Homework #5

TAs' email: dsata AT csie DOT ntu DOT edu DOT tw

RELEASE DATE: 05/03/2013
DUE DATE: 05/16/2013, noon

Specification of 5.2

We provide two data sets, namely `heart` and `wine` for you to test your program. Of course, we may use other data sets to evaluate your performance. If you are interseted in trying more data sets, you can check on UCI Machine Learning Repository.

`http://archive.ics.uci.edu/ml/datasets.html`

Input Format

The first argument of your program should be the training data file which contains the examples. The second argument of your program should be θ, the criterion of whether to stop branching.

```
./tree heart 3
```

Data Format

The first line contains two integers n and m, the former one (n) is the number of examples and the latter one (m) is the number of total factors. Each of the following n lines represents an example in the following format, where each number is separated by a space:

```
label factor[0] factor[1] ... factor[m-1]
```

For instance, for the line

```
1 14.23 1.71 2.43 15.6 127 2.8 3.06 0.28 2.29 5.64 1.04 3.92 1065
```

`1` is the label and the rest are the factors.

Output Format

Decision Tree

Please output your tree as a function in C/C++ language. The function must follow this signature:

int tree_predict(double *attr);

The only argument is a double array which contains the factors of one example in the same format as input. This function should return the label prediction of the example (1 or -1 for `heart`, for instance). Also, please name your output file as "tree_pred.h". Then, you can compile and run the provided "tree_predictor.cpp" to check how good your decision tree is (see README). For example, your "tree_pred.h" should look like:

```
int tree_predict(double *attr){
  if(attr[0] > 5){
    return 1;
  }
  else{
    return -1;
  }
}
```

Data Set Description

In this section we provide the meaning of factors and label in the data sets.

Wine

label means two types of wine from two different cultivars.
factors are:
1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10)Color intensity
11)Hue
12)OD280/OD315 of diluted wines
13)Proline

Heart

The data set describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images.
label means two catogories of patients : normal and adnormal.
factors are:
1. F1R: continuous (count in ROI (region of interest) 1 in rest)
2. F1S: continuous (count in ROI 1 in stress)
3. F2R: continuous (count in ROI 2 in rest)
4. F2S: continuous (count in ROI 2 in stress)
5. F3R: continuous (count in ROI 3 in rest)
6. F3S: continuous (count in ROI 3 in stress)
7. F4R: continuous (count in ROI 4 in rest)
8. F4S: continuous (count in ROI 4 in stress)

⋮

- all continuous attributes have integer values from the 0 to 100