Title

Authors

Editor:

Abstract

Abstract

1. Introduction

In High Energy Physics (HEP) and many other fields hypothesis testing is a key tool when reporting results from an experiment. Likelihood ratio test are the main technique for hypothesis testing and they are the most powerful statistic for simple hypothesis testing. For composite hypothesis testing the profile or generalized likelihood ratio test is commonly used. When computing the likelihood ratio the data distribution $p(x|\theta)$ must be evaluated, where θ are parameters of the probability distribution. However, it is common in HEP to have simulations that describe the distribution $p(x|\theta)$ while not having a description that can be directly evaluated. This simulations are used to obtain a high dimensional observation x by emulating the underlying physics of the process. Commonly it is impossible or computationally expensive to evaluate the likelihood ratio in this setting.

It has been recently proved that discriminative classifiers can be used to solve this problem by constructing an equivalent version of the likelihood ratio in this likelihood-free setting (Cranmer, 2015).

In this work we show how this results can be used to approximate the likelihood ratio when the underlying distribution is a weighted sum of probability distributions (mixture model). This case is common in physics when identifying a signal process over one or more background process. We also show that by training a set of classifiers in a pairwise manner on the components of the mixture model it is possible to considerably improve the results of the approximation. Finally, we demonstrate how this technique can also be used to estimate the unknown weights coefficients in the mixture model or equivalently the signal and background respective contributions.

This note is organized as follows. Section 2 gives a brief overview of how likelihood ratios are used at the LHC and how this ratios can be approximated by using supervised learning techniques. In Section 3 the method to approximate likelihood ratios on mixture models is explained and in Section 4 a complete study of the results obtained by the algorithm for several toy distributions is presented.

2. Approximating Likelihood Ratios using Discriminative Classifiers

A common use of likelihood ratios in HEP is signal process identification. In this task the hypothesis testing procedure is used to evaluate the signal process significance by contrasting the only background (null) hypothesis versus the signal plus background (alternative) hypothesis. In this setting, the underlying distribution can be seen as a signal and background mixture model defined as

$$p(x \,|\, \mu, \nu) = \mu p_s(x \,|\, \nu) + (1 - \mu)\, p_b(x \,|\, \nu), \tag{1}$$

where $p_x(x \,|\, \nu)$ correspond to the signal distribution and $p_b(x \,|\, \nu)$ is the background distribution, both parametrized by nuisance parameters ν which describe uncertainties in underlying physics predictions or response of measurement devices. The parameter μ is the mixture coefficient corresponding to the signal component of the distribution. In this case the generalized likelihood ratio test takes the form of

$$T(D) = \prod_{e=1}^{n} \frac{p(x_e|\mu = 0, \hat{\hat{\nu}})}{p(x_e|\hat{\mu}, \hat{\nu})}, \tag{2}$$

where D is a data set of i.i.d observations x_e, $\hat{\hat{\nu}}$ is the conditional maximum likelihood estimator for ν under the null hypothesis θ_0 ($\mu = 0$) and $\hat{\nu}, \hat{\mu}$ are the maximum likelihood estimators for ν and μ. This approach has been used extensively to assert the discovery of new particles in HEP (Cowan et al., 2010), such as in the discovery of the Higgs boson (The ATLAS Collaboration, 2012; The CMS Collaboration, 2012).

As previously mentioned, the original distributions for signal and background can only be approximated by forward simulations. Most of the likelihood ratio tests at the LHC are made on the distribution of a single feature that discriminate between signal and background observations. For this, the simulated data is used together with interpolations algorithms in order to approximate the parametrized model and then use it in the hypothesis testing procedure (Cranmer et al., 2012).

To obtain a discriminative feature between signal and background various experiments use supervised learning methods by training a discriminative classifier which can learn to classify between signal and background given the original high-dimensional simulated data. In HEP methods like boosted decision trees and multilayer perceptron has been implemented in libraries such as TMVA and used extensively (Hocker et al., 2007).

Noticeably, it has been shown that a discriminative classifier trained to classify between signal and background can be used to obtain an equivalent likelihood ratio test (Cranmer, 2015). Let $s(x; \theta_0, \theta_1)$ to represent the classification score learned by the classifier, parametrized by θ_0 and θ_1 the parameters of the statistical model, let $p(s(x; \theta_0, \theta_1)|\theta)$ the probability distribution of the score on the data from the original distribution $p(x|\theta)$. Then, the likelihood ratio test

$$T(D) = \prod_{e=1}^{n} \frac{p(x_e|\theta_0)}{p(x_e|\theta_1)}, \tag{3}$$

is equivalent to the test

$$T'(D) = \prod_{e=1}^{n} \frac{p(s(x_e;\theta_0,\theta_1)|\theta_0)}{p(s(x_e;\theta_0,\theta_1)|\theta_1)}, \tag{4}$$

where θ_0 correspond to the null hypothesis and θ_1 to the alternative hypothesis. The only requirement is that the discriminative classifier learn a monotonic function of the per event ratio $p(x_e|\theta_0)/p(x_e|\theta_1)$ (Cranmer, 2015). This is the usual case since many of the commonly used classifiers learn to approximate some monotonic function of the regression function $s(x) \sim p(y|x) = p(x|\theta_1)/(p(x|\theta_0) + p(x|\theta_1))$ which is monotonic to the desired ratio.

In the next section we show how this result can be used to decompose a likelihood ratio for mixture models into the likelihood ratio of its components obtained by training a set of classifiers pairwise.

3. Decomposed Likelihood Ratio Test for Mixture Models

A generalized version of the signal and background mixture model of Eq. 1 for several components is

$$p(x|\theta) = \sum_{i=1}^{k} w_i(\theta) p_i(x|\theta), \tag{5}$$

where $w_i(\theta)$ are the mixture coefficients for each one of the components parametrized by θ. In (Cranmer, 2015) it is shown that the likelihood ratio between two mixture models

$$\frac{p(x|\theta_0)}{p(x|\theta_1)} = \frac{\sum_{i=1}^{k} w_i(\theta_0) p_i(x|\theta_0)}{\sum_{j=1}^{k'} w_j(\theta_1) p_j(x|\theta_1)}, \tag{6}$$

is equivalent to the composition of pairwise ratios for each one of the components

$$\sum_{i=1}^{k} \left[\sum_{j=1}^{k'} \frac{w_j(\theta_1)}{w_i(\theta_0)} \frac{p_j(x|\theta_1)}{p_i(x|\theta_0)} \right]^{-1}, \tag{7}$$

and by Eq. 4 this is equivalent to the composition of ratios on the score distribution of pairwise trained classifiers

$$\sum_{i=1}^{k} \left[\sum_{j=1}^{k'} \frac{w_j(\theta_1)}{w_i(\theta_0)} \frac{p_j(s_{i,j}(x;\theta_0,\theta_1)|\theta_1)}{p_i(s_{i,j}(x;\theta_0,\theta_1)|\theta_0)} \right]^{-1}. \tag{8}$$

In the case that the only free parameters of the mixture model are the coefficients $w_i(\theta)$, then each distribution $p_i(s_{i,j}(x;\theta_0,\theta_1)|\theta)$ is independent of θ and can be pre-computed and used after in the evaluation of the likelihood ratio. Moreover, in this case each ratio

$p_j(s_{i,j}(x;\theta_0,\theta_1)|\theta_1)/p_i(s_{i,j}(x;\theta_0,\theta_1)|\theta_0)$ with $i=j$ can be replaced by 1. Also, for a two-class classifier the values of $s_{j,i}(x;\theta_0,\theta_1)$ for one of the classes can be replaced by the values of $s_{i,j}(x;\theta_0,\theta_1)$ for the opposing class, then it is only necessary to train the classifiers for $i<j$. This saves a lot of computation time and avoid possible variance that can be introduced by differences between $s_{i,j}(x;\theta_0,\theta_1)$ and $s_{j,i}(x;\theta_0,\theta_1)$ due to imperfect training.

In the common case of only background null hypothesis versus signal plus background alternate hypothesis, the coefficient $w_i(\theta_0)$ corresponding to the signal component will be equal to zero under the null hypothesis, while the coefficients on the alternate hypothesis will be all bigger than zero. Additionally, it is common for the signal coefficient $w_j(\theta_1)$ to be a very small number compared to the background coefficients under the alternate hypothesis. In this conditions a classifier trained in data from the full mixture model will have a lot of problems to identify the signal since most of the useful discriminative data will lay in a small region of the feature space.

It is possible to estimate the signal and background coefficients by using maximum likelihood estimation on the ratios, let $W(\theta_1)$ be the vector of coefficients under the alternate hypothesis, then for the pseudo data $D=\{x_i,\ldots,x_n\}$ generated from $p(x|\theta_1)$ and for a fixed value of θ_0

$$\hat{W}(\theta_1)=\arg\max_{W(\theta_1)}\prod_{e=1}^{n}\frac{p(x_e|\theta_1)}{p(x_e|\theta_0)}, \tag{9}$$

equivalently, the decomposed ratio can be used

$$\hat{W}(\theta_1)=\arg\max_{W(\theta_0)}\prod_{e=1}^{n}\left[\sum_{j=1}^{k'}\left[\sum_{i=1}^{k}\frac{w_j(\theta_0)}{w_i(\theta_1)}\frac{p_j(s_{j,i}(x_e;\theta_0,\theta_1)|\theta_0)}{p_i(s_{j,i}(x_e;\theta_0,\theta_1)|\theta_1)}\right]^{-1}\right]. \tag{10}$$

The complete algorithm to estimate the likelihood ratio using pairwise trained classifiers can be separated into three stages, classifier training, score distribution estimation and composition formula computation.

First, given some data sets $X_i,\ldots X_l$ generated from the component distributions $p_i(x),\ldots,p_l(x)$ (possibly by simulations), a set of classifiers is trained in each pair $[X_i,X_j]$ with $i<j$ where samples from X_i are labeled as signal and samples from X_j are labeled as background. It should be noted that the selection of classifiers and the training of this classifiers factorizes from the score distribution estimation and the computation of the composed ratio. The only requirement is that the trained classifier approximates to some degree a monotonic function of the regression function.

Each classifier $s_{i,j}(x;\theta_0,\theta_1)$ is used to estimate the score distribution $p(s_{i,j}(x;\theta_0,\theta_1)|\theta_0)$ and $p(s_{i,j}(x;\theta_0,\theta_1)|\theta_1)$ with x belonging to $X_i'(X_j')$ a dataset generated from $p_i(x)$ $(p_j(x))$ (possibly different from the one used in the training stage), by using an univariate density estimation technique such as histograms or kernel density estimation (Verkerke and Kirkby, 2003; Cranmer, 2001).

Finally, the estimated distributions $p(s_{i,j}(x;\theta_0,\theta_1)|\theta)$ are used in the composition formula and the values of the coefficients can be estimated using Eq. 10.

In the next section we will show how the method works in toy data, for a simple one-dimensional case and a harder multidimensional case, very similar to what can be found in real physics experiments.

4. Experiments

4.1 Univariate Case

4.2 Multivariate Case

References

Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *Eur.Phys.J.*, C71:1554, July 2010. doi: 10.1140/epjc/s10052-011-1554-0. URL `http://arxiv.org/abs/1007.1727`.

K. Cranmer. Approximating Likelihood Ratios with Calibrated Discriminative Classifiers. *ArXiv e-prints*, June 2015. `http://arxiv.org/abs/1506.02169`.

Kyle Cranmer, George Lewis, Lorenzo Moneta, Akira Shibata, and Wouter Verkerke. HistFactory: A tool for creating statistical models for use with RooFit and RooStats. 2012. CERN-OPEN-2012-016, `http://inspirehep.net/record/1236448`.

Kyle S. Cranmer. Kernel estimation in high-energy physics. *Comput. Phys. Commun.*, 136:198–207, 2001. doi: 10.1016/S0010-4655(00)00243-5.

Andreas Hocker, J. Stelzer, F. Tegenfeldt, H. Voss, K. Voss, et al. TMVA - Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

The ATLAS Collaboration. 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. doi: 10.1016/j.physletb.2012.08.020.

The CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys.Lett.*, B716:30–61, 2012. doi: 10.1016/j.physletb.2012.08.021.

Wouter Verkerke and David P. Kirkby. The RooFit toolkit for data modeling. *eConf*, C0303241:MOLT007, 2003.