GEORGIA INSTITUTE OF TECHNOLOGY
SCHOOL of ELECTRICAL and COMPUTER ENGINEERING

ECE 6254 Fall 2022
Project #1

Assigned: 23 Aug
Due Date: 30 Aug

Please contact the TAs for clarification on the instructions in the assignments.

Decision Trees

1. Set up an environment for running Python and Jupyter notebooks. To do this, I have secured access to the COC-ICE PACE cluster for every student in ECE6254. To use the COC-ICE PACE cluster for these homework assignments, please do the following:

   - Become familiar with the PACE Instructional Cluster Environment COC-ICE

     `https://docs.pace.gatech.edu/training/img/ICE_orientation_fall2021.pdf.pdf`

   - VPN into Georgia Tech (see Slide 7 of the ICE Orientation slides above)

     `https://faq.oit.gatech.edu/content/how-do-i-get-started-campus-vpn/`

   - Follow these steps for a general setup for your Project assignments this semester
     (a) ssh <gt-userID>@coc-ice.pace.gatech.edu (see Slide 7 of the ICE Orientation slides)
     (b) module load anaconda3/2020.11
     (c) conda create --name ece6254 python=3.8
     (d) conda activate ece6254
     (e) conda install -c anaconda jupyter
     (f) conda install jupyterlab
     (g) conda install -c anaconda scikit-learn
     (h) pip install turicreate
     (i) conda install -c conda-forge matplotlib
   - For Project01, do the following:
     (a) cp /storage/home/hcocice1/shared-classes/materials/ece6254/Project01.gz .
     (b) tar -xvf Project01.gz
     (c) cd Project01
     (d) jupyter notebook
     (e) Follow link (ctrl+click http://localhost:8888...)
     (f) (from your browser) click on fruit.ipynb

2. Complete the Jupyter notebook for the toy problem (fruit classification). You will need to implement 3 functions: $gini(rows), info\_gain(left, right, current\_uncertainty), build\_tree(rows)$. Look for $<< INSERT\ CODE\ HERE >>$

3. What is the output for the last cell (#43)?

```
for row in testing_data:
        print (”Actual: %s. Predicted: %s” %
                (row[-1], print_leaf(classify(row, my_tree))))
```

4. Now, let's use our decision tree to solve a more complicated problem. Replace the fruit training data with the Titanic dataset (located in 'data/titanic-train.real_valued.csv'). This data set has six binary features:

    - `PassengerId`
    - `Pclass` (which class did the passenger ride)
    - `Sex` (0 = Male, 1 = Female)
    - `Age`
    - `SibSp` (siblings + spouses aboard)
    - `ParCh` (parents + children aboard)
    - `Ticket`
    - `Fare`
    - `Embarked` (Left from Southhampton)
    - `Survived`(1=survived, -1=died)

    Based on these features, the Titanic task is to learn to predict the last column, whether or not the passenger survived (1 = survived).

5. What is the best question to ask first for the titanic dataset?

6. What is the accuracy of your decision tree classifier on the Titanic data set? To calculate this, generate a random 80/20 split, train the model on the 80% fraction and then evaluate the accuracy on the 20% fraction. Repeat this 100 times and average the result.