CS 165B – Machine Learning, Summer 2020

Homework 1

Due *Thursday, August 20, 2020 11:59pm*

Write a Python3 program called `hw1.py` that creates a decision tree classifier for the *Iris* flower data set. The training and testing data sets are available on Piazza.

1 Details

The program should first read in the contents of the training file, construct a decision tree following the C4.5 algorithm, and then use the decision tree to classify the instances in the testing file. To construct a decision tree, first find the best attribute and threshold to split on. Split the training dataset using the selected attribute and threshold, then recursively construct decision tree from the two subsets of data. Consider using entropy as the impurity function and information gain as the splitting criteria.

You are provided a training data set and a testing data set for this assignment. They are for development only. They contain 5 values on each line: sepal length in cm, sepal width in cm, petal length in cm, petal width in cm, and class, which may be either Iris Setosa, Iris Versicolour, or Iris Virginica.

You are also provided a skeleton code. Do not change the name of the functions. read_data(), c45(), and test() will be called to grade your code. The program must be able to run on CSIL and it must finish running within 1 minute. You are not allowed to use any third-party libraries or frameworks for this homework except those declared in the skeleton code.

2 Grading

The program will be compared to a decision tree classifier that uses entropy and information gain and does not prune. Your classifier needs to perform at least as well as this baseline classifier in order to get full score. A different set of files will be used for grading. The code must work on CSIL or you will receive 0%.

Grade Breakdown:

- 20% for outputting labels
- 30% for implementing the impurity function
- 30% for splitting using information gain
- 20% for performing at least as well as the baseline classifier and finish running in 1 minute.

3 Submission

Submit your solution on CSIL using this command:

```
turnin hw1@changhai_wang hw1.py
```

You may submit multiple times before the deadline. Only the last one will be used.