Assignment 4: Classification

Due: November 25 2019 11:59PM

Two datasets (Golf, Car) can be found on Piazza. In each dataset, each row corresponds to a record. The last column corresponds to the class label, and the remaining columns are the attributes. README presents the meanings of the attributes in these two datasets.

In this assignment, you are asked to implement Decision Tree algorithm. Template (*decisionTree_template.py*) is for Python 3. In the template, you are asked to fill in two functions: *chooseBestFeature* and *stopCriteria*. In *chooseBestFeature*, you need to use Gini index as the impurity measure to decide the best feature to split. In *stopCriteria*, you need to check whether the stopping criteria are satisfied, and return the class label assigned to the leaf node.

The attributes in the provided datasets are either nominal or ordinal. You need to consider multi-way split strategy to split the attributes in this assignment.

Do not directly call a function or package that implements Decision Tree algorithm. You need to implement the algorithm by yourself. If you are not sure about whether it is OK to use a certain function, please post your question on Piazza.

Please take the following steps:

1. Implement Decision Tree algorithm as follows:

DTree(*records*, *attributes*) returns a tree
      If stopping criterion is met, return a leaf node with the assigned class.
   Else pick an *attribute* F based on Gini Index and create a node *R* for it
            For each possible value *v* of *F*:
                  Let *Sv* be the subset of records that have value *v* for *F*
                  call DTree(*Sv*, *attributes* – {*F*}) and attach the resulting tree as the subtree to the current node.
            Return the subtree.

2. Test your Decision Tree algorithm on Golf dataset. Based on your output, you can use the provided *treeplot.py* file to automatically draw the tree. Using multi-way split, the resulting tree for the Golf dataset should look like:
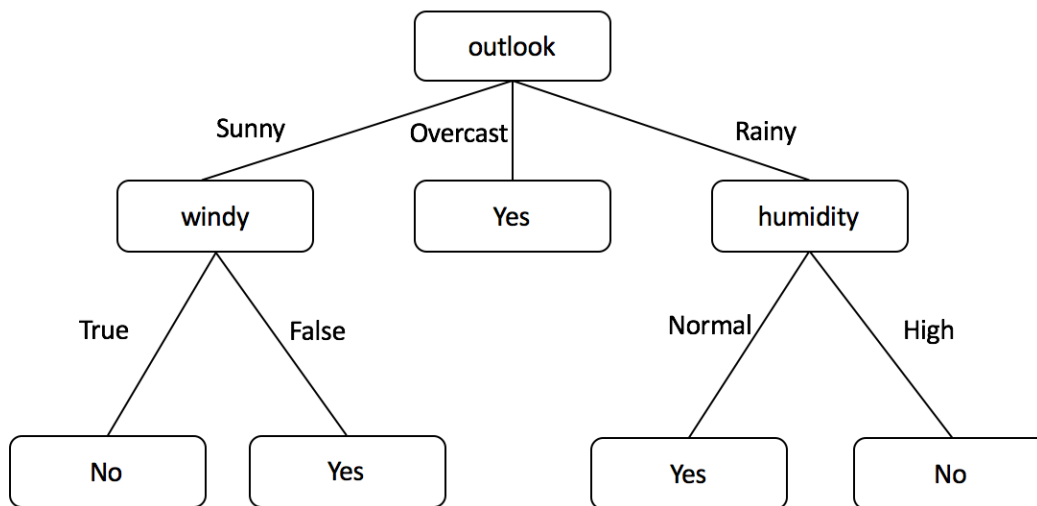
To draw the tree, you can use the provided *treeplot.py* file, or just draw the tree using Excel or PowerPoint, or draw the tree on a piece of paper and include a scanned copy of the tree in the report.

3. If you get the correct tree, then apply your algorithm on the Car dataset and draw the tree.

4. Prepare your submission. Your final submission should be a zip file named as Assignment4.zip. In the zip file, you should include:
   - A folder "Code", which contains all the codes used in this assignment. If you change the other functions that you are not asked to change, please have a file "README" which describes how to run your code.
   - Report: A doc or pdf file named as Assignment4.doc or Assignment4.pdf. The report should consist of the following parts: 1) The tree drawn based on the output obtained from the Car Dataset using your algorithm. 2) The code of the Decision Tree algorithm you implement.

5. Log in any CSE department server and submit your zip file as follows:
submit_cse469 Assignment4.zip

Please refer to Course Syllabus for late submission policy and academic integrity policy. We will take the submission time recorded by the server as the time of your submission. This assignment must be done independently. Running your submitted code should be able to reproduce the results in your report.