Glen Cowan
RHUL Physics
25 August, 2010

Statistics Problems for TAE 2010

Exercise 1: This is a problem from a recent University of London exam. Part (e) on Monte Carlo will not be in the lectures (but you may know anyway how to do this). The rest of the material is contained in lecture 1.

Consider the following two pdfs for a continuous random variable x that correspond to two types of events, signal (s) and background (b):

$$\begin{aligned} f(x|\mathrm{s}) &= 3(x-1)^2 \;, \\ f(x|\mathrm{b}) &= 3x^2 \;, \end{aligned}$$

where $0 \leq x \leq 1$. We want to select events of type s by requiring $x < x_{\mathrm{cut}}$, with $x_{\mathrm{cut}} = 0.1$.

(a) Find the efficiencies for selecting signal and background, i.e., the probabilities to accept events of types s and b, and evaluate numerically.

(b) Suppose the prior probabilities for events to be of types s and b are $\pi_{\mathrm{s}} = 0.01$ and $\pi_{\mathrm{b}} = 0.99$, respectively. Find the purity of signal events in the selected sample, i.e., the expected fraction of events with $x < x_{\mathrm{cut}}$ that are of type s and evaluate numerically.

(c) Suppose an event is observed with $x = 0.05$. Find the probability that the event is of type b and evaluate numerically.

(d) Again for an event with $x = 0.05$, find the p-value for the hypothesis that the event is of type b and evaluate numerically. Describe briefly how to interpret this number and comment on why it is not equal to the probability found in (c).

(e) Describe with the aid of a sketch how to generate values of x following using the acceptance-rejection method.

Describe how to generate values of x following using the transformation method, and find the required transformation.

In both cases assume one has available a generator of random numbers uniformly distributed in $[0, 1]$.

(f) Suppose in addition to x, for each event we measure a quantity y, and that the joint pdfs for the s and b hypotheses are:

$$\begin{aligned} f(x, y|\mathrm{s}) &= 6(x-1)^2 y \;, \\ f(x, y|\mathrm{b}) &= 6x^2(1-y) \;. \end{aligned}$$

Write down the test statistic $t(x, y)$ which provides the highest signal purity for a given efficiency by selecting events inside a region defined by $t(x, y) = t_{\mathrm{cut}}$, where t_{cut} is a specified constant.

Exercise 2: For this exercise you will do a simple multivariate analysis with the TMVA package together with ROOT routines. The code needed for the exercises can be found here:

`www.pp.rhul.ac.uk/~cowan/stat/root/tmva/`

First download the code in the subdirectories `generate`, `train`, `analyze` and `inc` from the website. Alternatively download the tarball `tmvaExamples.tar` to your work directory and type `tar -xvf tmvaExamples.tar`. To build the programs in the individual subdirectories, type `gmake`. The ROOT libraries need to be installed; if this does not work then please ask for help.

First, use the program `generateData` to generate two n-tuples of data whose values follow a certain three-dimensional distribution for the signal hypothesis and another for the background hypothesis. (The n-tuples are created and stored using the ROOT class `TTree`.) Using the macro `plot.C`, take a look at some of the distributions (run root and type `.X plot.C`).

Then use the program `tmvaTrain` to determine the coefficients of a Fisher discriminant. When you run the program, the coefficients of the discriminating functions are written into a subdirectory `weights` as text files. Take a look at these files and identify the relevant coefficients.

Finally use the program `analyzeData` to analyze the generated data. Suppose you want to select signal events, and that the prior probabilities of signal and background are equal. Suppose you select signal events by requiring $t_{\mathrm{Fisher}} > 0$. What are the signal and background efficiencies? What is the signal purity? (Insert code into `analyzeData.cc` to count the number of signal and background events that are selected.)

Modify the programs `tmvaTrain.cc` and `analyzeData.cc` to include a multilayer perceptron with one hidden layer containing 3 nodes. To book the multilayer perceptron you need a line of the form:

```
factory->BookMethod(TMVA::Types::kMLP, "MLP", "H:!V:HiddenLayers=3");
```

See the TMVA manual for more details. This will store the coefficients of the multilayer perceptron in a file in the `weights` subdirectory.

Next to analyze the data using the multilayer perceptron, you will need to add a call to `reader->BookMVA` using the corresponding names (replace `Fisher` with `MLP`). Then book and fill two more histograms to look at the distribution of the MLP statistic under both the signal and background hyothesis (do this in analogy with the histograms for the Fisher discriminant).

Finally, select signal events by requiring $t_{\mathrm{MLP}} > 0.5$. What are the signal and background efficiencies? What is the signal purity assuming equal prior probabilities for the two event types?

Exercise 3: In the lectures we considered an experiment that measured a number of events n, modeled as following a Poisson distribution with a mean value of $\mu s + b$. Here b is the contribution from background and s represents the mean number of events from the nominal signal model. The goal of the experiment is to determine whether the signal is present, i.e., whether μ is non-zero. The likelihood function is

$$L(\mu) = \frac{(\mu s + b)^n}{n!} e^{-(\mu s + b)} \ . \tag{1}$$

The test statistic for discovery q_0 can be written

$$q_0 = \begin{cases} -2 \ln \frac{L(0)}{L(\hat{\mu})} & \hat{\mu} \geq 0, \\ 0 & \hat{\mu} < 0 \ , \end{cases} \tag{2}$$

where $\hat{\mu} = n - b$.

(a) By using the asymptotic relation $Z = \sqrt{q_0}$ from the lecture, show that for $n > b$ the discovery significance Z can be written

$$Z = \sqrt{2 \left(n \ln \frac{n}{b} + b - n \right)} \ , \tag{3}$$

and that therefore the median significance assuming the nominal signal model can be approximated by

$$\mathrm{med}[Z_0|1] = \sqrt{2 \left((s+b) \ln(1 + s/b) - s \right)} \ . \tag{4}$$

(b) By expanding the logarithm show that this reduces to

$$\mathrm{med}[Z_0|1] = \frac{s}{\sqrt{b}} \left(1 + \mathcal{O}(s/b) \right) \ . \tag{5}$$

Although $Z_0 \approx s/\sqrt{b}$ has been widely used for cases where $s + b$ is large, one sees here that this final approximation is strictly valid only for $s \ll b$.