DRAFT 0.1

Glen Cowan
RHUL Physics
29 March, 2013

Thoughts on ATLAS MVA Challenge

This note records and extends some thoughts from the discussion at LAL on 29/3/13 (GDC, David Rousseau, Balazs Kegl, Cecile Germain) concerning a Data Challenge, in which participants would be given n-tuples of signal and background processes and try to construct an analysis with a maximum discovery sensitivity for the signal.

A possible set of rules for the analysis would be that the participant should define a single search region of the space of input variables in which one would count n events. These can be modeled as following a Poisson distribution with mean $\mu s + b$, where s and b are the expected numbers of events from the signal and background processes and μ is a strength parameter. In particular one is interested in testing the hypothesis $\mu = 0$.

A complication that one would like to include is that the background MC events will have weights, and the distribution of weights will not be known a priori. (In principle the signal events could also have weights, but in general it is much easier to generate a large sample of signal MC so let us assume for the moment that this can be produced without weights.)

Suppose the Challenge is to design an analysis for a certain integrated luminosity, $L_{\rm data}$, and that the MC samples correspond to effective integrated luminosities $L_{\rm s}$ and $L_{\rm b}$. These are related to the data luminosity by

$$\begin{aligned} L_{\rm s} &= \tau_s L_{\rm data} \ , \\ L_{\rm b} &= \tau_b L_{\rm data} \ , \end{aligned}$$

where $\tau_{\rm s}$ and $\tau_{\rm b}$ are known scale factors.

The analyst produces in the multivariable space a selection region and this results in $n_{\rm s}$ events found in the signal sample and $n_{\rm b}$ events in the background sample, with weights $w_1, w_2, \ldots, w_{n_{\rm b}}$.

From these values one can produce estimates[1] for s,

$$\tilde{s} = n_{\rm s}/\tau_{\rm s} \ , \tag{1}$$

and for b

$$\tilde{b} = \frac{1}{\tau_{\rm b}} \sum_{i=1}^{n_{\rm b}} w_i \ . \tag{2}$$

Given large enough MC samples, s and b could be estimated with arbitrary precision. Here we will assume $\tilde{s}$ is determined with negligible uncertainty but that this is not true for b.

[1] Here we use tildes for estimators rather than hats to distinguish them for other estimators for the same quantities that will enter later in the analysis.

Rather we will take $\tilde{b}$ as Gaussian distributed with a mean equal to the true b and a variance $\sigma^2_{\tilde{b}}$, which can be estimated as

$$\tilde{\sigma}^2_{\tilde{b}} = \frac{1}{\tau^2_{\rm b}} \sum_{i=1}^{n_{\rm b}} w_i^2 \; . \tag{3}$$

In the "real" experiment, one finds therefore n and $\tilde{b}$, which are modeled as

$$n \sim \text{Poisson}(\mu s + b) \; , \tag{4}$$

$$\tilde{b} \sim \text{Gauss}(b, \sigma_{\tilde{b}}) \; , \tag{5}$$

and for $\sigma_{\tilde{b}}$ we use the estimate from Eq. (3). The likelihood function is thus

$$L(\mu, b) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \, \frac{1}{\sqrt{2\pi}\sigma_{\tilde{b}}} e^{-(\tilde{b} - b)/2\sigma^2_{\tilde{b}}} \; . \tag{6}$$

The log-likelihood function is therefore found to be (up to an additive constant)

$$\ln L(\mu, b) = n \ln(\mu s + b) - (\mu s + b) - \frac{1}{2} \frac{(b - \hat{b})^2}{\hat{\sigma}^2_{\hat{b}}} \; . \tag{7}$$

To test a hypothetical value of μ we can use the profile likelihood ratio (see, e.g., [1]),

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{b}}(\mu))}{L(\hat{\mu}, \hat{b})} \; , \tag{8}$$

where the double hat notation indicates the value of b that maximizes L for the given value of μ, and single hats denote the (unconditional) maximum-likelihood estimators. In particular we are interested in testing $\mu = 0$, so we need $\lambda(0)$, and for this we require $\hat{\hat{b}}(0)$. The ingredients are found to be

$$\hat{\mu} = \frac{n - \tilde{b}}{s} \; , \tag{9}$$

$$\hat{b} = \tilde{b} \; , \tag{10}$$

$$\hat{\hat{b}}(0) = \frac{\tilde{b} - \sigma^2_{\tilde{b}}}{2} + \frac{1}{2}\sqrt{(\tilde{b} - \sigma^2_{\tilde{b}})^2 + 4\sigma^2_{\tilde{b}} n} \tag{11}$$

Using these quantities one can then evaluate the statistic

$$q_0 = \begin{cases} -2 \ln \lambda(0) & \hat{\mu} > 0 \; , \\ 0 & \text{otherwise} \; , \end{cases} \tag{12}$$

and from this one can obtain the discovery significance

$$Z = \sqrt{q_0} \, . \tag{13}$$

This is equivalent to the p-value of the $\mu = 0$ hypothesis; the two quantities are related by

$$Z = \Phi^{-1}(1 - p) \, , \tag{14}$$

where Φ^{-1} is the quantile of standard Gaussian.

For purposes of the Challenge one would like to maximize the sensitivity for discovery of the signal process, which can be quantified as the median discovery significance under assumption of $\mu = 1$ (i.e., assuming that the nominal signal model is correct). To obtain an estimate of this one can simply take the significance from Eq. (13) and evaluate it replacing the measured values n and $\tilde{b}$ by their expectation values, $\mu s + b$ and b (the so-called Asimov data set). For s and b one then uses the best available estimates, which here are simply $\tilde{s}$ and $\tilde{b}$, and in addition one can use $\sigma_{\tilde{b}}$ from Eq. (3). The task of computing the sensitivity (median significance) can thus be carried out with a few relatively simple formulas.

The procedure can be modified in a straightforward way to include more than one background component. One would in addition like to allow for a more realistic distribution of weights. That is, instead of assuming that $\tilde{b}$ follows a Gaussian distribution, a more realistic model would take, say, a log-normal distribution for the weights. Unfortunately this does not allow for a simple calculation of the ingredients required to obtain the significance (the profiled parameters must be determined numerically) and so this approach is perhaps less appropriate for the data challenge.

References

[1] Glen Cowan, Kyle Cranmer, Eilam Gross and Ofer Vitells, Eur. Phys. J. C 71 (2011) 1-19; arXiv:1007.1727.