TRISEP 2016, Vancouver
G. Cowan / TMVA Exercises

For this exercise you will do a simple multivariate analysis with the TMVA package together with ROOT routines. For instructions go to

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

and read the file `readme.txt` to find how to build the programs in the subdirectories `generate`, `train` and `analyze`.

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 train a Fisher discriminant and a boosted decision tree (BDT). When you run the program, the coefficients of the discriminating functions are written into a subdirectory `weights` as text files. You can take a look at these files and see the relevant coefficients.

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

Modify the program `tmvaTrain.cc` to include a BDT with 200 boosting iterations. To do this you need to add the line:

```
factory->BookMethod(TMVA::Types::kBDT, "BDT", "NTrees=200:BoostType=AdaBoost");
```

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

Next to analyze the data using the program `analyzeData.cc`. This is set up to make histograms of the Fisher discriminant for the signal and background events. For the BDT, you will need to add a call to `reader->BookMVA` using the corresponding names (replace `Fisher` with `BDT`). Then book and fill two more histograms for the BDT statistic under both the signal and background hypothesis (do this in analogy with the histograms for the Fisher discriminant). Make plots of the resulting histograms.

Suppose the cross sections for the signal and background processes are $\sigma_{\rm s} = 3\,{\rm fb}$ and $\sigma_{\rm b} = 15\,{\rm fb}$, respectively, and we have a data sample corresponding to an integrated luminosity of $L = 20\,{\rm fb}^{-1}$. Suppose we select signal events using by requiring that the test statistic t be greater than a given value $t_{\rm cut}$ using, e.g., either the Fisher discriminant or the BDT. Find and plot the expected number of signal and background events, s and b, as a function of $t_{\rm cut}$ using

$$s = \sigma_{\rm s} L \varepsilon_{\rm s} \,, \tag{1}$$

$$b = \sigma_{\rm b} L \varepsilon_{\rm b} \,, \tag{2}$$

where $\varepsilon_{\rm s}$ and $\varepsilon_{\rm b}$ are the signal and background efficiencies, i.e., the probabilities to accept signal and background events with the requirement $t > t_{\rm cut}$:

$$\begin{aligned} \varepsilon_{\rm s} &= P(t > t_{\rm cut}|{\rm s}) \;, && (3) \\ \varepsilon_{\rm b} &= P(t > t_{\rm cut}|{\rm b}) \;. && (4) \end{aligned}$$

To make the plots you can use the root class `TGraph` (see root documentation for details).

Suppose we want to choose the optimal value of $t_{\rm cut}$ to give the best test for the discovery of the signal process. This can be done by maximizing the expected discovery significance

$$Z = \sqrt{2\left((s+b)\ln\left(1+\frac{s}{b}\right) - s\right)} \;. \tag{5}$$

Make a plot of Z as a function of $t_{\rm cut}$. Find from the plot the value of $t_{\rm cut}$ that gives the maximum Z and determine the corresponding expected significance.