataset as a function of rotation angle.

4 Performance evaluation

Our experiments compare as classifier a plain Deep **Neural Network (NN)** with several neural network of the same architecture trained to be robust to systematic errors: **Neural Net w. Perturbed data (NNP)** with an identical size training set; **Neural Net w. Augmented data (NNA)**, similar to NNP but with 6 times more training data; **Pivot Adversarial Network (PAN)**, and **Tangent Propagation (TP)**.

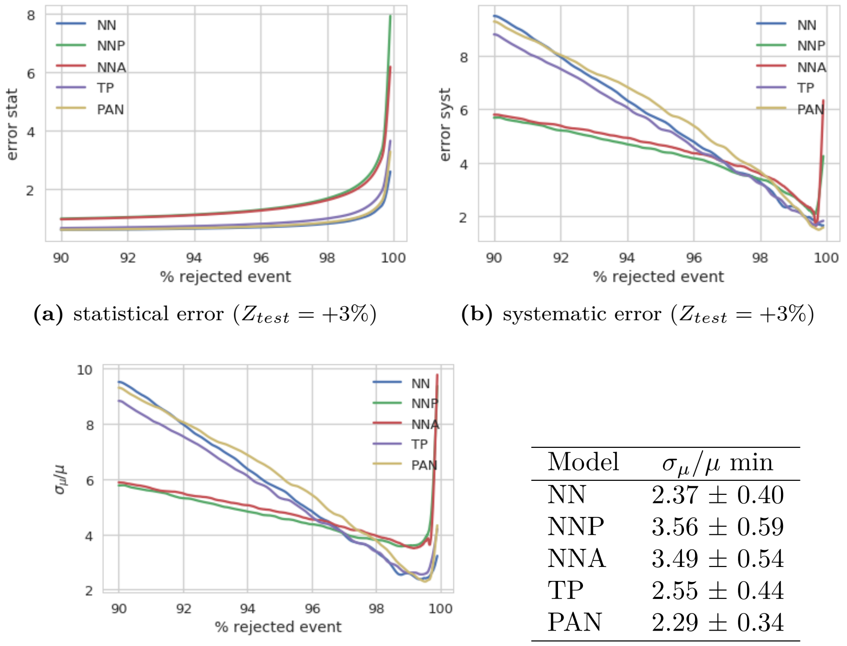
Experimental setting. In order to make the comparison manageable, the dimensioning hyper parameters are identical for all DNN architectures, with 3 hidden layers of 120 neurons each. The networks were all trained for 10000 iterations with a mini-batch size of 1024 and optimized with Adam method. Softplus and ReLU activations were tried and gave the same results.

NN is trained on the nominal dataset only. Pivot and Data Perturbation/Augmentation are trained on a mix of skewed data, with Z drawn from a normal distribution $\mathcal{N}(0, k10^{-2})$, with k can be 1, 3 or 5. TP is trained on the nominal dataset with the tangent vector initialized by the finite difference method. In all cases, the systematics are introduced in the test set with a fixed skewing, because the nuisance parameter is well defined, although unknown. All presented values are the mean and standard deviation of a 12-random-split cross validation (80% training, 20% testing).

MNIST benchmark. Figure 1 compares the accuracy for the multiclass classification problem the classical MNIST dataset [12]. Z is a rotation of a given angle of the digits. To compute the tangent vector, a smoothing is applied to the rotated image, as in [11]. NNA is best showing that data augmentation works, but at the price of training on more data. NNP shows robustness to systematic error, but obtains degraded performance on the original Z=0 distribution. TP performs better: it rips the full benefit of the a-priori knowledge of the geometric transformation, despite the fact that the manifold of rotations is likely to have high curvature [13]. PAN also shows good performances: it learns the invariance thanks to the adversarial training but TP is better for small angles.

The HEP benchmark. We report the errors and the overall figure of merit

σ_μ/μ not only at the optimum, but along the decision threshold t in a physically plausible region (expressed by the fraction of rejected events), in order to capture its behavior along the sensitivity/specificity trade-off.



(a) statistical error ($Z_{test} = +3\%$)

(b) systematic error ($Z_{test} = +3\%$)

(c) combined error ($Z_{test} = +3\%$)

Model	σ_μ/μ min
NN	2.37 ± 0.40
NNP	3.56 ± 0.59
NNA	3.49 ± 0.54
TP	2.55 ± 0.44
PAN	2.29 ± 0.34

(d) combined error minima

Figure 2: Performances along decision threshold for the HEP benchmark.

Figure 2 shows the complex tradeoff between the systematic and statistical errors. There are two different regimes. In the left part, up to $\sim 98\%$ of the examples were classified as negatives, $\sigma_{\rm sys}$ is much larger than $\sigma_{\rm sta}$. As expected, Tangent Propagation and Data Perturbation/Augmentation degrade the statistical error *w.r.t.* the basic NN (remember that the statistical error is computed on the nominal, not skewed, data). The degradation is worse for the later, but with limited impact as they both improve on NN on systematics. However, the interesting region is the right area, close to the minimum, where the classification threshold should be selected [14]. Here, the two errors are comparable. Data Perturbation/Augmentation is heavily penalized by statistical error, while the three others are very close, with a small penalty for Tangent Propagation. The Wilcoxon ranksum test (at the 95% confidence level) applied to the fold-wise minima confirms this ranking. In both regions, Pivot is close to the basic NN, hinting that the adversarial component did not succeed in learning an adequate target function.

5 Conclusion

This paper has presented a case study of data bias in simulation-based experimental science, and related it to the classical concepts of domain adaptation, systematic error and nuisance parameters. The problem consists of learning a representation that is insensitive to perturbations induced by nuisance parameters. The need for the adversarial techniques, assuming a completely knowledge-free approach, has been questioned. Our results contrast the superior performance of data augmentation or incorporating a priori knowledge −Tangent Propagation (TP) approach – on a well separated classes (MNIST data) with a real case setting in HEP. In the first case Augmentation/TP dominates. In the latter case, a plain neural network performs best. Adversarial learning with the Pivot method does not show a significant advantage in either case but never performs worse than the baseline.

References

[1] Adam-Bourdarios et al. The Higgs boson machine learning challenge. In *HEPML@ NIPS*, pages 19–55, 2014.

[2] Ben-David et al. A theory of learning from different domains. *Journal of Machine Learning*, (1-2):151–175, 2010.

[3] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[4] Cowan et al. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71:1554–1573, 2011.

[5] Baldi et al. Enhanced higgs boson to $\tau^+\tau^-$ search with deep learning. *Phys. Rev. Lett.*, 114:111801, Mar 2015.

[6] Estrade et al. Adversarial learning to eliminate systematic errors: a case study in hep. In *Deep Learning for Physical Sciences @ NIPS*, 2017.

[7] Ganin et al. Domain-Adversarial Training of Neural Networks. *arXiv:1505.07818 [cs, stat]*, May 2015. arXiv: 1505.07818.

[8] Goodfellow et al. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.

[9] G. Louppe, M. Kagan, and K. Cranmer. Learning to Pivot with Adversarial Networks. *arXiv:1611.01046 [physics, stat]*, November 2016.

[10] Hardt et al. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016.

[11] Simard et al. Tangent Prop - A Formalism for Specifying Selected Invariances in an Adaptive Network. In *NIPS*, pages 895–903. Morgan Kaufmann, 1991.

[12] Lecun et al. The MNIST database of handwritten digits.

[13] Y. Bengio and Y. LeCun. Scaling learning algorithms towards AI. In *Large-Scale Kernel Machines*. MIT Press, 2007.

[14] Gábor Melis. Dissecting the Winning Solution of the HiggsML Challenge. In *HEPML@ NIPS*, pages 57–67, 2014.