Statistical Methods for Particle Physics
Graduierten-Kolleg RWTH Aachen, February 2014
Problem sheet 2

Exercise 1: Charged particles traversing a gas volume produce ionization, the mean amount of which depends on the type of particle in question. Suppose a test statistic t based on ionization measurements has been constructed such that it follows a Gaussian distribution centred about 0 for electrons and about 2 for pions, with a standard deviation equal to unity for both hypotheses. A test of the pion hypothesis is constructed using the critical region $t < 1$.

(a) Find the size of the test, α (evaluate numerically).

(b) What is the power M of the test with respect to the alternative hypothesis that the particle is an electron?

(c) Suppose we want to select electrons (signal) and reject pions (background). That is, if t is found in the critical region, the pion hypothesis is rejected and the particle is accepted as an electron. What is the electron selection efficiency ε_{e}, i.e., the probability to accept a particle given that it is an electron?

(d) What is the probability that a pion will be accepted as an electron (the background efficiency ε_π)?

(e) Suppose a sample of particles is known to consist of 99% pions and 1% electrons. What is the purity of the electron sample selected by $t < 1$?

For this exercise you will need the cumulative Gaussian distribution, available e.g. from the ROOT routine `TMath::Freq` or from standard tables. Alternatively google for "Gaussian applet" or similar to find an online calculator for the cumulative normal distribution.

Exercise 2: The number of events observed having particular kinematic properties can be treated as a Poisson variable. Suppose that for a certain integrated luminosity (i.e. time of data taking at a given beam intensity), 3.9 events are expected from known processes and 16 are observed. Compute the p-value for the hypothesis that no new process is contributing to the number of events. To sum Poisson probabilities, you can use the relation

$$\sum_{n=0}^{m} P(n;\nu) \;=\; 1 - F_{\chi^2}(2\nu; n_{\mathrm{dof}})\,, \tag{1}$$

where $P(n:\nu)$ is the Poisson probability for n given a mean value ν, and F_{χ^2} is the cumulative χ^2 distribution for $n_{\mathrm{dof}} = 2(m+1)$ degrees of freedom. This can be computed using the ROOT routine `TMath::Prob` (which gives one minus $F_{\chi 2}$) or looked up in standard tables. If you have difficulty getting a program to return F_{χ^2}, you can simply carry out the sum of Poisson probabilities explicitly.

From the $p0$-value, find the equivalent significance $Z = \Phi^{-1}(1-p)$ (you can evaluate the standard Gaussian quantile Φ^{-1} with the routine `TMath::NormQuantile`).

Exercise 3: For this exercise you will do a simple multivariate analysis with the TMVA package together with ROOT routines. First download the code in the subdirectories `generate`, `train`, `analyze` and `inc` from the course website. Alternatively download the tarball `tmvaExamples.tar` to your work directory and type `tar -xvf tmvaExamples.tar`. As usual, to build the programs, 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.)

Make histograms of t_{Fisher} for both signal and background events. (You can superimpose two histograms on the same plot by using `h1->Draw(); h2->Draw("same");`).

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 for the MLP statistic under both the signal and background hyothesis (do this in analogy with the histograms for the Fisher discriminant). Make plots of the resulting histograms.

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?

G. Cowan
10 February, 2014