CS 4641 Homework 1

Budi Ryan — GTID : 903266310

1

Note: While classifying the attributes, '+' means 'Play' whereas '-' means 'Don't play'

(a) We begin analyzing the gain when we choose the root node using 'Outlook' attribute.
When we choose 'Outlook' to be the root, we will have [2+,3-] for 'Sunny', [4+, 0] for 'Overcast', and [3+,2-] for 'Rain'.

$$Entropy(S) \equiv -p_{\oplus}\log_2 p_{\oplus} - p_{\ominus}\log_2 p_{\ominus} \tag{1}$$

$$Entropy(S) = -\frac{9}{14}(\log_2\frac{9}{14}) - \frac{5}{14}(\log_2\frac{5}{14}) = 0.940 \tag{2}$$

$$Entropy(Sunny) = -\frac{2}{5}(\log_2\frac{2}{5}) - \frac{3}{5}(\log_2\frac{3}{5}) = 0.971 \tag{3}$$

$$Entropy(Overcast) = 0 \tag{4}$$

$$Entropy(Rain) = -\frac{2}{5}(\log_2\frac{2}{5}) - \frac{3}{5}(\log_2\frac{3}{5}) = 0.971 \tag{5}$$

$$Gain(S, Outlook) \equiv Entropy(S) \;\; - \sum_{v \in Values(Outlook)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{6}$$

$$Gain(S, Outlook) = Entropy(S) - \frac{5}{14}Entropy(Sunny) - \frac{5}{14}Entropy(Rain) = 0.246 \tag{7}$$

On the other hand, choosing humidity as our root node will yield us the output [5+,4-] for greater than 75% humidity and [4+,1-] otherwise. Calculating the gain:

$$Entropy(S) \equiv -p_{\oplus}\log_2 p_{\oplus} - p_{\ominus}\log_2 p_{\ominus} \tag{1}$$

$$Entropy(S) = -\frac{9}{14}(\log_2\frac{9}{14}) - \frac{5}{14}(\log_2\frac{5}{14}) = 0.940 \tag{2}$$

$$Entropy(> 75\%) = -\frac{5}{9}(\log_2\frac{5}{9}) - \frac{4}{9}(\log_2\frac{4}{9}) = 0.991 \tag{3}$$

$$Entropy(\leq 75\%) = -\frac{4}{5}(\log_2\frac{4}{5}) - \frac{1}{5}(\log_2\frac{1}{5}) = 0.722 \tag{4}$$

$$Gain(S, Humidity) \equiv Entropy(S) \;\; - \sum_{v \in Values(Humidity)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{5}$$

$$Gain(S, Humidity) = Entropy(S) - \frac{9}{14}Entropy(> 75\%) - \frac{5}{14}Entropy(\leq 75\%) = 0.04505 \tag{6}$$

(b) For the attribute 'Outlook'

$$SplitInformation(S, Outlook) \equiv -\sum_{i=1}^{c} \frac{Si}{S}\log_2\frac{Si}{S} \tag{1}$$

$$SplitInformation(S, Outlook) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{4}{14}\log_2\frac{4}{14} - \frac{5}{14}\log_2\frac{5}{14} = 1.577 \tag{2}$$

$$GainRatio(S, Outlook) = \frac{Gain(S, Outlook)}{SplitInformation(S, Outlook)} = \boxed{0.156} \quad (3)$$

For the attribute 'Humidity'

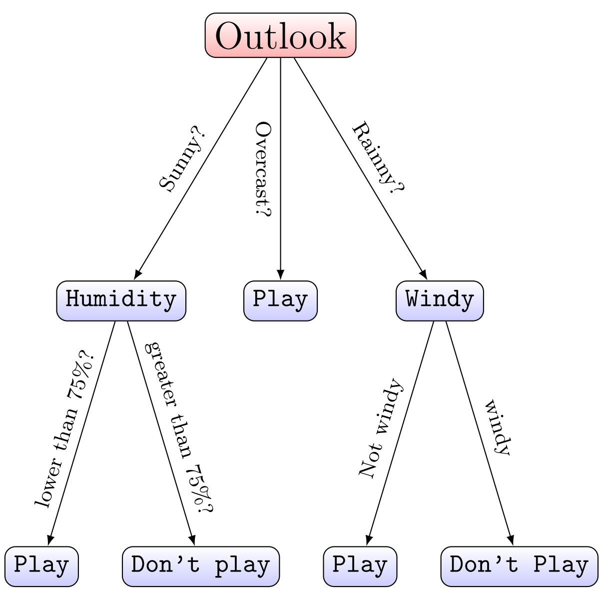
$$SplitInformation(S, Humidity) \equiv -\sum_{i=1}^{c} \frac{Si}{S} \log_2 \frac{Si}{S} \quad (1)$$

$$SplitInformation(S, Humiditym) = -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} = 0.940 \quad (2)$$

$$GainRatio(S, Humidity) = \frac{Gain(S, Humidity)}{SplitInformation(S, Humidity)} = \boxed{0.048} \quad (3)$$

(c) Explanation:
First we have to choose which feature is the best to be the root node.
In this case, we chose *Outlook* because the information gain of *Outlook* is the greatest among all other features such as *Humidity*, *Windy*, and *Temp*: (0.246 vs 0.04505, 0.04784 respectively).
For the next step, we just have to extend the tree for the *Outlook* instance of *Sunny* and *Rainy*.
We do not need to extend *Overcast* because the instance has successfully classified the outcome.
For instance *Sunny*, we then choose the feature *Humidity* because the information gain of it is larger compared to *Windy* (0.971 vs. 0.02 respectively), and it turned out to be a good choice because the branch perfectly classified into 2 discrete classes. Hence, we do not need to branch out anymore.
For the instance *Rainy*, we chose the feature *Windy* and it turned out to be a very good choice, The branches of this feature perfectly split into 2 discrete classes.

The tree:



Note: We do not need 'Temperature' feature in the decision tree because all of the tree's leafs perfectly classify all the examples.