Tyler Paulsen
10/3/2015
CSCI.420.01
HW04: Decision Trees

1) My guess is that the homework will take 4 hours.

3)

a. The weighted gini index and the weighted entropy were both tested on the training set to see what index would work better.
b. During the tests on the training set, the weighted gini index had the better results and is used for the test set.
c. The gini index was used because when run on the training data, the gini index got ~80% correct while the entropy only got ~65%. The results seem like they might be susceptible to overfitting because the entropy was so high for the indexes to split the data into fourths.
d. There were no ties to be broken during the training of the decision tree.
e. The program did not need to break any ties during the training
f. The only stopping criteria used was to split the graph up into 4 sections. The main split down the axis and two other splits perpendicular to the main split. The main split was determined by the attribute with the lowest Weighted Gini index. Once the main split was determined, the data was split into two different sections -- above the index, and below the index. The data was then split up based on the two other attributes that had lower weighted Gini indexes in the two seconds. In each section, the lower of the two weighted gini indexes was taken.
g. There was no noise cleaning on the data. When the second and third attribute were graphed together, there was a skew in the data.
h.

	Attribute Number Used	Threshold Used
a	0	4.2
b	1	3.4
c	1	9.5

The interesting part about the thresholds is that threshold c is extremely high. This would indicate that it will almost alway be classified as a zero if the value is above threshold a.

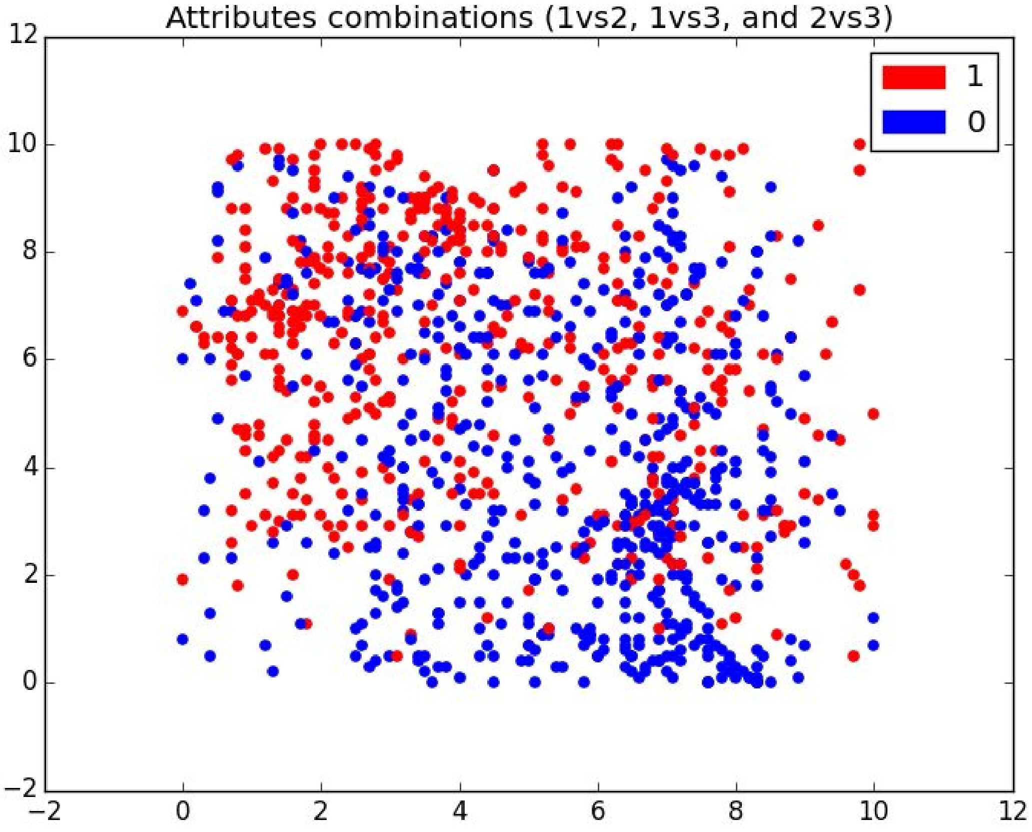
i. The result of my classifier for the training data was 81.71%.
j. The training program only generated the thresholds for the classify to use, it did not actually create the program.
k. The hardest part of getting this to work was ensuring the calculations that the training program created were accurate. The way to check this was to compare it to the scatter

plot of all attributes crossed with each other. This helped to visualize if they were correct or not.

l. The algorithm for the calculations was incorrect 3-4 times. If this homework was not started earlier, it would have been a disaster.

4) The homework took about 8 hours.

5) Bonus


Attributes combinations (1vs2, 1vs3, and 2vs3)
🟥 1
🟦 0