Connor Malley
09/15/20
CAP5610
HW2

Task 1

For the Titanic challenge we need to guess whether the individuals from the test dataset had survived or not. Please:

1) **Preprocess your Titanic training data;**

   Done in Python Code from HW1. (Link at end of homework)
   Essentially I just looked at the correlation between each feature and survived, and chose those with a large variance.

2) **Select a set of important features. Please show your selected features and explain how you perform feature selection.**

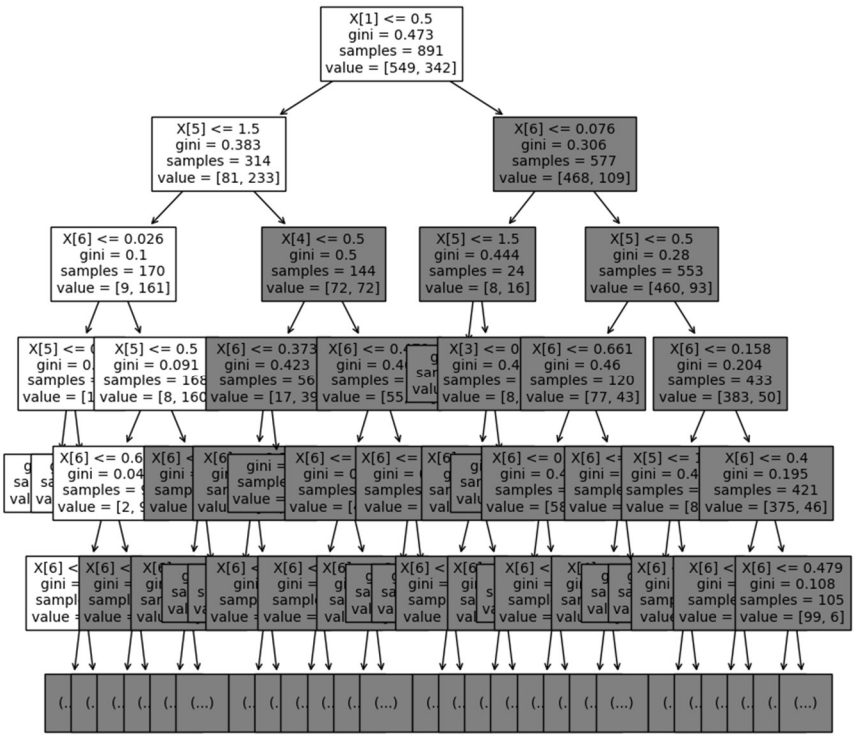
   For the features, I have selected those with high variance, high correlation with survived, and not a lot of missing data. All of the calculations are done in the python code referenced at the end of the homework. The features were selected as follows:

   - Sex
     - 65% male survived, 35% female survived (large difference)
   - Age
     - Not great correlation, but those between 0 to 10 were likely to survive, and those between 15 to 25 were unlikely to survive. So it would be good with banded age groups.
   - Embarked
     - The large majority of those embarking from port P survived. The other ports had relatively equal amounts of survived/not survived
   - Pclass
     - Those in the higher class had a much higher chance of surviving.

   I replaced nans with the mean for the continuous features, and the mode for the categorical features. I also used a one hot encoder for the nominal features.

Connor Malley
09/15/20
CAP5610
HW2

3) **Learn a decision tree model with the Titanic training data using Gini index, plot your decision tree;**



Plotted only a few layers so it's somewhat readable

Decision tree accuracy (no cross validation) : 0.547486

4) **Apply the five-fold cross validation of the decision tree learning algorithm to the Titanic training data to extract average classification accuracy;**

Accuracy of split # 0 == 0.820225
Accuracy of split # 1 == 0.747191
Accuracy of split # 2 == 0.758427
Accuracy of split # 3 == 0.842697
Accuracy of split # 4 == 0.765363

5) **Apply the five-fold cross validation of the random forest learning algorithm to the Titanic training data to extract average classification accuracy;**

Limiting max features to 4
Random forest accuracy : 0.513966

Connor Malley
09/15/20
CAP5610
HW2

6) **Which algorithm is better, Decision Tree or Random Forest?**

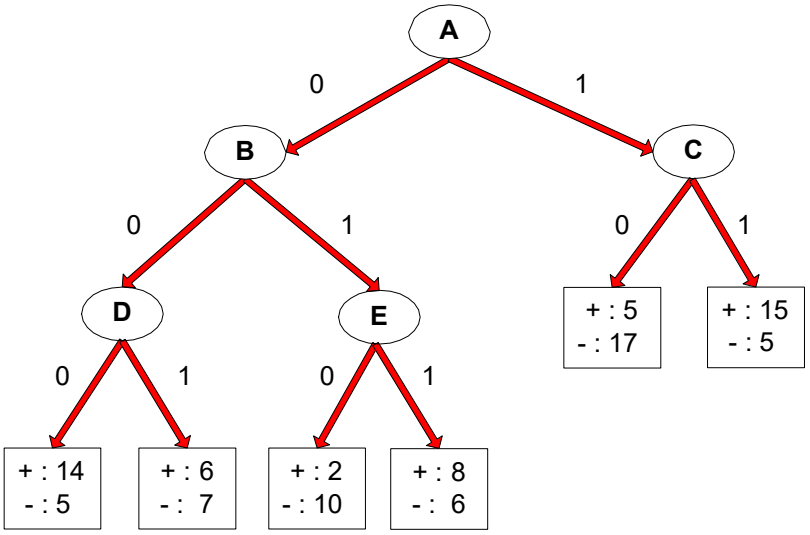
Random forest should in theory be better, so it's likely that I messed something up in the code causing random forest to have a low accuracy (especially for 2 class problem)

7) **What can you learn from the algorithm comparison and analysis?**

I Learned that cross validation is a valuable technique for reducing overfitting of data. I also learned that random forest should in theory help with both underfitting and overfitting, though it didn't help me in this case. Likely a problem with the code.

Task 2

Consider the decision tree shown in the diagram below. The counts shown in the leaf nodes correspond to the number of training records associated with the nodes.



(a) What is the training error rate for the tree (3 points)? Explain how you get the answer?

The training error rate is **29%.** This is calculated by taking the sum of the least probable answers from each leaf node and dividing by the total number of entries classified.

(b) Given a test instance T={A=0, B=1, C=1, D=1, E=0}, what class would the decision tree above assign to T? Explain how you get the answer?

The decision tree would end up with the **"-"** class. I got the answer by going from node A to node B to node E, using the given values to choose my path. Then the class that would be assigned is the most probable at that nod

Connor Malley
09/15/20
CAP5610
HW2

Task 3

Consider the following data set for a binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
T	F	−

Q1: What is the overall entropy before splitting?

0.97

$$-\frac{6}{10}\cdot\log_2\left(\frac{6}{10}\right)-\frac{4}{10}\cdot\log_2\left(\frac{4}{10}\right)$$

$$= 0.970950594455$$

Q2: What is the gain in entropy after splitting on A?

0.28

$$0.971-\left(-\frac{7}{10}\cdot\left(\frac{4}{7}\cdot\log_2\left(\frac{4}{7}\right)+\frac{3}{7}\cdot\log_2\left(\frac{3}{7}\right)\right)-\frac{3}{10}\cdot\left(\frac{3}{3}\cdot\log_2\left(\frac{3}{3}\right)\right)\right)$$

$$= 0.281340304776$$

Q3: What is the gain in entropy after splitting on B:

0.26

$$0.971-\left(-\frac{4}{10}\cdot\left(\frac{3}{4}\log_2\left(\frac{3}{4}\right)+\frac{1}{4}\log_2\left(\frac{1}{4}\right)\right)-\frac{6}{10}\cdot\left(\frac{1}{6}\log_2\left(\frac{1}{6}\right)+\frac{5}{6}\log_2\left(\frac{5}{6}\right)\right)\right)$$

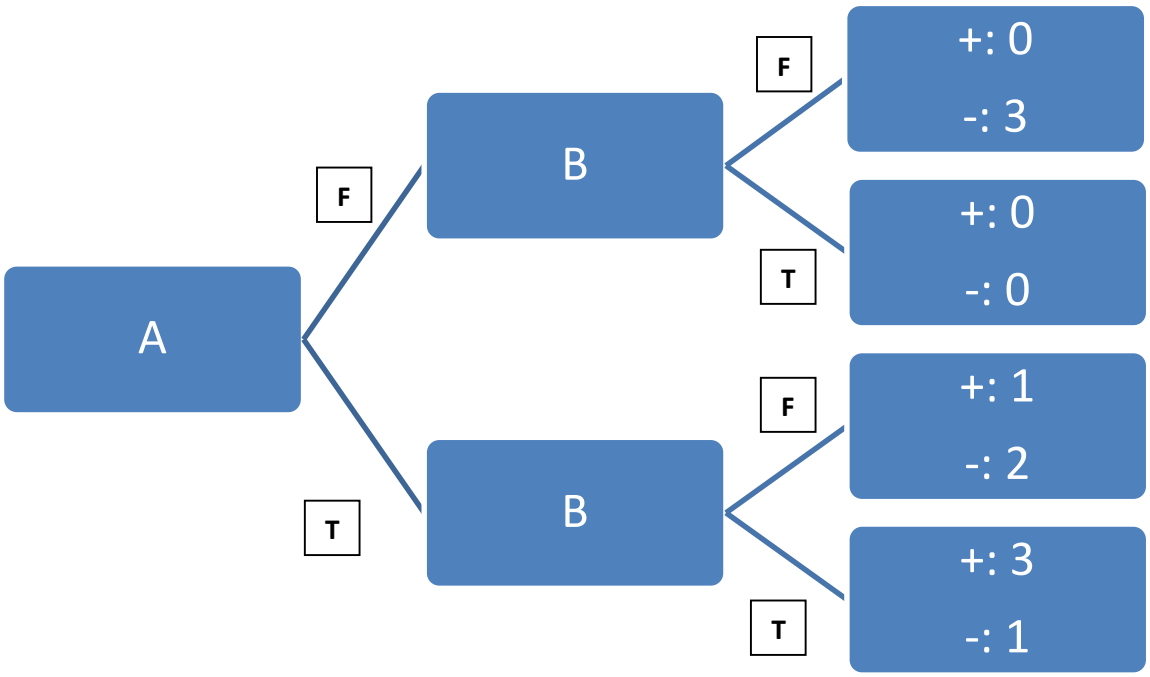
$$= 0.256475297227$$

Connor Malley
09/15/20
CAP5610
HW2

Q4: Which attribute would the decision tree choose?

The decision tree would choose the attribute A, because splitting on A has a higher gain in the entropy than splitting on B

Q5: Draw the full decision tree that would be learned for this dataset. You do not need to show any calculations. (We want to split first on the variable which maximizes the information gain until there are no nodes with two class labels.)



Task 4: Please answer and explain.

Q1: Are decision trees a linear classifier?

Decision trees aren't linear classifiers, as they don't use a linear combination of the input attributes to get a classification (e.g. $ax_1 + bx_2 + cx_3 + \ldots$)

Q2: What are the weaknesses of decision trees?

- Accuracy may not be good for complicated data sets.
- Data with low variance may yield poor results

Q3: Is Misclassification errors better than Gini index as the splitting criteria for decision trees?

Gini is generally better than misclassification error, but this may not always be the case. GINI index is more sensitive to changes in the data, as shown in the classification powerpoint on webcourses.

Connor Malley
09/15/20
CAP5610
HW2

GITHUB LINK:

https://github.com/malleyconnor/cap5610/tree/master/hw2