Homework #3 Due on 12/01/2021

Instructions: While discussion with classmates are allowed and encouraged, please try to work on the homework independently and direct your questions to me.

Part A

1. Sketch the decision tree corresponding to the partition of the predictor space illustrated in Figure 1. The numbers inside the boxes indicate the mean of Y within each region.

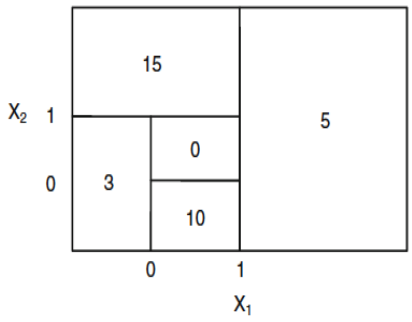

Figure 1

2. Create a diagram similar to Figure 1, using the decision tree illustrated in Figure 2. Please split up the predictor space into the correct regions, and indicate the mean for each region.

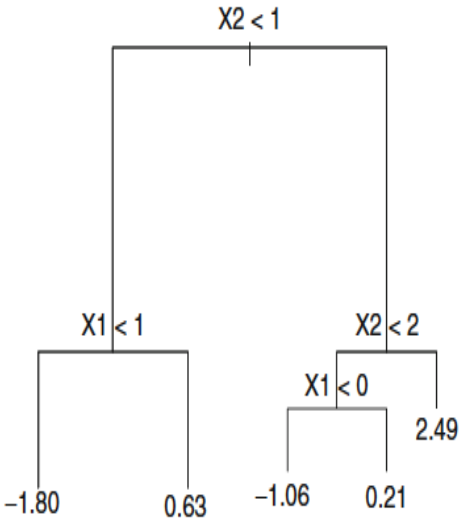

Figure 2

3. If a Decision Tree is overfitting the training set, is it a good idea to try decreasing max_depth?

4. If it takes two hours to train a Decision Tree on a training set containing 2 million instances, approximately how much time will it take to train another Decision Tree on a training set containing 20 million instances?

Part B

In this Lab we will be using the MNIST dataset, which is a set of 70,000 small images of digits handwritten by high school students and employees of the US Census Bureau. Each image is labeled with the digit it represents. Import the MNIST dataset using the following python code:

```
from sklearn.datasets import fetch_openml
mnist = fetch_openml('mnist_784', version=1)
```

1. Split the dataset into a training set, a validation set, and a test set using the ratio $5 : 1 : 1$.

2. Next train the following classifiers:

   a. Random Forest classifier

   b. Bagging classifier

   c. Decision tree classifier

3. Combine the classifiers into an ensemble on the validation set using hard voting.

4. Does the ensemble outperform the individual classifiers?

5. Next remove the individual classifier with the smallest accuracy score.

6. Now combine the classifiers into an ensemble on the test data using hard voting.

7. How much better does it perform compared to the individual classifiers? Comment on your results