COMPSCI 762 2022 S1 Week 3 Solution

Luke Chang

April 8, 2022

Question 1

- The maximum entropy is not 1 unless it is binary. "Entropy is 1" does not mean anything on data with multiple classes.
- Decision stump (R1) provides a non-trivial baseline. Later this course, we will learn that using multiple R1 models to build high performance ensemble models.
- Maximum entropy indicates the outcome is not better than random guess.
- 0 is the minimum entropy. It indicates you are 100% certain about the outcome.
- Entropy is not a probability, the range is $[0, \infty)$. It does not bound between $[0, 1]$.

Question 2

- The tree you draw by hand should match the plot from `sklearn`.
- By symmetry, "B" and "C" should have identical entropy. We only need compute one of them in order to save time.

Question 3

- The tree you draw by hand should match the plot from `sklearn`.
- Without prior condition (not looking at any attribute). There are 6 data points. 3 of them are positive, and others are 3 negative. This is a binary case, and the outcome is 50:50. Therefore, we can write $H(Y) = 1$ immediately without any computation (Max entropy).
- Let consider the formula of *Information Gain* (IG), it is a non-negative decreasing.
- The maximum IG cannot go above the parent entropy.
- Recall, 0 is the minimum entropy, and it indicates you are 100% certain about the outcome.
- By combining two bullet points above, we know the maximum IG appears when we choose the attribute with 0 entropy (all cases combine).
- After we found "Shape" maximize IG at the decision stump, we realize $H(Y|\text{Color} = \text{red}, \text{Shape} = \text{triangle})$ and $H(Y|\text{Color} = \text{blue}, \text{Shape} = \text{triangle})$ both have 0 entropy (the output labels are 100% certain). Thus, we can condition on "Color" without computing entropy for other attributes. Other attributes can do no better than "Color". Moreover, because we are 100% the output when we condition on "Color", this is the end of the decision tree. There will be no more leaf node.