Department of Computer Science and Engineering

CSE-454: Data Warehousing and Data Mining Sessional

Assignment-4 (Decision Trees)

In our fourth lab, we have seen basics of classification, theories of decision tree classification and implementation of decision tree on restaurant data using scikit-learn. We have seen pseudo-code for Learning Decision Tree algorithm which is as follows.

```
function DECISION-TREE-LEARNING (examples, attributes, parent_examples)
 returns a tree

   if examples is empty then return PLURALITY-VALUE(parent_examples)
   else if all examples have the same classification then return the classification
   else if attributes is empty then return PLURALITY-VALUE(examples)
   else
        A ← argmax_{a ∈ attributes} IMPORTANCE(a, examples)
        tree ← a new decision tree with root test A
        for each value v_k of A do
             exs ← { e : e ∈ examples and e.A = v_k }
             subtree ← DECISION-TREE-LEARNING(exs, attributes − A, examples)
             add a branch to tree with label (A = v_k) and subtree subtree
        return tree
```

In this assignment:

1. You have to implement this algorithm in python programming language.

2. You have to use restaurant dataset to build a decision tree model.

3. Create as much as necessary functions to build and test the model/algorithm. You might need to create the following functions:

i. DECISION TREE LEARNING
ii. IMMPORTANCE (using Information Gain)
iii. PLURALITY VALUE (Majority Learner)
iv. INFORMATION GAIN (to calculate IG for an attribute)
v. FIT MODEL (to be called with data, usually this function calls the learning function)
vi. PREDICT (to predict an unknown sample)
vii. ACCURACY (to find the accuracy of prediction)

And so on. You may adopt OOP concept to well structure your implementation. You may create one or two classes to encapsulate all data and functions.

4. Put sufficient markdown blocks and comments so that anyone can understand your code.

5. Use Jupyter notebook to create your project. Rename the notebook file with your student ID. File name must be your student ID (<std id>.ipynb). Do not add your name or your section in the file name.

6. Submit only one Jupyter notebook (<std id>.ipynb). file. Do not compress it nor include any other file with your submission. Answer the questions in a Markdown block at the end of the notebook.

7. DO NOT COPY FROM ANYWHERE

Marks Distribution

Ser	Description	Marks
1	Loading dataset and data preprocessing (with necessary data structure)	5
2	Implementing learning algorithm (with all supporting functions)	10
3	Implementing prediction function	5
Total		**20**
*	Successful execution of whole project (bonus)	5

* Deadline for submission is **Tuesday 13 October, 2020 11:55pm**
* This assignment will carry 10% weight in final grading.
* Successful execution of whole project will be rewarded 25% bonus of total marks.
* Don't do copy-and-paste programming. Severe actions will be taken against any sort of plagiarism.
* Please leave a comment if find any difficulties.