Assignment 4.4

Consider the 4 training instances from the EnjoySports example used in assignment 3.3. Build a decision tree using the ID3 algorithm. Write down all Information Gain values. Compare your result tree with the version space.

Sky	Temp	Humid	Wind	Water	Forecast	EnjoySport
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Let YES = + = ‘p’ and NO = - = ‘n’

$$\boxed{\mathbf{Entropy(S)} \equiv -p_{\oplus}\log_2(p_{\oplus}) - p_{\ominus}\log_2(p_{\ominus})}$$

Entropy([3+, 1-]) =

$$-\frac{3}{4}log_2\left(\frac{3}{4}\right) - \frac{1}{4}log_2\left(\frac{1}{4}\right) = 0.811$$

For Gain(S, Attribute=’Sky’)

Sky	p	n	Entropy(Sky)
Sunny	3	0	0
Rainy	0	1	0
∑	3	1	⇒4

$$Entropy(Sky) = \frac{p_{Sunny} + n_{Sunny}}{p+n} \cdot E(Sky=Sunny) + \frac{p_{Rainy} + n_{Rainy}}{p+n} \cdot E(Sky=Rainy)$$

$$Entropy(Sky) = \frac{3}{4}\cdot(0) + \frac{1}{4}\cdot(0) = 0$$

$$Gain(S, Sky) = Entropy(S) - Entropy(Sky) = 0.811 - 0 = 0.811$$

For Gain(S, Attribute='Temp')

Temp	p	n	Entropy(Temp)
Warm	3	0	0
Cold	0	1	0
$\sum$	3	1	$\Rightarrow 4$

$$Entropy(Temp) = \frac{p_{Warm} + n_{Warm}}{p+n} \cdot E(Temp=Warm) + \frac{p_{Cold} + n_{Cold}}{p+n} \cdot E(Temp=Cold)$$

$$Entropy(Temp) = \frac{3}{4} \cdot (0) + \frac{1}{4} \cdot (0) = 0$$

$$Gain(S, Temp) = Entropy(S) - Entropy(Temp) = 0.811 - 0 = 0.811$$

For Gain(S, Attribute='Humid')

Humid	p	n	Entropy(Humid)
Normal	1	0	0
High	2	1	0.918
$\sum$	3	1	$\Rightarrow 4$

$$Entropy(Humid) = \frac{p_{Normal} + n_{Normal}}{p+n} \cdot E(Humid=Normal) + \frac{p_{High} + n_{High}}{p+n} \cdot E(Humid=High)$$

$$Entropy(Humid) = \frac{1}{4} \cdot (0) + \frac{3}{4} \cdot (0.918) = 0.688$$

$$Gain(S, Humid) = Entropy(S) - Entropy(Humid) = 0.811 - 0.688 = 0.122$$

For Gain(S, Attribute='Wind')

Wind	p	n	Entropy(Wind)
Strong	3	1	0.811

$$Entropy(Wind) = \frac{p_{Strong} + n_{Strong}}{p+n} \cdot E(Wind=Strong)$$

$$Entropy(Wind) = \frac{4}{4} \cdot (0.811) = 0.811$$

$$Gain(S, Wind) = Entropy(S) - Entropy(Wind) = 0.811 - 0.811 = 0$$

For Gain(S, Attribute='Water')

Water	p	n	Entropy(Water)
Warm	2	1	0.918
Cool	1	0	0
∑	3	1	⇒4

$$Entropy(Water) = \frac{p_{Warm} + n_{Warm}}{p+n} \cdot E(Water = Warm) + \frac{p_{Cool} + n_{Cool}}{p+n} \cdot E(Water = Cool)$$

$$Entropy(Water) = \frac{3}{4} \cdot (0.918) + \frac{1}{4} \cdot (0) = 0.688$$

$$Gain(S, Water) = Entropy(S) - Entropy(Water) = 0.811 - 0.688 = 0.122$$

For Gain(S, Attribute='Forecast')

Forecast	p	n	Entropy(Water)
Same	2	0	0
Change	1	1	1
∑	3	1	⇒4

$$Entropy(Forecast) = \frac{p_{Same} + n_{Change}}{p+n} \cdot E(Forecast = Same) + \frac{p_{Change} + n_{Change}}{p+n} \cdot E(Forecast = Change)$$

$$Entropy(Forecast) = \frac{2}{4} \cdot (0) + \frac{2}{4} \cdot (1) = 0.5$$

$$Gain(S, Forecast) = Entropy(S) - Entropy(Water) = 0.811 - 0.5 = 0.311$$

$G(S, Sky) = 0.811$

$G(S, Temp) = 0.811$

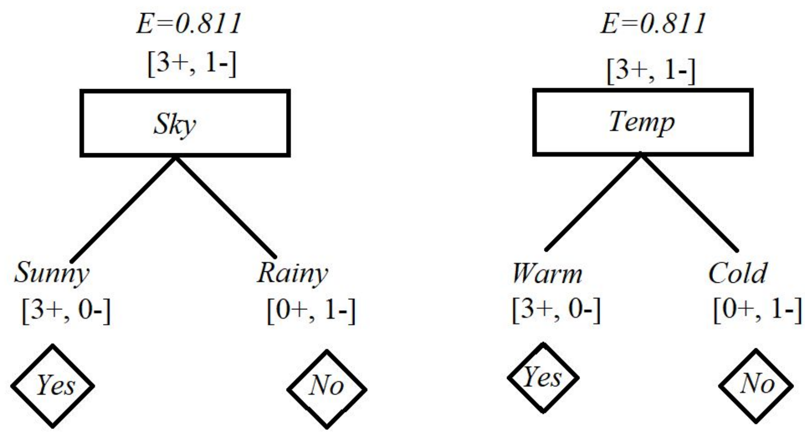
$G(S, Humid) = 0.122$

$G(S, Wind) = 0$

$G(S, Water) = 0.122$

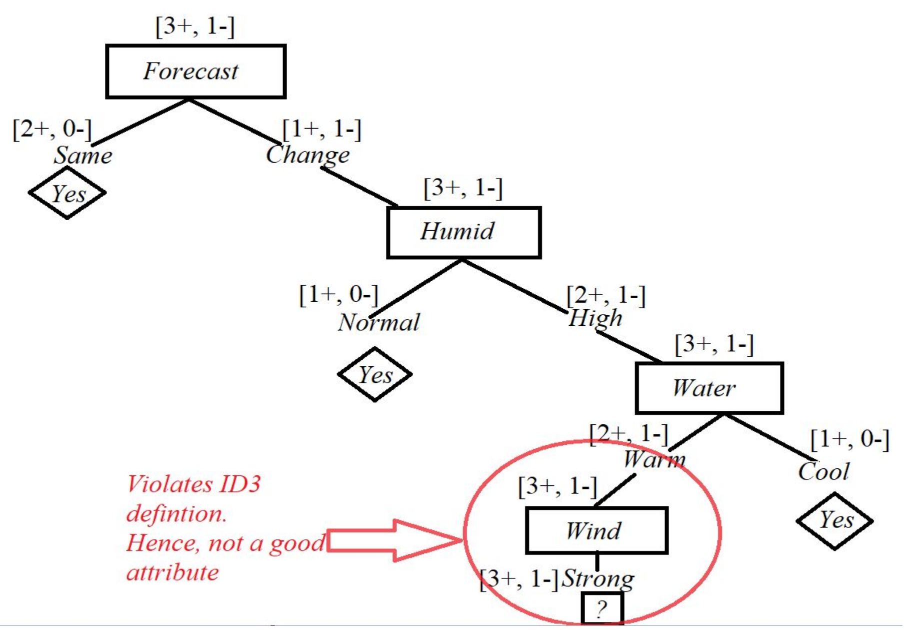
$G(S, Forecast) = 0.311$

From the Gain values we can conclude that attributes *'Sky'* and *'Temp'* are the most useful for classifying examples.

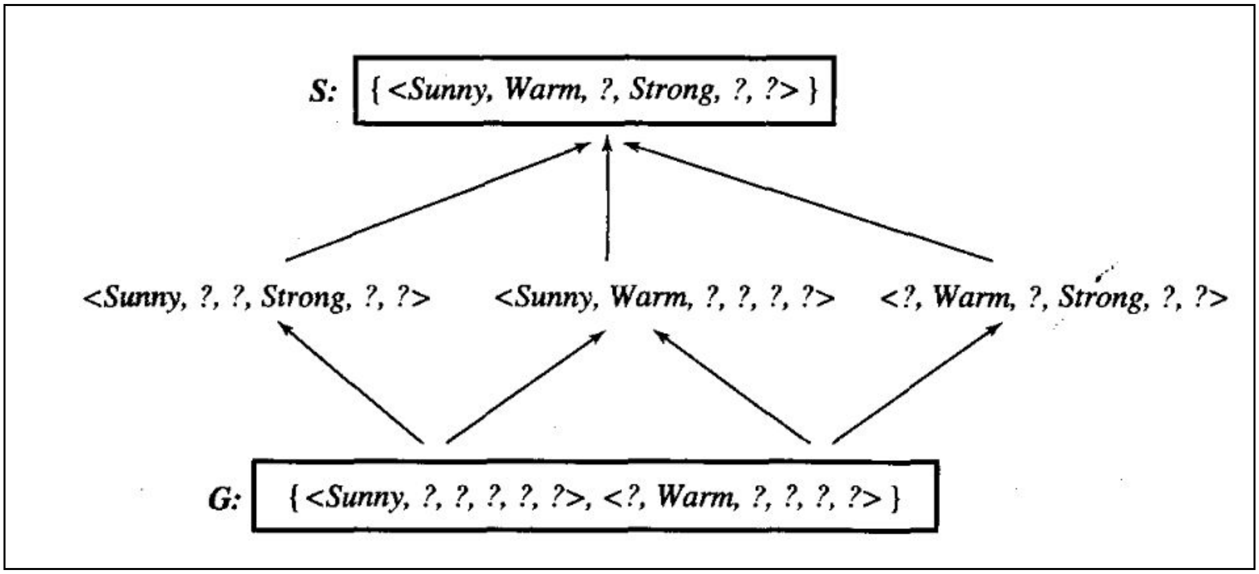


We accept any one of these since ID3 searches for just one consistent hypothesis.

If we build a tree using '*Forecast'* as the root node we obtain a tree like,



In comparison to version space,



Credits: Tom Mitchell

Since, version space includes all the six hypothesis, out learned tree is consistent with the one which is present inside the G boundary.