Code details

My code is about decision tree.

My code is 100% made by myself, I didn’t copy any information from the internet.

My code can work correctly.

Limitation: the type column must be unique. Otherwise, it cannot work.

My code has two parts. I use the first to get the winner of first question, and the second part is used to get the winner of the second, third question, etc. It depends on how many attributes the set have.

First part:

Use array ‘v’ to store the value of information content and array ‘E’ to store the entropy of each column.

I use the map to store the key and value. Key is the type in each column, for example, column, 0 and 1, represents to simple and compound. Value is the total size of key in the column, for example, in column1, simple is 206 and compound is 44. Because the key is unique, each time I add the entry to the map, if they are the same, then the value will be increased. Finally, for the first column, I will get 0 – 206, 1 – 44. Later I will use the value to calculate the ‘v’; and ‘E’.

After I get the winner of the first question. I will store the column index of winner into LinkedList ‘finished’, and clean ‘myMap’, ‘v’ and ‘E’. And ‘finished’ will be used to store the index of each round best choice.

Second part:

In my second part, I have a big loop. This loop can be dynamic. If I have four attributes, it will run at most 3 times because the first winner I have already gotten in the first time. If I have five attributes, it can run at most 4 times. I have a set ‘newChoiceIndex’. I use it to store the index of column which I choose in the loop. By using it and ‘finished’, I can know which column I haven’t choose and which column need to choose to test the result of ‘E’.

In the big loop, I have a ‘innerloop’, use it to select every column. In the innerloop, I create a small table. I will use it to do the test instead of the big table which I used in the beginning. This small table made by 2d array. I call it ‘tempSet’. It is dynamic. It has three parts. The rightmost the is count column in the original table. The leftmost column is the choice column in each round. And the middle part is the winner column , it always increases. For example, because the first winner is column 4. Then in my second question round. Column 4 will be the middle column in ‘tempSet’. If I got the winner is column3 in the second question, then, the middle part will be column3 and 4 in the next round. I use this ‘tempSet’ to store the key and value for ‘myMap’ and ‘tempMap’. I just convert the entries from the integer to String and concatenate them. For example, row 1 column1 is 0, and row1 column2 is 0, then , it will be ‘00’ as a key stored in map. This time, I have two maps. They may have the same key, like 00, 01,000, etc, depends on what the ‘tempSet’ has and what turn it is. But they have the different values. ‘myMap’ do the same thing as what I did in code first part. And ‘tempMap’ needs to store the value of information content ‘iv’. I need to use two values in two maps which have the same key to calculate the ‘E’.

After I got the smallest ‘E’, I will add the index to the ‘finished’ Linkedlist. In the next round, all indexes in ‘finished’ will not be consider to choose for column and their columns will be added to ‘tempSet’. Then, clean the ‘tempMap’, ‘myMap’ and continuous to find all best questions.

Output of my code:

Total samples number are: 250

Get the number of unique type each column: 2\3\4\3\

max number of type in each column is: 4

E1: 1.087942952277379

E2: 1.022965596927147

E3: 0.8967499262436704

E4: 0.237262742772967

The column 4 -- Margin Sub-Category -- is the 1 best choice.

E1: 0.13139056971549412

E2: 0.10935146969818636

E3: 0.032

The column 3 -- Margin -- is the 2 best choice.

E1: 0.0

E2: 0.0

The column 1 -- Overall Type -- is the 3 best choice.

Find E equal to 0. Program ends.

About Python sklearn library, I watched some videos and tried to make the code in jupyter notebook, but I didn’t make it work. I can import the library.

A picture containing text

Description automatically generated

For some reasons,

First, my data part may not input correctly. I checked the system set ‘wine’. It has a big dictionary, but my set, I have no idea how I can add 250 rows inside. So, I just use 8 rows to test.

A picture containing chart

Description automatically generated

Next part, I use the code to do train, but I got 0.0. I think I may not understand how the function work.

Graphical user interface, text, application

Description automatically generated This part, I just want to check I have the table.

Graphical user interface, text, application

Description automatically generated

Finally, I want to use graphviz to get the decision tree.

Graphical user interface, text

Description automatically generated

But I got error, as a conclusion, I didn’t check the decision tree by using sklearn in python successfully. I think I need to learn more how to use it in the future.

Graphical user interface, text, application, email

Description automatically generated