Weakly Supervised Classification For High Energy Physics

Lucio Dery

E-mail: ldery@stanford.edu

Abstract. We present a classification algorithm that applies the machine learning paradigm of Learning from Label Proportions (LLP) [1] to enable learning on unlabelled data. Our algorithm, Weakly Supervised Classification, receives as its only input the class proportions of batches of data but makes per-instance classification decisions matching the performance of fully supervised approaches. We apply our model to the problem of Quark-Gluon tagging and show that it is robust to underlying mismodelling of the simulated data unlike fully supervised learning.

1. Introduction

Classification problems abound in High Energy Physics; from Quark-Gluon tagging [2] and b-tagging [3] to boosted W tagging [4], there is a strong need to discriminate the types of particles produced by collision events. Traditionally, these classification problems are approached by training a model that takes in individual feature-label pairs and learns to generalize from the supervision provided to unseen instances. The feature-label pairs used for training are generated via high fidelity Monte-Carlo (MC) simulations that model physical processes at distances ranging from 10^{-25} meters all the way to the macroscopic dimensions of detectors.

Since these simulations are only approximate, the underlying physical processes are not always perfectly model and as a consequence, the performance of classifiers trained via this approach and applied to actual data on collision events are in some cases sub-optimal. Comparisons of tagging efficiencies between data and simulation show large $(10\% - 30\%)$ differences for b-tagging [5] quark/gluon tagging [6], boosted W tagging [7], and high p_T top quark tagging [8]. Since using collider data for supervised training is not an option due to the absence of labels, there is the need to develop unsupervised or weakly supervised approaches that directly leverage collider data.

In some cases, though the per-instance class labels for data are unknown, we know the relative proportions in which the classes occur in. For example, at a fixed order in perturbation theory, the probability for an outgoing parton to be a quark or a gluon depends on well-known parton distribution functions and matrix elements. Relying only on the weak supervision of class proportions and leveraging insights from the areas of Multiple Instance Learning (MIL) [9] and LLP we propose an algorithm that matches the performance of fully supervised classifiers. More specifically, given a batch of data where only the relative proportions of classes are known, we learn a per-instance mapping from features to labels under the constraint that statistics of the predicted labels in a batch match the expected proportions.

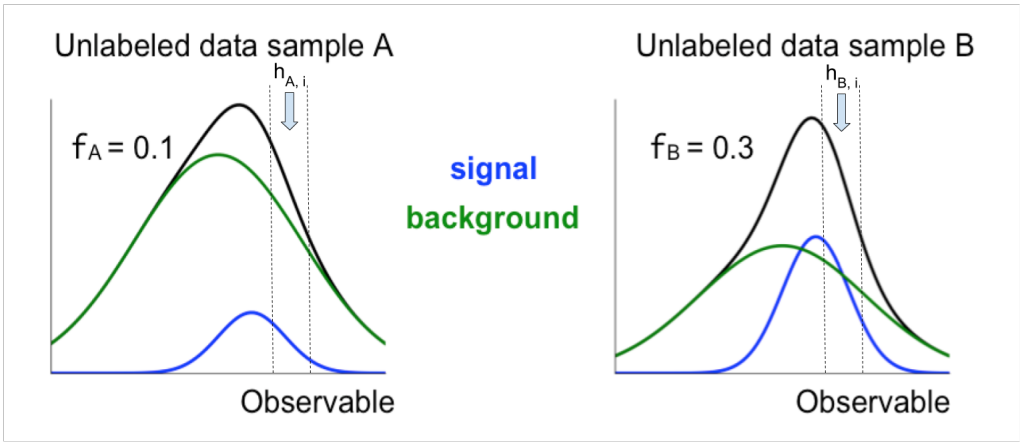


Figure 1: A rendering of the setup in section 2.1. The black distribution is obtained by binning the features of all examples, both signal and background, in a batch. f_A, f_B give the background to signal proportion in batches A and B respectively. The same bin i is chosen for in both histograms for resolving the true distributions of classes A and B

2. Method

Though proportions are a much weaker learning signal than per-instance labels, we will prove in the section that follows that it is a sufficient signal in finding a decision boundary that optimally classifies the data.

2.1. Probability Density Functions from Fractions

We show the feasibility of our approach by providing an analytic solution to a simple 2-batch problem where the class proportions are known. Consider two batches of data, A and B, each of which contains a mixture of two classes {0,1} in proportions f_A and f_B respectively. Each input is a feature vector $x_i \in \mathcal{R}^d$. In order to identify the two components, we need to resolve the Probability Density Functions (PDFs) corresponding to the individual classes. To do this, we first build the PDF corresponding to each of the individual batches. By considering m bins for each dimension d, we adopt a binning strategy, creating a 1-dimensional histogram (with $d \times m$ bins) representing a density function over feature bins. Since we know that each batch is a mixture of inputs from both classes, the height of a particular bin i is a linear combination of the true heights $h_{0,i}$ and $h_{1,i}$ for each class in their relative proportions. Figure 1 illustrates the setup. Using $h_{A,i}$ and $h_{B,i}$ as the height of bin i in batches A and B respectively, formally

$$h_{A,i} = f_A h_{0,i} + (1 - f_A) h_{1,i}$$
$$h_{B,i} = f_B h_{0,i} + (1 - f_B) h_{1,i}$$

Given the system of linear equations above, we can solve for the independent heights of each class in each bin i, $h_{0,i}$ and $h_{1,i}$ respectively. This allows us to extract the PDFs for each of the individual classes. Having obtained the PDFs for each class, we can easily build an optimal classifier that distinguishes the two classes based on the likelihood ratios.

The above procedure provides an analytic approach to learning from batch proportions. It is however not a practical approach. Specifically, the binning strategy is intractable when the

feature space is large. The system of equations set up also becomes over-constrained when more than two batches of proportions are provided, which means there might not exist an exact solution. We get around both these problems by using a neural network to approximate both the binning strategy and the optimal decision function.

2.2. Weakly Supervised Classifiers

The supervised binary classification problem is set up as follows. Given a training set $\mathcal{T}$, consisting of M_{train} tuples $(x_i, y_i)^{M_{train}}$ where x_i is a feature vector of an instance and y_i is the corresponding true label, we learn a classifier $\mathcal{F}_{sup}$ which generalizes to a unseen set of M_{test} examples $\mathcal{D} = \{x_j\}^{M_{test}}$ of unknown labels. To learn $\mathcal{F}_{sup}$ we set up a loss function l_{sup} that is independently evaluated for each member of a batch of data N.

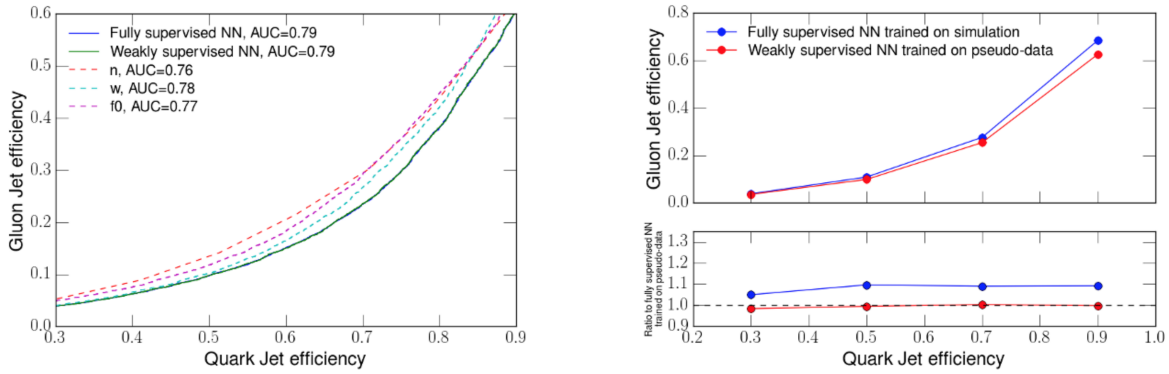
$$\mathcal{F}_{sup} = argmin_{\mathcal{F}} \sum_i l_{sup}\Big(\mathcal{F}(x_i), y_i\Big)$$

In the weakly supervised setup however, the training set consists of tuples (X_B^j, y_j) where X_B^j is the jth batch of examples containing M_{batch} instances. The label y_j in this case is the proportions of members of X_B^j that belong to class A. We still need the classifier to produce a per-instance binary decision though it is fed batched data. To achieve this, we introduce a loss that ties together the predictions of a batch. That is, $\mathcal{F}_{weak}$ makes a prediction for each member of the batch, but the loss, l_{weak} is evaluated on how well the distribution of batch predictions match the expected proportions y_j. More formally,

$$\mathcal{F}_{weak} = argmin_{\mathcal{F}} \sum_j l_{weak}\left(\frac{\sum_{x_i \in X_b^j} \mathcal{F}(x_i)}{M_{batch}}, y_j\right)$$

3. Results

To test our approach, we apply weak supervision to the quark-gluon tagging problem. We generate a dataset of $2 \to 2$ quark-gluon scattering (dijet) events, simulated using the Pythia 8.18 event generator [10]. Jets are clustered using the $anti - k_t$ algorithm [11] with distance parameter R = 0.4 via the FastJet 3.1.3 [12] package. Jets are classified as quark- or gluon-initiated by considering the type of the highest energy quark or gluon in the full generator event record that is inside a 0.3 radius of the jet axis. For simplicity, one transverse momentum range is considered: 45 GeV $<$ pT $<$ 55 GeV. Additionally, there is a pseudo-rapidity requirement that mimics the usual detector acceptance for charged particle tracking: $|\eta| < 2.1$. Heuristically, gluons have twice as much strong-force charge as quark jets, resulting in more constituents and a broader radiation pattern. More precisely, at leading logarithm the ratio of the average gluon and quark jet constituent multiplicity is equal to the ratio of the QCD color-factor (Casimir) associated with quarks (CF) and gluons (CA): $CA/CF = 9/4 \approx 2$ [13], [14]. Due to Casimir scaling, the ROC curve for any p_T and angle-weighted moments of the jet radiation pattern is set by CA/CF [15]. The following variables are useful for quark/gluon discrimination: the number of jet constituents n, the first radial moment in p_T (jet width) w, and the fraction of the jet p_T carried by the leading $anti - k_T$ R = 0.1 subjet f_0. The constituents i considered for computing n and w are the hadrons in the jet with $p_T > 500MeV$. A weakly supervised classifier with one hidden layer of size 30 is trained by considering 12 bins of the distribution of the absolute difference in pseudorapidity between the two jets [16]. The proportion of quark initiated jets varies between 0.21 and 0.32



(a) (b)

Figure 2: In (a) Fully and Weakly supervised classifiers are trained on identical simulated data and evaluated on a test sample drawn from the same population. The weakly supervised classifier matches the performance of the fully supervised one. The curves corresponding to the three input observables used as discriminant are shown as reference. In (b) Fully supervised classifier (blue line) is trained on a labeled simulated training sample. The weakly supervised classifier (red line) is trained on an unlabeled pseudo-data training sample. In both cases, the performance is evaluated on the same pseudo-data test sample. The ratios to the performance of a fully supervised classifier trained on a labeled pseudo-data sample are shown in the bottom pad

As can be seen from Figure 2(a) our classifier achieves the same performance as a fully supervised classifier. The key application we envision for Weak Supervision is the circumvention of the use of imperfect simulation data to build classifiers for high energy physics. To highlight this, we setup an experimental scenario to evaluate the performance of a fully supervised classifier trained on simulation but evaluated on pseudo-data against a weakly supervised classifier trained directly on pseudo-data. We build the pseudo-data samples by distorting the probability distributions of n and w in the training sample to emulate the difference in efficiency measured in [2] between simulation and data. As can be seen from Figure 2(b) weak supervision outperforms fully supervised, demonstrating ability to circumvent mis-modeling issues presented by training on simulated data.

The demonstrations provided so far have been on relatively low dimensional inputs. Weak supervision has also been applied to high dimensional jet images for the Q-G problem, achieving competitive results as in [17]

4. Conclusion

We have presented a new approach to classification in High Energy Physics in cases where class proportions are known but individual labels are not readily available. This weakly supervised classification has broad applicability and has been demonstrated in one important discrimination task in high energy physics: quark versus gluon jet tagging. Our method serves to lay the groundwork for circumventing the mis-modelling issues that plague training on data from simulations, allowing us to train directly on collider data.

References

[1] Giorgio Patrini, Richard Nock, Paul Rivera, and Tiberio Caetano. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, pages 190–198, 2014.

[2] Melissa Franklin, John Huth, Valerio Ippolito, David Lopez Mateos, Kevin Mercurio, Masahiro Morii, William Spearman, Andy Yen, and G Zevi Della Porta. Light-quark and gluon jet discrimination in pp collisions at $\sqrt{s} = 7$TeV with the atlas detector. 2014.

[3] Atlas Collaboration et al. Performance of b-jet identification in the atlas experiment. *Journal of instrumentation*, 11(04):P04008, 2016.

[4] Georges Aad, B Abbott, J Abdallah, O Abdinov, R Aben, M Abolins, OS AbouZeid, H Abramowicz, H Abreu, R Abreu, et al. Identification of boosted, hadronically decaying w bosons and comparisons with atlas data taken at $\sqrt{s} = 8$TeV. *The European Physical Journal C*, 76(3):154, 2016.

[5] CMS collaboration et al. Identification of b-quark jets with the cms experiment. *Journal of Instrumentation*, 8(04):P04013, 2013.

[6] CMS collaboration et al. Performance of quark/gluon discrimination in 8 tev pp data. *CMS Physics Analysis Summary CMS-PAS-JME-13-002, CERN*, 2013.

[7] CMS collaboration et al. Identification techniques for highly boosted w bosons that decay into hadrons. *arXiv preprint arXiv:1410.4227*, 2014.

[8] CMS collaboration et al. Boosted top jet tagging at cms. Technical report, CMS-PAS-JME-13-007, 2013.

[9] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997.

[10] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. A brief introduction to pythia 8.1. *Computer Physics Communications*, 178(11):852–867, 2008.

[11] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-kt jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063, 2008.

[12] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. Fastjet user manual. *The European Physical Journal C*, 72(3):1896, 2012.

[13] Jason Gallicchio and Matthew D Schwartz. Quark and gluon tagging at the lhc. *Physical review letters*, 107(17):172001, 2011.

[14] Keith A Olive, Particle Data Group, et al. Review of particle physics. *Chinese Physics C*, 38(9):090001, 2014.

[15] Andrew J Larkoski, Jesse Thaler, and Wouter J Waalewijn. Gaining (mutual) information about quark/gluon discrimination. *Journal of High Energy Physics*, 2014(11):129, 2014.

[16] Georges Aad, B Abbott, J Abdallah, O Abdinov, B Abeloos, R Aben, M Abolins, OS AbouZeid, NL Abraham, H Abramowicz, et al. Measurement of the charged-particle multiplicity inside jets from $\sqrt{s} = 8$TeV pp collisions with the atlas detector. *The European Physical Journal C*, 76(6):322, 2016.

[17] Patrick T Komiske, Eric M Metodiev, Benjamin Nachman, and Matthew D Schwartz. Learning to classify from impure samples. *arXiv preprint arXiv:1801.10158*, 2018.