Analysis Centre Statistics School 2014 Hamburg
Multivariate Analysis Tutorial session – Part D

4 TMVA Exercise – Function Estimation or Regression

Note: the example is listed on `http://pi.physik.uni-bonn.de/~evt/tmva/`, please have a look there if you are interested in more complicated examples.

Regression analysis provides an estimate of one (or several in case of multi-objective regression) 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 an ensemble of measurements from jets and single particles. There is always only one cluster present in each event. The energy measurements of each layer are labeled *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 theory our target variable. In practice it is better to target the correction factor for *esum*, the ratio *etruth/esum*. Also available are the cluster centre-of-gravity in η and ϕ (variables *eta* and *phi*), where eta_0 is the reconstructed value. The data file for this exercise is `regressionTestData.root`, and the example macro is `TMVACalorimeterRegression.C`

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

- This is a full example that requires you to go through all the steps.
- Please form groups of at least three and pick one of the following methods:
  - MLP with BFGS training (option TrainingMethod=BFGS)
  - BDT with BoostMethod=Grad
- The macro `TMVARegGui.C` is the collection of macros for regression. Use this macro to display the average standard deviation.
- 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 $\sigma[\frac{etruth}{esum}] = 0.175$ (verify the numbers!)
- Your regression estimate should be significantly better than 0.175!
- Run the initial script `TMVARegressionExample.C`, fix it, so that it works.
- Modify the script to make a regression estimate.
  Then take a look at the current training and test error and the final result.
- If you are using MLP, you can modify the number of inputs, the number of hidden nodes and the training.
- If you are using BDT, you can modify the splitting level, the number of nodes and cuts.

Discuss and find answers to the following questions:

- Open the data file and plot the correlation between `etruth` and `esum`.
- What does this tell you?
- How would a perfect (or perfectly calibrated) calorimeter look like in this plot?
- Physics: why is $E[\frac{etruth}{esum}] > 1$?
- Take a close look at all the different input distributions, do they make sense to you? Can you explain the features, on each single one and also in comparison to the related others?