The Format of the Data Sets

The data sets in this repository are provided in two separate files. One file with extension ***.data** contains the cases available for training, while the other file with extension ***.domain** contains information on the variables (attributes) used to describe each case.

The file with the training cases (***.data**), contains one case per line. Each line is a set of attribute values separate either by comma, space or tab characters. The last value in this list is the goal variable value (i.e. the variable to be predicted by the learned model).

The file with the information on the used variables (***.domain**) contains as many lines as there are variables. Each line describes one variable. The order of the lines is exactly the order in which the variable values appear in each case of the data file. This means that the first line describes the variable that appears in the first position in the list of values of each case, and so on. Thus the last line describes the goal variable.

The description of each variable has the following format:
variable_name **:** *variable_type***.**

The types of the variables can be:

- The word **continuous**, meaning that the variable is a number.
- A list of discrete variable values separated by commas, in case the variable is discrete.
- The word **nominal**, if the variable is discrete and you don't want to validate and restrict the values appearing in the data file.

An example:

- The file with information on the variables:
  x1: continuous.
  x2: green,red,blue.
  y: continuous.
- The file with the cases:
  12.4, green, 14
  -4.5, blue, 12
  0, green, 56
  9,blue,78
  etc

smail
LIACC-CIUP
R. Campo Alegre, 823

UP
LIACC
MLgroup

email : ltorgo@liacc.up.pt
WWW : http://www.liacc.up.pt/~ltorgo
phone : (+351) 2 6078830

4150 PORTO
PORTUGAL

Luís Torgo

fax : (+351) 2 6003654