Analysis Centre Statistics School 2011 Mainz
Multivariate Analysis Tutorial session

These exercises are optional, since there might not be enough time to do them. Exercise 6 is a regression task, exercise 5 is a classification task. They are to be solved in different groups that compete against each other, trying to achieve the best result

Exercise 5 – optional Regression competition

Regression analysis provides an estimate of one (or several) continuous observables based on input variables. In this exercise the data represent measurements in a **toy calorimeter**.

The observable to be estimated is the energy of the calorimeter cluster. All energies are given in GeV. The calorimeter is segmented into five thin layers followed by eight thicker layers. The calorimeter is imperfect in many ways, making the energy measurement more challenging. There are indications of leakage at the end of the calorimeter, dead regions and non-compensation. The data represents a ensemble of measurements from jets and from single particles. There is always just one cluster present in each event. The energy measurements of each layer are labelled *e0* through *e12*. The sum over all layers is called *esum*. The true energy deposition in the training tree is called *etruth*.
The quantity *etruth* is in principle our target variable. In practice it is better to target the correction factor for *esum*, namely the ratio *etruth/esum*. Also available are the cluster centre-of-gravity in η and ϕ (variables *eta* and *phi*). The data file for this exercise is `regressionTestData.root`, and the example macro is `TMVARegressionExample.C`

Try to find a classifier that provides the smallest standard deviation of target vs estimated value. Hints:

- The macro `TMVARegGui.C` is the collection of macros for regression. Use this macro to display the average standard deviation.
- Regression is not yet available for all methods, consider to use:
  - MLP with BFGS training (option TrainingMethod=BFGS)
  - BDT with BoostMethod=Grad
  - PDEFoam
  - FDA
- The measure of success/performance is the regression estimate, the standard deviation of the regression target w.r.t. to the true value.
- For this example $E[\frac{etruth}{esum}] = 1.06$ and $\sigma[\frac{etruth}{esum}] = 0.175$
- Your regression estimate should be significantly better than 0.175!

Exercise 6 – optional Classification competition

The three dimensional data is a lot more complicated than the examples before. The signal has a complex shape, with non-linear correlations. The background is flat. Your goal is to find the best possible classification, as measured by the ROC curve integral.

The groups should be split to cover (evenly) the methods:

- Likelihood
- Multi Layer Perceptron (MLP)
- Boosted Decision Trees (BDT)

The common data file is `classificationTestData.root`.

Hints:

- Take a close look at the signal and the background in the file
- Use the macro `TMVAClassificationExample.C` and modify it!
- For visual aid, use the macro `TMVAGui.C`