Page No.
Date: / /

MIES - CT1 18EC10054
SHIKHAR MOHAN

1. a) Total instances possible = 72
Total concepts possible = 2^{72}

b) Adding two other attribute values (?, ϕ) to each space:
Total hypotheses = $4 \times 5 \times 4 \times 4 \times 5 = 1600$

All hypotheses containing even a single ϕ are the same

$\therefore$ statistically different = $1 + (3 \times 4 \times 3 \times 3 \times 4) = 1 + 432$
$= 433$ Ans

c) We apply Find-S on the given data.

$S_0 = \langle \phi, \phi, \phi, \phi, \phi \rangle$ [Most specific hypothesis]

Training on ex-1: Nothing changes (-ve ex), $S_1 = S_0$

Training on ex-2: $S_2 = \langle$ many, big, no, exp, one $\rangle$

Training on ex-3: $S_3 = S_2$ (-ve ex, nothing changes)

Training on ex-4: $S_4 = \langle$ many, ?, no, exp., ? $\rangle$

Training on ex-5: $S_5 = \langle$ many, ?, no, ?, ? $\rangle$

This is the most specific final hypothesis.

2.

S1	A1	A2	A3	A4	T
1	High	Good	High ·	73	Pass
2	High	Fair	Low x	60	Fail
3	Low	Good	High ·	64	Pass
4	Low	Good	low x	57	Fail
5	High	Fair	High ·	66	Pass
6	Low	Fair	High ·	59	Fail

Values (A1) = High, Low

$A1 = [3+, 3-]$

$A1_{High} = [\cancel{3}2+, 1-]$

$A1_{Low} = [1+, 2-]$

$$Gain(S, A1) = Entropy(A1) - \frac{1}{2} Entropy(A1_{High}) - \frac{1}{2} Entropy(A1_{Low})$$

$$= -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\Big[-\frac{2}{3}\log \frac{2}{3} - \frac{1}{3}\log \frac{1}{3} - \frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}$$

$$= 1 + \left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right)$$

$$= 1 - .918$$

$$= 0.082$$

Similarly for A2,

$A2 = [3+, 3-]$

$A2_{good} = [2+, 1-]$

$A2_{fair} = [1+, 2-]$

Being exactly similar to A1, $gain(S, A2) = 0.082$

For A3,

$$A3 = [3+, 3-]$$

$$A3_{High} = [3+, 1-]$$

$$A3_{Low} = [0+, 2-]$$

$$gain(S, A3) = 1 + \left(\frac{4}{6}\cancel{\log}\left(\frac{3}{4}\log\left(\frac{3}{4}\right) + \frac{1}{4}\log\left(\frac{1}{4}\right)\right) + \frac{2}{6}(0)\right)$$

$$= 1 - 0.541$$

$$= \underline{\underline{0.459}}$$

We use the comparator $c = 65$ to convert A4 from a continuous attribute to a discrete attribute.

$$A4 = [3+, 3-]$$
$$A4_{<65} = [1+, 3-]$$
$$A4_{>65} = [2+, 0]$$

Being exactly similar in statistics to A3, we find

$$Gain(S, A4) = 0.459$$

By having the lowest attribute number for the highest entropy gain, we see that A3 is our first node.

$Entropy(A3_{long}) = 0$, hence no further nodes are needed in that branch

$$Entropy(A3_{high}) = 0.811$$

Now we evaluate A1, A2 and A4 under $A3_{high}$

$$Gain(A3_{high}, A1) = 0.811 + \frac{1}{2}\left(\frac{1}{2}\log\frac{1}{2}\right)$$
$$= 0.561$$

$$Gain(A3_{high}, A2) = 0.811 + \frac{1}{2}\left(0 + \frac{1}{2}\log\left(\frac{1}{2}\right)\right)$$
$$= 0.561$$

$$Gain(A3_{high}, A4) = 0.811 + \frac{1}{2}\left(0 + \frac{1}{2}\log\frac{1}{2}\right)$$
$$= 0.561$$

By similar rules as before, we pick A1 as the second node on $A3_{high}$ branch.

$A3_{high} \rightarrow A1_{high}$ has 0 entropy, so no further nodes needed here

We evaluate A2 and A4 on $A3_{high} \rightarrow A1_{low}$

$$Gain(A3_{high} \rightarrow A1_{low}, A2) = 1 - 1$$

$= 0$, which is perfect classification for both $A2_{good}$ & $A2_{poor}$.

Final tree