CIS 6930 - Applied Machine Learning Using Python (Fall 2019)
Project 1 - Due Date: Oct. 18, 2019 Fri 11:59 pm

For this project, you are implementing two decision tree classifiers, namely, Quinlan's ID3 and C4.5 as defined in our lecture. Once implemented, they will be experimented on the Mushroom dataset from the University of California Irvine Machine Learning Repository.

1 Input Dataset

We will use the UCI Mushroom (`https://archive.ics.uci.edu/ml/datasets/mushroom`) datasets for this project. This dataset contains about 8124 mushroom examples over 22 categorical attributes, with each example labeled either edible ("e") or poisonous ("p"). In this dataset, there are 4208 (51.8%) edible examples and 3916 (48.2%)poisonous examples. For your experiments once the decision tree model is implemented, this dataset is randomly split to a training set of 80% and a testing set of 20%. Both have very similar distribution: 51.9% edible and 48.1% poisonous for the training set, and 51.4% edible and 48.6% poisonous for the testing set. Both training.data and testing.data can be downloaded from Canvas and will be used for this project.

2 Requirements

1. The program should implement Quinlan's ID3 and C4.5 decision tree learning algorithms using the template shown in page 5 of slides dt.pdf, where the IMPORTANCE method would be based on information gain and gain ratio, respectively.
2. For both ID3 and C4.5, the program should implement a 10-fold cross validation process over the training set, pick the best model with the highest F_1 score, and run it finally on the testing set and report the F_1 score.
3. Based on the results, compare ID3 and C4.5.
4. Note that, to ensure non-randomness, the 10 partitions used in cross validation are provided as 10 files for you to download, i.e., training_aa.data through training_aj.data.

3 Deliverables

Zip the following to [your-last-name]_Project1.zip and submit to Canvas.

1. A directory that contains all your Python programs that are may be .py or .ipynb.
2. A README file that contains instructions to run your Python programs.
3. A PDF report that describes what your program does and explains the results on the Mushroom dataset for both training (10-fold cross validation) and testing.