Likelihood ratio tests constructed with discriminative classifiers and calibrated with generative models

Kyle Cranmer

November 18, 2014

1 Introduction

In many areas of science, likelihood ratio tests are established tools for statistical inference. Directly constructing the likelihood ratio for high-dimensional observations is often not possible or is computationally impractical. Here we demonstrate how discriminative classifiers can be used to construct equivalent likelihood ratio tests when a generative model for the data is available for calibration. We use the following notation

- x: a vector of features
- D: a dataset of $D = \{x_1, \ldots, x_n\}$, where x_e are assumed to be i.i.d.
- θ: parameters of a statistical model
- $f(x|\theta)$: probability density (statistical model) for x
- $s(x; \theta_0, \theta_1)$: real-valued discriminative classification score, parametrized by θ_0 and θ_1
- $g(s|\theta)$: The probability density for $s(x; \theta_0, \theta_1)$ implied by $f(x|\theta_0, \theta_1)$

We will assume the x_e are i.i.d., so that $f(D|\theta) = \prod_{e=1}^{n} f(x_e|\theta)$.

In the setting where one is interested in simple hypothesis testing between a null $\theta = \theta_0$ against an alternate $\theta = \theta_1$, the Neyman-Pearson lemma states that the likelihood ratio

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

is the most powerful test statistic. In order to evaluate $T(D)$, one must be able to evaluate the probability density $f(x|\theta)$ at any value x. However, it is increasingly common in science that one has a complex simulation that can act as generative model for $f(x|\theta)$, but one

cannot evaluate the density directly. For instance, this is the case high energy physics where the simulation of particle detectors can only be done in the 'forward mode'.

Our main result is that one can form an equivalent test based on

$$T'(D) = \prod_{e=1}^{n} \frac{g(s_e|\theta_0)}{g(s_e|\theta_1)} \tag{2}$$

if

$$s_e = s(x_e; \theta_0, \theta_1) = \frac{f(x_e|\theta_0)}{f(x_e|\theta_1)} \tag{3}$$

or some monotonic function of that ratio. This will be proven below. This allows us to recast the original likelihood ratio test into an alternate form in which a discriminative classifier is used to learn $s(x; \theta_1, \theta_0)$. The discriminative classifier can be trained with data $(x, y = 0)$ generated from $f(x|\theta_0)$ and $(x, y = 1)$ generated from $f(x|\theta_1)$.

While the original goal for frequentist hypothesis testing is to make a decision to accept or reject the null hypothesis based on the entire dataset D, the machine learning problem is an event-by-event classification problem. Of course, this follows from the fact that we assume the x_e to be i.i.d.

2 Dimensionality reduction

The target hypothesis test is based on

$$\ln T = \sum_{e=1}^{n} \underbrace{\log \left[\frac{f(x_e|\theta_0)}{f(x_e|\theta_1)} \right]}_{q(x_e)} . \tag{4}$$

Here we see that the optimal T for the experiment is composed of a sum over events of a linear function of the per-event function $q(x)$. A monotonic, but non-linear function of $q(x)$ would not lead to an equivalent hypothesis test.

The important part of the per-event function $q(x)$ is that it defines iso-contours in the feature space x. As we will show, our goal is to learn a monotonic function of $f(x|\theta_0)/f(x|\theta_1)$ that shares the same iso-contours. Then the remaining challenge is to find the appropriate rescaling that gives back linear function $q(x)$. Our claim is that the generative model $f(x|\theta)$ can be used to calibrate $g(s|\theta)$ and that

$$\ln T' = \sum_{e=1}^{n} \underbrace{\log \left[\frac{g(s_e|\theta_0)}{g(s_e|\theta_1)} \right]}_{q(s_e)} , \tag{5}$$

leads to an equivalent test. In particular, we need to show the density

$$f(q_x|\theta) = \int dx \delta(q_x - q_x(x)) f(x|\theta) / |\hat{n} \cdot \nabla q_x| \tag{6}$$

is the same as

$$f(q_s|\theta) = \int dx \delta(q_s - q_s(s(x)))\, f(x|\theta)\, /|\hat{n} \cdot \nabla q_s| \,. \tag{7}$$

It is sufficient to show that $q(x_e) = q(s(x_e))\ \forall x \in \Omega_c$.

For notational simplicity, let $f_0(x) = f(x|\theta_0)$, $f_1(x) = f(x|\theta_1)$, and $s(x) = s(x; \theta_1, \theta_0)$. The distribution of x totally determines the distribution of s via the change of variables $x \to s$. In the application at hand, the function s maps a high-dimensional feature vector x to $\mathbb{R}^+$. Let Ω_c be the level set $\{x \mid s(x) = c\}$ and $\hat{n} = \nabla s(x)/|\nabla s(x)|$ be the orthonormal vector to Ω_c at the point x. The induced density $g_1(c)$ is given by

$$g_1(c) = \int dx \delta(c - s(x)) f_1(x) = \int d\Omega_c f_1(x)/|\hat{n} \cdot \nabla s| \tag{8}$$

and a similar equation for $g_0(c)$.

Theorem 1: We have the following equalities

$$\frac{g_1(c)}{g_0(c)} = s(x) = \frac{f_1(x)}{f_0(x)} \qquad \forall x \in \Omega_c. \tag{9}$$

Proof We can factor out of the integral $s(x) = f_1(x)/f_0(x)$ since it is constant over Ω_c. Thus

$$g_1(c) = \int dx \delta(c - s(x)) f_1(x) = \int d\Omega_c f_1(x)/|\hat{n} \cdot \nabla s| = s(x) \int d\Omega_c f_0(x)/|\hat{n} \cdot \nabla s| \,, \tag{10}$$

and the integrals cancel in the likelihood ratio

$$\frac{g_1(c)}{g_0(c)} = \frac{s(x) \int d\Omega_c f_0(x)/|\hat{n} \cdot \nabla s|}{\int d\Omega_c f_0(x)/|\hat{n} \cdot \nabla s|} = s(x) = \frac{f_1(x)}{f_0(x)} \qquad \forall x \in \Omega_c. \tag{11}$$

In the case of simple hypothesis testing, θ_0 and θ_1 are specified and there is a unique map $s(x) = s(x_e; \theta_0, \theta_1)$. In that case, the equivalent likelihood ratio test can be performed by first transforming the data to $D_s = \{s_1, \ldots, s_e\}$, constructing the likelihoods

$$g(D_s \,|\, \theta) = \prod_{e=1}^{n} g(s_e \,|\, \theta) \tag{12}$$

for $\theta = \{\theta_0, \theta_1\}$, and constructing the likelihood ratio based on $g(D_s|\theta_0)/g(D_s|\theta_1)$.

3 Composite hypotheses and the generalized likelihood ratio

In the case of composite hypotheses $\theta \in \Theta_0$ against an alternative $\theta \in \Theta_0^C$, the generalized likelihood ratio[1] test is commonly used

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} f(D|\theta)}{\sup_{\theta \in \Theta} f(D|\theta)} \; . \tag{13}$$

This generalized likelihood ratio can be used both for hypothesis tests in the presence of nuisance parameters or to create confidence intervals with or without nuisance parameters. Often, the parameter vector is broken into two components $\theta = (\mu, \nu)$, where the μ components are considered parameters of interest while the ν components are considered nuisance parameters. In that case Θ_0 corresponds to all values of ν with μ fixed.

Denote the maximum likelihood estimator

$$\hat{\theta} = \arg\max_{\theta} f(D|\theta) \tag{14}$$

and the conditional maximum likelihood estimator

$$\hat{\hat{\theta}} = \arg\max_{\theta \in \Theta_0} f(D|\theta) \; . \tag{15}$$

It is not obvious that if we are working with the distributions $g(s|\theta)$ (for some particular $s(x; \theta_0, \theta_1)$ comparison) that we can find the same estimators. Fortunately, there is a construction based on $g(s|\theta)$ that works. The maximum likelihood estimate is the same as the value that maximizes the ratio with respect to $f(D|\theta_1)$ for some fixed value of θ_1. This allows us to use Theorem 1 to find

$$\hat{\theta} = \arg\max_{\theta} \frac{f(D|\theta)}{f(D|\theta_1)} = \arg\max_{\theta} \sum \ln \frac{f(x_e|\theta)}{f(x_e|\theta_1)} = \arg\max_{\theta} \sum \ln \frac{g(s(x_e; \theta, \theta_1)|\theta)}{g(s(x_e; \theta, \theta_1)|\theta_1)} \; . \tag{16}$$

It is important that we include the denominator $g(s(x_e; \theta, \theta_1)|\theta_1)$ because this cancels Jacobian factors that change as we vary θ.

4 Learning the correct mapping and its distribution

Thus far we have shown that likelihood ratio tests based on $f(x|\theta_0)/f(x|\theta_1)$ with high dimensional features x can be reproduced via hypothesis tests based on the univariate densities $g(s|\theta)$ for the very special dimensionality reduction map $s(x|\theta_0, \theta_1)$. The motivation for this is that often it is not possible to evaluate the density $f(x|\theta)$ at a given point x. This approach is not useful if it is not possible to approximate $s(x|\theta_0, \theta_1)$ and $g(s|\theta)$ without

[1] Also known as the profile likelihood ratio.

evaluating the density $f(x|\theta)$. In order for this approach to be useful, we need to be able to approximate both based on samples $\{(x, \theta)\}$ drawn from the generative model $f(x|\theta)$.

Denote the approximate dimensionality reduction map $\hat{s}(x; \theta_0, \theta_1)$ and its distribution $\hat{g}(\hat{s}|\theta)$. In general we will be interested in the machine learning problem that approximates these distributions based on samples $\{(x, \theta)\}$ drawn from the generative model $f(x|\theta)$. In particular, is there a loss function such that the function $\hat{s}(x)$ that minimizes the expected loss leads to a function that is one-to-one with $f(x|\theta_0)/f(x|\theta_1)$?

4.1 The standard discriminative classification setting

For fixed θ_0 and θ_1 we can generate large samples from each model and train a classifier. To be concrete, let's use $f(x|\theta_0)$ to generate training data $(x_i, \theta_i = \theta_0, y_i = 0)$ and $f(x|\theta_1)$ to generate training data $(x_i, \theta_i = \theta_1, y_i = 1)$. If we use the squared-loss function, then the expected loss is

$$\mathbb{E}[L] = \int dx f(x|\theta_0)(\hat{s}(x))^2 + \int dx f(x|\theta_1)(1 - \hat{s}(x))^2 \ . \tag{17}$$

The function $\hat{s}(x) = f(x|\theta_1)/(f(x|\theta_0) + f(x|\theta_1))$ minimizes this expected loss.

Proof: Consider a variation about $\hat{s}(x) = f(x|\theta_1)/(f(x|\theta_0) + f(x|\theta_1))$ given by $\hat{s}'(x) = \hat{s}(x) + h(x)$. The change in the expected loss is given by

$$\Delta = \mathbb{E}[L_{\hat{s}'}] - \mathbb{E}[L_{\hat{s}}] = \int dx f(x|\theta_0)(h^2(x) + 2h(x)\hat{s}(x)) + f(x|\theta_1)(h^2(x) - 2h(x) + 2h(x)\hat{s}(x) \ . \tag{18}$$

using $\hat{s}(x) = f(x|\theta_1)/(f(x|\theta_0) + f(x|\theta_1))$ we obtain

$$\Delta = \int dx (f(x|\theta_0) + f(x|\theta_1))h^2(x) > 0 \ . \tag{19}$$

Thus any variation on $\hat{s}(x) = f(x|\theta_1)/(f(x|\theta_0) + f(x|\theta_1))$ has a larger expected loss.

The conclusion is that standard classification with a quadratic loss function and training data as described above will approximate a discriminative classifier needed to produce an equivalent likelihood ratio test.

Once the classifier is trained, we can use the generative model and any univariate density estimation technique (e.g. histograms or kernel density estimation) to approximate $\hat{g}(x|\theta)$.

This treatment shows that in the asymptotic limit of samples from the generative model, that we can approximate arbitrarily well the original likelihood ratio test. With finite training data for $\hat{s}(x)$ and samples to approximate $\hat{g}(x)$ it will be necessary to be more specific about the what loss function we are interested in for approximating the likelihood ratio test. This will depend in general on the ultimate goal of the test and is left as a subject for future work.

5 Typical usage of machine learning in HEP

In high-energy physics (HEP) we are often searching for the properties of some class of events, generically referred to as *signal*, in the presence of a separate class of *background* events. For each event we measure some quantities x that have corresponding distributions $f_b(x|\nu)$ for background and $f_s(x|\nu)$ for signal. Often machine learning classification algorithms are trained on large samples of synthetic data $\{x_i, c_i\}$ generated with some nominal values of the parameters ν_0, where $c = 0$ corresponds to background and $c = 1$ corresponds to signal. The resulting classifier is denoted $s(x)$. Based on this classifier and large samples of synthetic data drawn from $f_c(x|\nu)$ we construct the distribution $g_c(s|\nu)$. An example of the distributions of the distribution of s for the signal and background events with $\nu = \nu_0$ is shown in Figure 1.

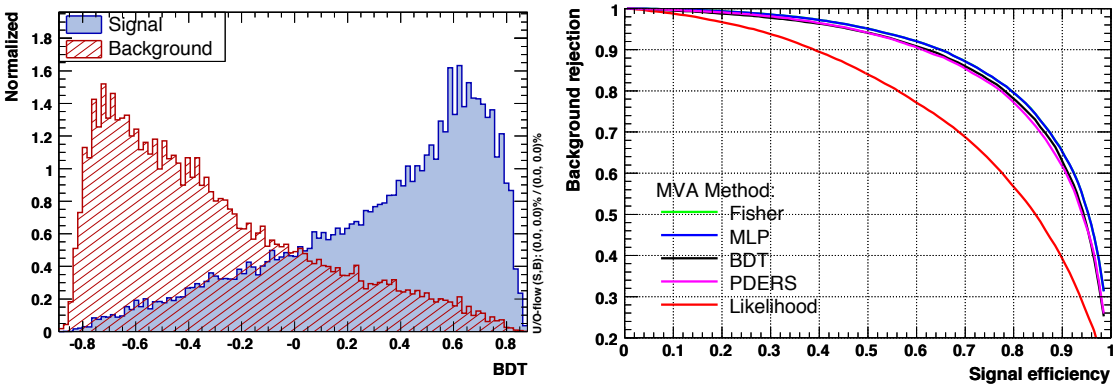

Figure 1: Left: an example of the distributions $f_b(s|\nu)$ and $f_s(s|\nu)$ when the classifier s is a boosted-decision tree (BDT). Right: the corresponding ROC curve (right) for this and other classifiers. Figures taken from TMVA manual.

These steps lead to a subsequent statistical analysis where one observes in data $D = (x_1, \ldots, x_n)$. For each event, the classifier is evaluated and one performs inference on a parameter μ related to the presence of the signal contribution. In particular, one forms the statistical model

$$f(D\,|\,\mu,\nu) = \prod_{e=1}^{n} \left[\,\mu g_s(s(x_e)\,|\,\nu) + (1-\mu)\, g_b(s(x_e)\,|\,\nu)\,\right] \,, \tag{20}$$

where $\mu = 0$ is the null (background-only) hypothesis and $\mu > 0$ is the alternate (signal-plus-background) hypothesis.[2] Typically, we are interested in inference on μ and ν are nuisance parameters.

[2]Sometimes there is an additional Poisson term when expected number of signal and background events is known, which is referred to as an extended likelihood.

5.1 Comments on typical usage of machine learning in HEP

Nuisance parameters are an after thought in the typical usage of machine learning in HEP. In fact, most machine learning discussions would only consider $f_b(x)$ and $f_s(x)$. However, as experimentalists we know that we must account for various forms of systematic uncertainty, parametrized by nuisance parameters ν. In practice, we take the classifier as fixed and then propagate uncertainty through the classifier as in Eq. 20. Building the distribution $g(s|\nu)$ for values of ν other than the nominal ν_0 used to train the classifier can be thought of as a calibration necessary for classical statistical inference; however, this classifier is clearly not optimal for $\nu \neq \nu_0$.

5.2 A more powerful approach

The idea here is to combine the calibration of the distributions of the classifier output and a more optimal family of classifiers $s(x; \nu_0, \nu_1)$. Creating the family of classifiers is straight forward, one simply augments the training data with x examples drawn from several values of ν and then includes the corresponding value of ν as an input to the classifier. Thus $\{x_e, c_e\} \to \{x_e, \nu_{0,e}, \nu_{1,e}, c_e\}$ leading to a parametrized learner $s(x) \to s(x; \nu_0, \nu_1)$. This leads to a complication: one does not know the value of ν to use when evaluating the parametrized classifier $s(x; \nu_0, \nu_1)$, so one cannot simply pre-compute s_e before performing the likelihood fit. However, when performing a likelihood ratio test, one can calculate the maximum likelihood estimate and conditional maximum likelihood estimate according to Eq. 14 and then form the equivalent likelihood ratio test according to Eq. 2.

6 Conclusions

We have shown that a parametrized family of discriminative classifiers $s(x; \theta_0, \theta_1)$ trained and calibrated with a generative model $f(x|\theta)$ can be used to approximate statistical inference likelihoodased on the ratio $f(x|\theta_0)/f(x|\theta_1)$ when it is not possible to evaluate the densities $f(x|\theta)$ for an arbitrary x. This approach leverages the power of machine learning in a classical statistical setting.

Acknowldgements

KC would like to thank in particular Daniel Whiteson for discussions and encouragement throughout the project, Alex Ihler for challenging discussions that led to a reformulation of the initial idea, and Shimon Whiteson for patient feedback in that process. KC would also like to thank Yann LeCun, Philip Stark, and Pierre Baldi for their feedback on the project early in its conception, Balázs Kégl for discussions about the Kaggle challenge and non-standard loss functions, and Yuri Shirman for reassuring discussions on elementary vector calculus. KC is supported by the US National Science Foundation grants PHY-0854724

and PHY-0955626. KC is grateful to UC-Irvine for their hospitality while this research was carried out and the Moore and Sloan foundations for their generous support of the data science environment at NYU.