ID3 Machine Learning Write up

Advay koranne

October 2019

0.1 Mushroom Data Set

I will first discuss the mushroom data set which consisted of 8124 elements. I first ran it with 10% of the total set for training purposes which was 812 examples for training. I then ran the test 10 times and calculated the average accuracy which was approximately 99.677%. The average depth of the tree was 5 and the average number of nodes were 26.2 nodes for the all of the trees which were constructed. When I ran the same data set with 50% training the average accuracy went up to 99.985%. This time however, the average depth of the tree was 5 and the average number of nodes was 28.2.

0.2 Titanic Data Set

After running titanic data set which consisted of 1045 elements I ran 10% of the data for training purposes which was 104 elements. The average depth of the tree was 5. There a average of 34.0 number of nodes. Doing a similar process as above I ran the test 10 times and got a average accuracy rate of 72.4017%. When I ran the same data file except with 50% training the average accuracy was 78.356% which was higher than with only 10%. The average depth of the tree was 5 and the average number of nodes was 64.0.

0.3 Primary Tumor

Primary tumor gave much lower results for accuracy than the other files. Using 10% of the data for training which was 33 examples the average accuracy rate was 21.83% which is much lower than the others. However, unlike the other two files above which only had to outcomes, (EDIBLE, POISONOUS or Yes, No) the primary tumor data set had a lot more categories(locations). Just by counting the first couple of categories by hand there were more than 10. This would mean that if I were to guess where a tumor could be located I would probably have less than a 10% chance. However, by only training it on 33 examples, I was able to more than double the accuracy showing that in fact my program is capable of taking examples and “learning.” The average depth of the tree was 11.7 and the average number of nodes was 47.3. After training it using 50% of the data I was able to further improve my accuracy to 35.235%. The average depth of the tree was 18 and the average number of nodes was 253.1.

1 Conclusion

Overall my implementation of the ID3 algorithm shows that it can do a better job than taking a random guess. For example, the probability in the mushroom data set of categorizing one element correctly is 50% however, with mine it was almost 99% showing that the decision tree is serving its purpose.

1.1 Graph

Percentage of Total set used for Training vs. Percentage Accuracy

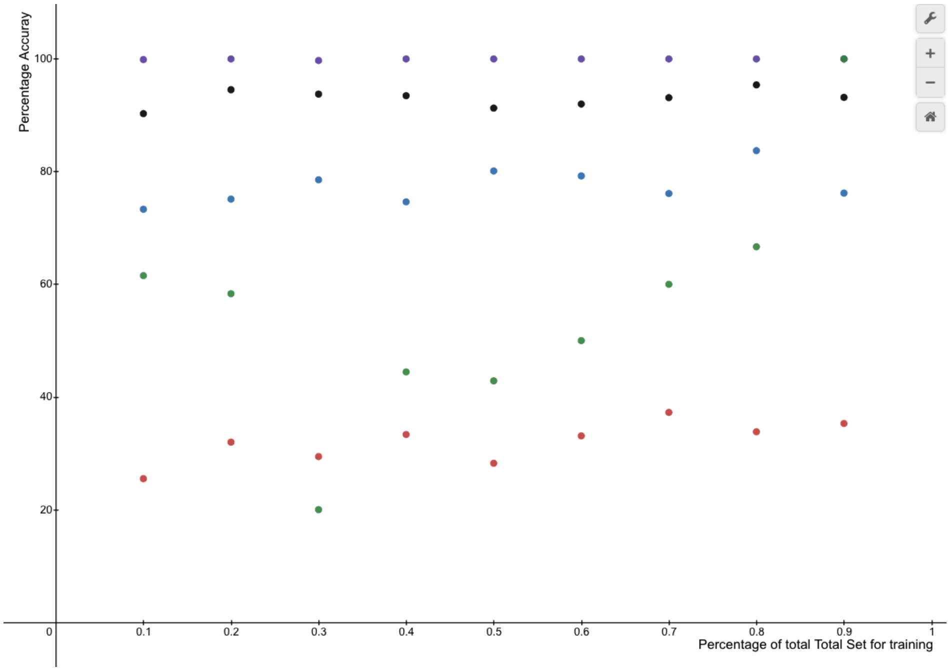

Figure 1: purple = mushrooms, green = tennis , blue = titanic ,red = tumor

2 Tennis Data Output

```
outlook
    sunny
       humidity
         normal
            yes
         high
            no
    overcast
       yes
    rain
       wind
         weak
            yes
         strong
            no
depth:  3  number of nodes:  8
---------Average for 10 trials------------
Average percentage:  100.0
Average depth:  3.0
Average number of nodes:  8.0
```