

## Praktikum zu Moderne Methoden der Datenanalyse

### Exercise 7: Data Mining Cup: Cuts

Each year since 2000 there is organised a competition to extract the relevant information from a large amount of data, the Data Mining Cup: [www.data-mining-cup.de](http://www.data-mining-cup.de). 2005 the challenge was to predict whether a customer of an online shop will pay his order or not. This is also the subject of this and the following exercises.

A text file containing a detailed description of the exercise can be found on the web page of this course. The other files needed for this exercise are provided there as well: a root file containing the training data where it is known whether the customer paid, a root file containing the test data with unknown customer behaviour and a text file containing the description of variables in the datasets. These files are also available at `/home/staff/tkuhr/Praktikum/DMC`.

The simplest and most intuitive way of selection or classification is the application of cuts. For many problems this approach is absolutely sufficient. And even if other methods are superior in case of more complex problems, like the one we are facing here, a cut based study can help to understand the data.

- **Exercise 7.1:**

Explore the (training) data and try to find out which variables can be used to predict whether a customer will pay or not. Plot the distribution of variables for good and bad customers. Make profile plots of the target for all variables.

Study correlations of variables as well.

- **Exercise 7.2:**

Invent some cuts for the classification of orders. Evaluate the quality of these cuts by calculating the score as defined in the detailed description of the exercise.

Use the method `MakeClass` of the `TTree` object `h1` to generate the source file for this task. Add the score calculation code to the `Loop` method of the generated tree analysis class.

- **Exercise 7.3:**

Apply your cut based classification to the test dataset in the root file `class.root`. Create a text file containing per line one order number and the corresponding decision separated by a space. Use 1 for high and 0 for low risk orders.

Probably the simplest way to produce such a text file is to open the `class.root` file in the analysis class created in exercise 7.2 and print the two numbers to standard output in the `Loop` method. Then the output can be redirect to a file with the `>` operator.

- **Exercise 7.4:**

Give the text file produced in exercise 7.3 to a tutor. He will calculate a score for you.