Exercises 2

1. Download TMVA-v3.5.6.tgz from my general account and install:

   a) tar -zxvf TMVA-v3.5.6.tgz <return>

   b) cd TMVA <return>

   c) source setup.sh (or setup.csh depnding on the shell you use)

   d) download macros_exercise_1.tgz

   e) tar –zxvf macros_exercise_1.tgz

   f) cd src

   g) make

   h) while TMVA compiles you can already start in another window/shell with the exercises

2. cd macros_exercise_1 i

   a) in this directory you find a file data_uncorrelatedGauss.root and a little root macro called FindCuts_uncorrelatedGauss.C. The data file contains signal and background events with four observables (var1..var4), where each variable is distributed as a Gaussian. The macro allows you to:

      - plot the distribution of the input variables for signal and background
      - place cuts on all the different variables (see in the macro which values to change) and get the corresponding efficiency and background rejection (the macro will print it for you) … or should I let them program this themselves ????

   b) now if TMVA is already compiled. Run TMVA's cut analysis by typing: root TMVAnalysis_uncorrelatedGauss.C\(\"CutsGA\"\)

3. CutsGA is an cut optimization that uses a Genetics Algorithm in the maximization process of the background rejection for each given efficieny.

   a) Compare your cut performance (the background rejection which you achieved at your signal efficiency) with the one found by TMVA.

   b) Have a look in the "weight file" (weights/TMVAnalysis_CutsGA.weights.txt) which cuts TMVA chose for the efficiency bin you are comparing with. Who won ? Can you see why one or the other was better ?

4. Now run the the Cut optimization and the 1Dimensional likelihood (naïve Bayesian) classifier both together in one job, by typing

   a) root TMVAnalysis_uncorrelatedGauss.C\(\"CutsGA,Likelihood\"\)

b) Select from the GUI the plot of the ROC-curve. This allows you to compare the performance of Cuts with the Likelihood selection.

c) How should the "ideal" ROC curve look like?

d) Which one performes better, Cuts or the Likelihood ?

e) Try to formulate in words why the Likelihood selection performs better than the Cut selection. (remember yesterdays exercise ?)

f) Look at the Reference distributions of the Likelihood.

g) what "are" these reference distributions. Try to give a verbal explanation of what they represent.

h) what do you notice when looking at these distributions ?

i) suggest a way to improve the performance of the Likelihood classifier ? Try it! (Please ask immediately if you don't know how to actually "implement" this option. I hope your suggested option is available in TMVA already :). Otherwise :(

5. repeat 2.) 3.) and 4) with the data file data_correlatedGauss. (use the data file data_correlatedGauss.root and macro FindCuts_correlatedGauss.C and TMVAnalysis_correlatedGauss.C. !! Before you make plots with the GUI, please copy the old directory "plots" to plots_uncorrelated. Otherwise the plots will be overwritten and you cannot compare between correlated and uncorrelated anymore.

   a) How do your own cuts and those found by TMVA differ? What about their performance?

   b) What about the comparison between the performance of the cuts and the Likelihood ? What has changed? Look at the input variable scatter plots

   c) Compare also the "Classifier output distributions" for the two cases (Likelihood with correlated and uncorrelated gauss distributions). What is the "characteristic difference between these two?"

6. Use the KNN method and compare the performance with the 1D likelihood

   a) modify the "size" parameter of the Kernel (Here default is a guassian Kernel)