Assignment-4: Learning Decision Trees

Problem Statement:

A car manufacturing company X wants to design a system that predicts people's acceptability of a car given its diverse features (number of doors, maximum passengers, boot space, cost of car, cost of maintenance, and safety). The acceptability ratings are expressed using four classes: unacceptable, acceptable, good and very-good. Given a dataset collected by the company X with the aforementioned features, develop a decision tree that predicts the acceptability of a car.

Implementation: [3+3+2=8]

- Implementation of decision tree algorithm [Dec_Tree_Mod].
- Evaluate the role of entropy threshold parameter on accuracy and size of decision tree.
- Devise early stopping method to address over-fitting.

Implement [Dec_Tree_Mod] from scratch. You may make use of the numpy library to perform basic operations. (To create the tree graph structure *only*, you may use the code that we discussed in the tutorial).

**In general, you may use libraries to process and handle data.

Experiments: [3+3+2=8]

The dataset will be split into Train:Validation:Test with 60:20:20 ratio.

1. **Experiment 1:** Report the effect of varying entropy threshold hyperparameter in [Dec_Tree_Mod] for deciding to stop splitting or proceed. Choose threshold values from [0, 0.25, 0.5, 0.75, 1].
   (a). Plot Percentage Accuracy vs threshold hyperparameter on training and validation data.
   (b). Plot size of decision tree vs threshold parameters.

   Using validation data, find the best value of the hyperparameter, entropy threshold, based on percentage accuracy.

2. **Experiment 2:** We now implement an early stopping method to tackle overfitting.
   (a). With the optimal hyparameter found in the earlier experiment, find the overall percentage accuracy on the training and testing data.
   (b). Now repeat the decision tree formation using the optimal parameter step by step, and plot the percentage accuracy on training and validation datasets after every branch formation.
   (c). Repeat the decision tree formation, and stop at the instance where the percentage accuracy on validation dataset starts decreasing. In the new decision tree, analyse whether the overall testing accuracy improved (as compared to that calculated in 2(a))? Similarly, analyze the training accuracy.

3. **Experiment 3:** For the Experiments 1 and 2 traverse the decision trees and print the rules for classification in a readable format ANTECEDENT (IF) => CONSEQUENT (THEN). Use AND and OR operators to print the rules in a compact way.

Datasets:

This dataset comprises four acceptability levels as well as some properties about each car. You can find the dataset here, arranged in the following order.

- Price Buying: buying price (v-high, high, med, low)
- Price Maintenance: price of the maintenance (v-high, high, med, low)
- Doors: Number of doors (2, 3, 4, 5-more)
- Persons: Capacity in terms of persons to carry (2, 4, more)
- Lug_boot: The size of luggage boot (small, med, big)
- Safety: Safety rating of the car (low, med, high)
- Acceptability: People's acceptability (unacceptable, acceptable, good and very-good)

Problem: Predict people's acceptability of a car

Submission:

A .zip file containing the python source code and a PDF report file. The final name should follow the template: <Assign-No>_<Your Roll No>.zip. For example, if your roll no is 15CE30021, the filename for Assignment 4 will be: `Assign-4_15ce30021.zip`

1. A **single python code (.py)** containing the implementations of the models and experiments with comments at function level. The first two lines should contain your name and roll no.

2. A report [PDF] containing **[3 points]**
    a. Experiment 1: Two plots. (a) Bar-chart that shows percentage Accuracy vs threshold parameters (on training and validation data), (b). Size of decision tree vs threshold parameters. (c). Mention the optimal parameter based on the percentage accuracy (on validation data).

    b. Experiment 2: Plot of Percentage accuracy on train and validation data for each branch generation step. Mention the total number of nodes which correspond to the case when validation percentage starts to decrease.

    c. Experiment 3: Print the rules for classification for the decision trees in Experiments 1 and 2.

Responsible TAs:

Please write to the following TAs for any doubt or clarification regarding Assignment 4
Soumyadipto Banerjee: soumyadiptobnrj071@gmail.com
Ankit Katewa: ankitmatrix3@iitkgp.ac.in

Deadline:

The deadline for submission is **10th SEPT, 11:55 PM, IST**. Irrespective of the time in your device, once submission in moodle is closed, no request for submission post-deadline will be entertained. No email submission will be considered. So, it is suggested that you start submitting the solution at least one hour before the deadline.

Plagiarism Policy

1. <= 40: 0
2. In range of 41-50: -3
3. In range of 51-60: -6
4. In range of 61-70: -9
5.In range of 71-80: -12
6. >80 => negative marks without checking ==> -2 (absolute)