Decision Trees Self-Check

The purpose of this self-check is to make sure you understand key concepts for the algorithms presented during the module and to prepare you for the programming assignment. As you work through problems, you should always be thinking “how would I do this in code? What basic data structures would I need? What operations on those basic data structures?”

ID3 Algorithm

Suppose we have the following training data where Shape, Size and Color are the features (attributes) and Safe? is the class label:

#	Shape	Size	Color	Safe?
1	round	large	blue	no
2	square	large	green	yes
3	square	small	red	no
4	round	large	red	yes
5	square	small	blue	no
6	round	small	blue	no
7	round	small	red	yes
8	square	small	green	no
9	round	large	green	yes
10	square	large	green	yes
11	square	large	red	no
12	square	large	green	yes
13	round	large	red	yes
14	square	small	red	no
15	round	small	green	no

Use the ID3 algorithm to construct the decision tree for it. The formulas are provided below:

Entropy:

$$E(S) = -\sum_i p_i log_2(p_i)$$

Information Gain:

$$G(S,A) = E(S) - \sum_{v \in V_A} \frac{|S_v|}{|S|} E(S_v)$$

And for completeness, the formula for Split is provided below but you don't have to use it. Split is the normalizer for normalized information gain or gain ratio:

$$Split(S,A) = -\sum_{v \in V_A} \frac{|S_v|}{|S|} log_2(\frac{|S_v|}{|S|})$$