TDT4171 Assignment 3

erlingrj

February 2018

1 Introduction

I have implemented the decision-tree algorithm in Python. It is based on the algorithm in Fig 18.5 in [1]. To run the code you need to have the *training.txt* and *test.txt* in the same folder as *main.py*.

2 Random Importance

First I implemented the algorithm with random importance appointed to the attributes. This results in quite large decisions tree's and variable performance. See Fig. 1 for a decision tree created with Random Importance. This specific tree classified correctly 18 out of 28 ($64,3\%$) test examples.

3 Entropy Importance

Next I implemented an importance function that calculated the entropy gain for each attribute. It is based on the algorithm described in Chapter 18.5.4 in [1] and calculates at each node the entropy gain for splitting the remaining examples for each remaining attribute. Then we find the attribute which yields the highest entropy gain.

See Fig. 2 for the decision tree with entropy based importance. This tree classified correctly 26 out of 28 ($92,5\%$) test examples.

4 Conclusion

It is clear that splitting the examples by attributes based on their entropy gain yields much better performance. The random importance will sometimes give perfect performance, however, when running the random importance alternative a large number of times, say $10,000$ we get an average of 75.0% correct classification of the test examples.

As observed earlier, the entropy based decision tree will always be the same for the same training data. This is because it is a deterministic algorithm and evaluates an attribute, entropy, which doesn't change from simulation to simulation.

However, when running the random importance version we get a new decision tree each time, this is because we have a (quasi) random algorithm constructing the tree.

5 Appendix

```
A3
A5A5
A2A2A1A4
A1A4A6A1A0A2A2A2
C1A6A1A0A1A4A4C1C1A4A0C1C1A0A0A6
XXA0A4A0C1C1A6C1A4A1A0A0A6XXXXA6A6C1C2XXXXC1A6C1C2A1A1
XXXXC1C2C1A0C1C2XXXXC2A1XXA0C2C1C2C1A1C1C2C2A0XXXXXXXXA2A2A2A2XXXXXXXXXXXXXXXC1A1XXXXA0C1A0C1
XXXXXXXXXXXXXXC1C2XXXXXXXXXXXXXXXC2C1XXXXC1C2XXXXXXXXXC1C2XXXXXXXC1C2XXXXXXXXXXXXXXXXXC1C2C1C2C1C2C1C2XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXC2C1XXXXXXXXXC1C2XXC1C2XX
```

Figure 1: Decision tree with random importance function

```
A0
C1A4
XXA2A5
XXXXA1A1A1A3
XXXXXXXXC1C2C2C1A2A2A1A1
XXXXXXXXXXXXXXXXXXXXXXXXC1C2C2C1A2A2C2A2
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXC1C2C2C1XXC2C1
```

Figure 2: Decision tree with entropy-based importance function

References

[1] Stuart Russell & Peter Norvig, *Artificial Intelligence - A Modern Approach*, Pearson 3rd edition, 2009.