COMPSCI 762 2022 S1 Week 3 Tutorial 2 – Decision Tree

Luke Chang

March 10, 2022

Question 1

Answer the following questions:

1. Explain how entropy is calculated.
2. What do Entropy = 1 and Entropy = 0 mean?
3. Explain what overfit and underfit are, and how they relate to decision tree pruning.
4. What is the effect of different size decision tree? If you train 4 decision trees on one dataset: (a) R0 – the baseline rule, just using Y, (b) R1 – just using 1 variable, (c) a pruned tree, and (d) an unpruned tree. What will be the difference in performance? Which are going to be more likely to underfit? Which are going to be more likely to overfit?

Question 2

2.1 Make a decision tree by hand

Give a decision tree for the following Boolean function using information gain and entropy:

- $A \vee (B \wedge C)$

Answer this question by the following steps

1. Create a table with all combinations;
2. Compute the root entropy without any prior;
3. Find the decision stump with the best score;
4. Split into two subsets based on the stump;
5. Keep finding the next decision stump until you obtain the complete decision tree.

2.2 Coding Practice

Answer the question above using the *DecisionTreeClassifier* method from *sklearn*. Plot the decision tree and compare it with your result.

Question 3

3.1 Make a decision tree by hand

Give a decision tree for Table 1 using information gain and entropy:

Table 1: Verifying gemstones

Colour	Length	Size	Brightness	Shape	Class
red	long	larger	bright	triangle	TRUE
red	long	small	bright	circle	FALSE
red	long	small	bright	triangle	TRUE
red	short	larger	dull	circle	FALSE
red	short	larger	bright	triangle	TRUE
blue	short	larger	bright	triangle	FALSE

3.2 Coding Practice

Answer the question above using the *DecisionTreeClassifier* method from *sklearn*. Plot the decision tree and compare it with your result.

Important Notes

- Check the demo before you start answering questions.
- This is a group activity. You should divide the work into equal shares, and everyone should present their answer during the tutorial.
- The marks are based on the results and your explanation. There is no mark for slides. You may choose whatever you think is suitable for presenting your solution, such as Jupyter Notebook or drawing on a whiteboard.