Introduction to Machine Learning

Sudhamshu B N
1DS18EC091
7th B

Apply the best classification technique to determine the highest information among the given attributes.

Instances	A_1	A_2	Target class
1	T	T	+
2	T	T	+
3	T	F	−
4	F	F	+
5	F	T	−
6	F	T	−
7	F	F	−
8	T	F	+
9	F	T	−

$$Entropy(S) = -P_{\oplus}\log_2 P_{\oplus} - P_{\ominus}\log_2 P_{\ominus}$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Entropy(S) = \begin{array}{l} +ve \rightarrow 4 \\ -ve \rightarrow 5 \end{array}$$

$$Entropy([4+,5-]) = -\left(\frac{4}{9}\right)\log_2\left(\frac{4}{9}\right) - \left(\frac{5}{9}\right)\log_2\left(\frac{5}{9}\right)$$

$$= 0.51996 + 0.4711$$

$$= \underline{\underline{0.99107}}$$

Apply

Entropy[S(A1)]

$$S_T = -\frac{3}{4}\log_2(3/4) - \frac{1}{4}\log_2(1/4)$$

$$= 0.31127 + 0.5$$

$$= \underline{\underline{0.81127}}$$

$$S_F = -\frac{1}{5}\log_2\left[\frac{1}{5}\right] - \frac{4}{5}\log_2\left[\frac{4}{5}\right]$$

$$= 0.4643 + 0.2575$$

$$= \underline{\underline{0.7218}}$$

$$\text{Gain}(S, A_1) = \text{Entropy}(s) - \left\{ \frac{4}{9}\,\text{Entropy}(S_T) + \frac{5}{9}\,\text{Entropy}(S_F) \right\}$$

$$= 0.99107 - \left\{ \frac{4}{9} * 0.81127 + \frac{5}{9} * 0.7218 \right\}$$

$$= \underline{\underline{0.2294}}$$

Entropy[S(A2)]

$$S_T = -\frac{2}{5}\log_2(2/5) - \frac{3}{5}\log_2(3/5)$$

$$= \underline{\underline{0.9709}}$$

$$S_F = -2/4\log_2(2/4) - 2/4\log_2(2/4)$$

$$= 1//$$

$$\text{Gain}(S, A_2) = \text{Entropy}(S) - \left\{ \frac{5}{9}\,\text{Entropy}(S_T) + \frac{4}{9} * \text{Entropy}(S_F) \right\}$$

$$= 0.99107 - \left\{ 5/9 * 0.97095 + \frac{4}{9} \times 1 \right\}$$

$= 0.99107 - 0.98386$

$= \underline{0.00721}$

$Gain(S, A_1) = \underline{0.2294}$

$Gain(S, A_2) = \underline{0.00721}$

A_1 has got more information gain. So that will be the root node.