Devon James November 4th, 2018
CSI 431
Prof. Petkov

Solution Observations

Parts

a) Solution

1) Run through dataset and finds how many classes are of class ‘0’.
2) Find the probability of the “Yes” class by dividing the # of ‘0’ classes by the length of the dataset
3) Find the probability of the “No” class by subtracting the probability of the “Yes” class from 1.
4) Calculate entropy based on those probabilities
5) Make empty arrays to place split and class data (for when the split is actually made)
6) Loop through each row and check to see if the values at the given index are greater or less than the given value and splits them into their respective arrays
7) Find out how many samples are in the split datasets
8) Run through the new data sets and find the number of class ‘0’’s and ‘1’’s.
9) Implement function to avoid “divide by zero” error: If a class has a 0 probability, change the result to 1 (since log(1) = 0)
10) Calculate split entropy using those probabilities
11) Calculate information gain
12) Return value

Observation

I observed that the entropy and split entropy are both needed to calculate information gain. I also observed that many arrays are needed to successfully split the dataset. I also observed that there was a “divide by zero” error, which was fixed by changing the result to 1 for a logically equivalent calculation.

b) Solution

1) Get the length of the dataset
2) Make empty arrays to place split and class data (for when the split is actually made)
3) Loop through each row and check to see if the values at the given index are greater or less than the given value and splits them into their respective arrays
4) Find out how many samples are in the split datasets
5) Run through the new data sets and find the number of class ‘0’’s and ‘1’’s.
6) Calculate Gini Index
7) Return value

Observation

I observed that many arrays are needed to successfully split the dataset. Basically the same logic for the previous class with adjustments for the Gini Index calculation.

c) Solution

1) Get the length of the dataset
2) Make empty arrays to place split and class data (for when the split is actually made)
3) Loop through each row and check to see if the values at the given index are greater or less than the given value and splits them into their respective arrays
4) Find out how many samples are in the split datasets
5) Run through the new data sets and find the number of class ‘0’’s and ‘1’’s.
6) Calculate CART
7) Return value

Observation

I observed that many arrays are needed to successfully split the dataset. Basically the same logic for the previous class with adjustments for the CART calculation.

d) Solution

1) Transpose data
2) Loop through each row in the dataset
3) Loop through each column in the range of the transposed dataset
4) Evaluate criterion – for “G”, minimize by looking for the smallest value starting from b (or 1 in the code); for “IG/CART”, maximize by looking for the largest value starting from a (or 0 in the code).
5) Store the value of the tuple that matches the criterion
6) Return tuple

Observation

I observed that for maximizing and minimizing, you do need a starting point for what the value needs to be greater than or less than.

e) Solution

1) Load the file based on the filename
2) Returns as an array containing the dataset and the class (last col)
3) Ran the best split function for their respective criterions using the train dataset (beginning of classify function)

Observation

I observed that

f) Solution

Observation

I observed that

g) Solution

1) Make an empty array to hold the predicted classes
2) Ran the best split function for their respective criterions using the

train dataset
3) Goes through the dataset and splits the data based on the index found in the best split function
4) Append data to new array
5) Return array

Observation

I observed that