Lab04

PREDICTION WITH DECISION TREE

1. Description

You are given a small dataset, called TRAIN, of M1 examples. There are N nominal attributes and one target in the dataset (also nominal). The target has only two possible values (so that the maximum entropy value is 1). Use the ID3 algorithm to build a decision tree from the given dataset.

Later, you are given another dataset, called TEST, of M2 examples. Its format is the same as TRAIN, yet the target values are missing. Use your decision tree to predict these missing values.

2. Specifications

- **Input1:** The training examples are stored in the **train.txt** file, whose format is described as follows:
  - The first line contains a list of N attribute names separated by white spaces. Attributes names are case-insensitive and contain a-z letters only.
  - Each of M1 next lines represent one example. Each example has N values, corresponding to N attributes. These values are case-insensitive and are separated by white spaces.
- **Input2:** The testing examples are stored in the **test.txt** file, whose format is the same as the train.txt file. However, there are M2 examples instead of M1 examples and the missing target values are represented by the question mark ("?")
- **Output1**: The **tree.txt** file stores the decision tree constructed from the training data in the train.txt file. There is no specific format (i.e. you can decide how to represent the tree). For example,

cuu duong than cong . com

```
outlook = sunny
|  humidity = high: no
|  humidity = normal: yes
outlook  =  overcast:  yes
outlook = rainy
|  windy = true: no
|  windy = false: yes
```

- **Output2**: The **predict.txt** file stores the predicted target values for M2 examples in the test.txt file. Each of M2 lines contains the value for the corresponding example.
- The **main function** must perform the following basic actions:
  - Read the TRAIN and TEST datasets from the input files and store the data in appropriate data structures.
  - Call the function **ID3**, which implements the construction of ID3 decision tree from the given training data.
  - Call the function **Prediction**, which implements the prediction of target values for the testing examples using the constructed decision tree.
  - Show the outputs.

An example of the input and output data

train.txt	Note
outlook,temperature,humidity,windy,play sunny,hot,high,false,no sunny,hot,high,true,no overcast,hot,high,false,yes rainy,mild,high,false,yes rainy,cool,normal,false,yes rainy,cool,normal,true,no overcast,cool,normal,true,yes sunny,mild,high,false,no sunny,cool,normal,false,yes rainy,mild,normal,false,yes sunny,mild,normal,true,yes overcast,mild,high,true,yes overcast,hot,normal,false,yes rainy,mild,high,true,no	4 attribute names and 1 target name 14 examples

cuu duong than cong . com

test.txt	Note
outlook,temperature,humidity,windy,play overcast,hot,high,false,? sunny,mild,normal,true,?	4 attribute names and 1 target name 2 examples

<table>
<tr><th>tree.txt</th><th>Note</th></tr>
<tr><td>outlook = sunny
| humidity = high: no
| humidity = normal: yes
outlook = overcast: yes
outlook = rainy
| windy = true: no
| windy = false: yes</td><td></td></tr>
</table>

prediction.txt	Note
yes yes	for the first example for the second example

3. Grading

No.	Specifications	Scores
1	Successfully read the training and testing examples and store them in some data structures	20%
2	Write a complete ID3 function	10%
3	Write a complete Prediction function	10%
4	The tree.txt contains a readable and correct ID3 decision tree	30%
5	The prediction.txt contains correct predicted target values	20%
6	The input and output files strictly follow the specifications	10%
Total		100%

cuu duong than cong . com

4. Notice

- This is an **INDIVIDUAL** assignment.
- 10% bonus will be given as an award for students who can submit a perfect solution 5 days before the submission deadline.
- You are allowed to use data structure functions/libraries (e.g. queue, stack), yet **you must implement the ID3 decision tree by yourself**.
- Report can be written in English or Vietnamese.