Kirchhoff-Insitut für Physik
Physikalisches Insitut

Winter semester 2013-14
KIP CIP Pool (1.401)

Exercises for Statistical Methods in Particle Physics

http://www.physi.uni-heidelberg.de/~nberger/teaching/ws13/statistics/statistics.php

Dr. Niklaus Berger (nberger@physi.uni-heidelberg.de)
Dr. Oleg Brandt (obrandt@kip.uni-heidelberg.de)

Exercise 12: Multivariate analysis techniques

20. January 2014
Hand-in solutions by 14:00, 26 January 2014

Please send your solutions to `obrandt@kip.uni-heidelberg.de` by 26.1.2014 14:00, punctually. Make sure that you use *SMIPP:Exercise12* as subject line. If plots are requested, please include print statements to produce pdf files in your code, and provide the plots separately. Please add comments to your source code explaining the steps. Test macros and programs before sending them off...

1 TMVA – Introduction

ROOT comes with an extensive package for multivariate analysis called Toolkit for Multivariate Analysis (TMVA), for which Ref. [1] is very helpful. More general information on MVAs and other statisics questions can be found in great detail in Ref. [2]...
On the CIP-Pool machines, TMVA is installed (for ROOT version 5.28). To test and explore it, copy the directory `/opt/root-5.28/tmva/test` into a location that is writable to you. Change into the directory and run

```
$ root -l 'TMVAClassification.C("Fisher")'
```

You should get a window with lots and lots of buttons, each of which either produces plots or another set of buttons producing even more plots. In this example, TMVA uses the test sample containing four input variables for 6000 signal and background events that comes with TMVA for demonstration purposes. Have a look at the output of the buttons (you may want to start with "Input variables (training sample)", "Input variable correlations", "Classifier Output Distributions", and "Classifier Background Rejection vs Signal Efficiency (ROC Curve)") and try to find out what they mean (there is extensive documentation on the TMVA website).

2 TMVA – simple application

In your TMVA test directory, run

```
$ root -l TMVAClassification.C
```

which will train and run *all* classifiers available in TMVA and take about 15 minutes or so. At 90% efficiency, which classifier has the highest purity?

3 Separating signal from background using TMVA

On the course website, there are three files:

- `ex_12_sig.root` a pure signal file,
- `ex_12_bkg.root` a pure background file,
- `ex_12_dat.root` a file containing pseudo-data,

for a search for a resonance in the invariant mass spectrum of three body decays. As one can verify from the signal sample, it shows a nice resonance. We want to isolate the signal sample from background using multivariate analysis techniques (i.e. not applying straight forward cuts). For each event, besides the invariant three-body mass, there are three measured variables for each of the three decay particles, namely the energy loss per path length $\frac{\mathrm{d}E}{\mathrm{d}x}$ in a gaseous drift chamber, a time-of-flight (TOF) measurement relative to some reference trigger, and the energy fraction deposited in the electromagnetic calorimeter.
To get a feeling for the behaviour of the input variables for signal and background, one typically starts by comparing the signal and background distributions in interactive ROOT.
As next step, modify `TMVAClassification.C` to use the signal and background samples from the course website (use the mass as a spectator variable, and utilise all others as variables as input to TMVA) and train a few classifiers. How easy do you find it to change little bits and pieces in a large package written by someone else?
In the last step, use the classifiers to classify the data sample from the website (you can do this fairly easily by using the data tree as the test sample - classifier values will be saved to the `TMVA.root` output file. Note that in this case, per constructionem, outputs like ROC curves become meaningless.

As a solution,
please provide the modified `TMVAClassification.C` file and a few representative plots - especially the mass spectrum for signal and background compared to data, before and after cutting on some discriminant value for a couple of classifiers. For each of the classifiers, please indicate which operation point (discriminant value) you picked and motivate briefly why based on the ROC curve for signal and background).

Literatur

[1] `http://tmva.sourceforge.net`.

[2] `http://www.statsoft.com/textbook`.