Scalable Decision Tree Learning

Implementation of the SLIQ Algorithm

Implement the SLIQ Algorithm, without pruning, using the heuristic.

Use *binary* splits for all types of attributes.

Dataset:

We use a modified dataset from the UCI machine learning repository*, which has 9 attributes and a **binary** target attribute. You can find and download the data under the name

data_exercise_2.csv

from the course website (see category "others"). We have made a few modifications to get the following format:

- The first line contains a sequence of attribute type declarations of the form:
  $$a_1 : T_1, a_2 : T_2, \ldots, a_k : T_k, a_{k+1} : T_{k+1}$$
  where a_i is the id of attribute i and T_i is its type $(1 \leq i \leq k+1)$. All attributes have one of the following types:
  - "n": numerical
  - "c": categorical with at least three attribute values
  - "b": binary with attribute values "yes" and "no"

  In the uploaded data file "data_exercise_2.csv", the first line looks like this:

  a:n, b:c, c:c, d:n, e:b, f:b, g:c, h:c, i:b, j:t

  The last attribute a_{k+1} (in the example above: "j") is always the target attribute and has attribute type "b".
- Each subsequent line describes an example, e.g.,

  24, bb, cc, 3, yes, no, gb, hc, yes, yes
  33, bd, cc, 5, yes, no, gb, hd, yes, no

- The **output** must be a table in ASCII text file format, where each line describes a node n of the decision tree. In particular, the first entry of the line for n specifies the identifier of n, the second the result of the parent's test for n, the third the *Test* (for internal nodes) or the class label (for leaves) associated with n, and the next entries the identifiers of the children of n. *Test* should be in one of the following forms:
  - "i in $\{x_1, \ldots, x_n\}$" if attribute i is categorical and the test is $i \in \{x_1, \ldots, x_n\}$
  - "$i < v$" if attribute i is numerical
  - "i" if attribute i is binary

  where i is the identifier of the attribute used in *Test*. For example, the line for node 5

  5 yes $a < 30$ 6 7

  means that node 5 is true for the test from its parent. Now itself splits by the test $a < 30$ and has children node 6 and 7 corresponding to the different outcomes of $a < 30$.
- Print out all middle steps including updates of histograms during constructing the decision tree.
- Split your data randomly into two parts: use 2/3 of the data as *training examples* for building the decision tree and 1/3 as *test examples* on which you evaluate the predictive accuracy of your decision tree.

[*] The description of the original dataset can be found in the UCI machine learning repository under `https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice`