DRAFT 0.2

Glen Cowan
17-18 December 2013

More notes on the ATLAS MVA Challenge

This note summarizes discussions on the criteria used to define the multivariate classifier for the Higgs Machine Learning Challenge. It is written with the idea that parts of the text can be incorporated into the documentation available to the participants.

1 Introduction

The goal of the Challenge is to design a statistical test to discover the decay of a Higgs boson into two tau leptons: $H \to \tau^+\tau^-$ (the "signal process"). Although the Higgs boson itself is well established, the existence of this particular decay mode is not. To construct an optimal analysis one should maximize the expected statistical significance of the discovery under the assumption that the signal indeed exists. Details of the statistical procedure are discussed in Sec. 3 and the ingredients required for the Challenge are summarized in Sec. 4.

It is not possible to reproduce all of the complexities of the search for the signal process in the Challenge. In particular some simplifications are made concerning the accuracy of estimated rates for background processes. Further, the search is carried out by counting the number of events found satisfying certain criteria; one may extend this to multiple counts with different criteria or to use of a more sophisticated likelihood-ratio test. The simplifying assumptions used here preserve the main features of a more complicated analysis and are expected to have only a minor influence on its sensitivity and on the broader question of how machine learning can be used to improve the search.

2 Physics background

Brief description of Higgs physics, the LHC and the ATLAS detector (from existing note).

3 Statistical treatment of the measurement

Simulated data samples are provided for the signal and background processes. Each instance ("event") is characterized by a set of measured quantities (the input variables). A simple but realistic type of analysis is where one counts the number of events found in a given region in the space of input variables (the "search region", denoted below as Ω), which is defined by the multivariate analysis. The number of events n found in this region is assumed to follow a Poisson distribution with mean $s + b$,

$$P(n|s,b) = \frac{(s+b)^n}{n!} e^{-(s+b)} \ , \tag{1}$$

where s and b are the expected numbers of events from the signal and background, respectively. To establish the existence of the signal process, we test the hypothesis of $s = 0$ (the

background-only hypothesis) against the alternative where the signal exists and predicts an expected number of events of s. That is, the statistical test in question does not refer to individual events but rather to the entire sample of events collected. We test the hypothesis

$$H_0 \ : \text{ all events are of the background type}$$

versus the alternative

$$H_1 \ : \text{ the events are a mixture of signal and background.}$$

To construct such a test one needs to know the values of s and b corresponding to given search region Ω. These can be estimated using samples of Monte Carlo data generated for the signal and background processes and processed through the simulation of the ATLAS detector. In practice, each simulated event comes with an associated weight w. The values of s and b can be estimated as the sum of the weights for events found in Ω,

$$\tilde{s} = \sum_{\text{sig.},i\in\Omega} w_i \ , \tag{2}$$

$$\tilde{b} = \sum_{\text{bkg.},i\in\Omega} w_i \ . \tag{3}$$

Here the tildes are used to distinguish between values estimated from the simulated data and the true parameter values that would be obtained using an infinite sample. (In Sec. 4 to simplify the notation the tildes will be dropped.)

The statistical accuracy (or variance) of the estimate of b has a direct influence on the statistical significance in a test of the background-only hypothesis. This variance stems from the limited amount of simulated data available and is related both to the binomial fluctuations in the number of background events found in the search region as well as to the distribution of weights, which is not known a priori. Furthermore the Monte Carlo has imperfections due to various approximations used. The estimate of b can be improved by using control measurements based on real data, the details of which are not relevant for this Challenge. We can approximate the achievable improvement in accuracy by taking the estimate of b to be of the form

$$\sigma_b^2 = m\phi \ . \tag{4}$$

Here m is the number of simulated background events found in the search region and ϕ is a scale factor, the exact value of which will depend on the nature of the control measurements carried out. For purposes of the Challenge we take $\phi = 0.2$, which is a realistic value for this analysis. The general structure of a successful multivariate analysis is expected to be relatively insensitive to the value of ϕ used.

Once the search region Ω is defined, one can determine from the real data the number n of events in Ω. Furthermore one will have an estimate $\tilde{b}$ for the expected number of background events and its variance σ_b^2. For purposes of constructing a statistical test we can treat σ_b^2 as fixed to the value estimated from Eq. (4) and model n and $\tilde{b}$ as following the probability distribution

$$P(n, \tilde{b}|s, b, \sigma_b^2) = \frac{(s+b)^n}{n!} e^{-(s+b)} \frac{1}{\sqrt{2\pi}\sigma_b} e^{-(\tilde{b}-b)^2/2\sigma_b^2} \; . \tag{5}$$

Regarding the probability (5) as a function of the parameters s and b thus gives the likelihood function $L(s, b)$. This in turn can be used to construct a statistical test of the background-only hypothesis using the profile likelihood ratio

$$\lambda(0) = \frac{L(0, b_0)}{L(\hat{s}, \hat{b})} \; . \tag{6}$$

Here $\hat{s}$ and $\hat{b}$ are the values of s and b that maximize $L(s, b)$ (the ML estimators) and b_0, called the profiled value of b, is the value which maximizes $L(0, b)$, i.e., with s fixed to zero. From this one defines the test statistic

$$q_0 = -2 \ln \lambda(0) \tag{7}$$

for $\hat{s} > 0$ and $q_0 = 0$ otherwise. Higher values of q_0 correspond to increasing incompatibility between the data and the background-only hypothesis and to greater evidence in favour of the existence of the signal.

According to Wilks' theorem [1], providing certain regularity conditions are satisfied, the distribution of q_0 approaches a simple asymptotic form related to the chi-squared distribution in the large-sample limit. In practice the asymptotic formulae are found to provide a useful approximation even for moderate data samples (see, e.g., [2]). Assuming that these hold, one finds that the p-value of the background-only hypothesis from an observed value of q_0 is

$$p = 1 - \Phi\left(\sqrt{q_0}\right) \; , \tag{8}$$

where Φ is the standard Gaussian cumulative distribution.

In particle physics it is customary to convert the p-value into the equivalent *significance* Z, defined as

$$Z = \Phi^{-1}(1 - p) \; , \tag{9}$$

where Φ^{-1} is the standard normal quantile. Equations (10) and (9) lead therefore to the simple result

$$Z = \sqrt{q_0} \; . \tag{10}$$

The quantity Z measures the statistical significance in units of standard deviations or "sigmas". Often in particle physics a significance of at least $Z = 5$ (a five-sigma effect) is regarded as sufficient to claim a discovery. This corresponds to finding the p-value less than 2.9×10^{-7}.[1]

[1] This extremely high threshold for statistical significance is a subject of some debate in the particle physics community. It is motivated by a number of factors related to multiple testing, accounting for mismodeling and the high standard one would like to require for an important discovery.

4 Summary of ingredients

To determine the optimal search region, one must maximize the significance expected under assumption that the signal process is present. This can be approximated by simply evaluating the significance Z from Eq. (10) with n and $\tilde{b}$ replaced with the estimates of their expectation values, $s + b$ and b, respectively.

We will use $\mathcal{M}$ to denote the expected value of Z under assumption that signal exists. Furthermore in this section to simplify the notation we will use s, b, b_0 and σ_b^2 to refer to the values estimated from the simulated data. The quantity to optimize is therefore

$$\mathcal{M} = \sqrt{2\left((s+b)\ln\frac{s+b}{b_0} - s - b + b_0\right) + \frac{(b-b_0)^2}{\sigma_b^2}} \tag{11}$$

where

$$s = \sum_{\text{sig},i\in\Omega} w_i \,, \tag{12}$$

$$b = \sum_{\text{bkg},i\in\Omega} w_i \,, \tag{13}$$

$$b_0 = \frac{b-\sigma_b^2}{2} + \frac{1}{2}\sqrt{(b-\sigma_b^2)^2 + 4\sigma_b^2(s+b)} \tag{14}$$

$$\sigma_b^2 = m\phi \,, \tag{15}$$

$$m = \text{number of background events in search region } \Omega, \tag{16}$$

$$\phi = 0.2\,. \tag{17}$$

In the limit where the expected number of background events is very well determined (i.e., $\phi \to 0$), the expected significance reduces to

$$\mathcal{M} \approx \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)}\,. \tag{18}$$

For the value of $\phi = 0.2$ this should usually give a reasonable approximation. Equation (18) in turn reduces to $\mathcal{M} \approx s/\sqrt{b}$ for $s \ll b$. These approximations often hold in practice and may, depending on the search region chosen, be valid in the Challenge.

References

[1] S.S. Wilks, *The large-sample distribution of the likelihood ratio for testing composite hypotheses*, Ann. Math. Statist. **9** (1938) 60-2.

[2] Glen Cowan, Kyle Cranmer, Eilam Gross and Ofer Vitells, Eur. Phys. J. C 71 (2011) 1554; arXiv:1007.1727.