CS 165B – Machine Learning, Spring 2021

Machine Problem #2
Due Tuesday, May 18 by 11:59pm PST

Note: This assignment should be completed individually. You are not supposed to copy code directly from anywhere else, and we will review the similarity of your submissions. You should write the code from scratch, and you are not allowed to directly use any third-party tools such as sklearn. But you can use numpy if needed.

Problem description:

Write a **Python3** program called **`mp2.py`** that creates a decision tree classifier for the abalone age classification problem. The training and testing data sets are available on the Assignments page of the course web site. The file **extra_info.txt** has the information about the dataset, label, attribute.

What you need to do:

The syntax of your program should be:

```
% def run_train_test (training_data, training_labels, testing_data)
```

The **`training_data`** and **`testing_data`** are the features for training and testing, and **`training_labels`** are the labels for the training samples. Basically, you need to do the following two things:

- build the decision tree using training_data and training_labels
- predict the labels for the given testing_data

You need to deal with both categorical(integer) and continuous(float) features. You may refer https://www.coursera.org/lecture/ml-classification/threshold-splits-for-continuous-inputs-tn6M9 about how to handle the continuous features.

You are encouraged to explore different strategies to improve the performance of the decision tree such as trying different criteria to select feature to split, do pruning to alleviate over-fitting, and assemble multiple decision trees like random forest.

How to evaluate:

To evaluate your code in local machine, and this will output the accuracy:
python3 evaluate.py

To submit your code: upload **mp2.py** in the Gradescope.

About the grade:

The score on Gradescope are computed based on the accuracy using the following rule:

- If your accuracy is above 50%: Score = min(100, 85 + 100*(Accuracy - 50%))
- If your accuracy is below 50%: Score = max(0, 85 - 0.5 * (100*(Accuracy - 50%))^ 2)

You can also check the leaderboard about the performance of others, and you don't have to use to real name for leaderboard. **Top-3 performers will get extra credit 50%, 25% and 25%.**

Late submission are accepted with 20% reduced credit each day. For example, you will get 80% of full credit if one day late, 60% if two days late and so on so forth.