Module - 5

Decision - tree problems.

1) Classification

Sore throat	Fever	Swollen glands	Congestion	Headache	Diagnosis
Yes	Yes	Yes	Yes	Yes	Strep throat
No	No	No	Yes	Yes	Allergy
Yes	Yes	No	Yes	No	Cold
Yes	No	Yes	No	No	Strep Throat
No	Yes	No	Yes	No	Cold
No	No	No	Yes	No	Allergy
No	No	Yes	No	No	Strep throat
Yes	No	No	Yes	Yes	Allergy
No	Yes	No	Yes	Yes	Cold
Yes	Yes	No	Yes	Yes	Cold

~~Inf Gain,~~

$$I(p(n)) = -\frac{p}{s}\log_2\frac{p}{s} - \frac{n}{s}\log_2\frac{n}{s}$$

$E(S)$

$S = p + n$

Entropy

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} \; I(p_i, n_i)$$

$$\text{Gain}(A) = E(S) - E(A)$$

$$\log_2 x = \frac{\log_{10} x}{\log_{10} 2}$$

$$S = S.T + A + C. = 10.$$

$$\begin{matrix} | & | & | \\ 3 & 3 & 4 \end{matrix}$$

$$E(S) = -\left[\frac{3}{10}\log_2\frac{3}{10} + \frac{3}{10}\log_2\frac{3}{10} + \frac{4}{10}\log_2\frac{4}{10}\right]$$

$$= -\left[0.3\log_2 0.3 + 0.3\log_2 0.3 + 0.4\log_2 0.4\right]$$

$$= -\left[0.6\log_2(0.3) + 0.4\log_2(0.4)\right]$$

$$= -\left[0.6\,\frac{\log_{10} 0.3}{\log_{10} 2} + 0.4\,\frac{\log_{10} 0.4}{\log_{10} 2}\right]$$

$$= -\left[0.6\left(\frac{-0.522}{0.301}\right) + 0.4\left(\frac{-0.397}{0.301}\right)\right]$$

$$= \left[0.6\,(1.73) + 0.4\,(1.318)\right]$$

$$= 1.038 + 0.5272$$

$$= 1.562$$

Finding the splitting attribute [Select attribute with highest gain]

1) Sore Throat

	S T	A	C	
Yes	2	1	2	Inf. Gain × P
No	1	2	2	Inf. Gain × P

} Entropy

E(Sore throat)

$$I(Yes) = -\left[\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]$$

$$I(Yes) = 1.52$$

$$I(No) = -\left[\frac{1}{5}\log\frac{1}{5} + \frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right]$$

$$= 1.52$$

$$E(\text{sore throat}) = P(10 \to 5)\ 0.5 \times 1.52 + 0.5 \times 1.52$$

$$= 1.52$$

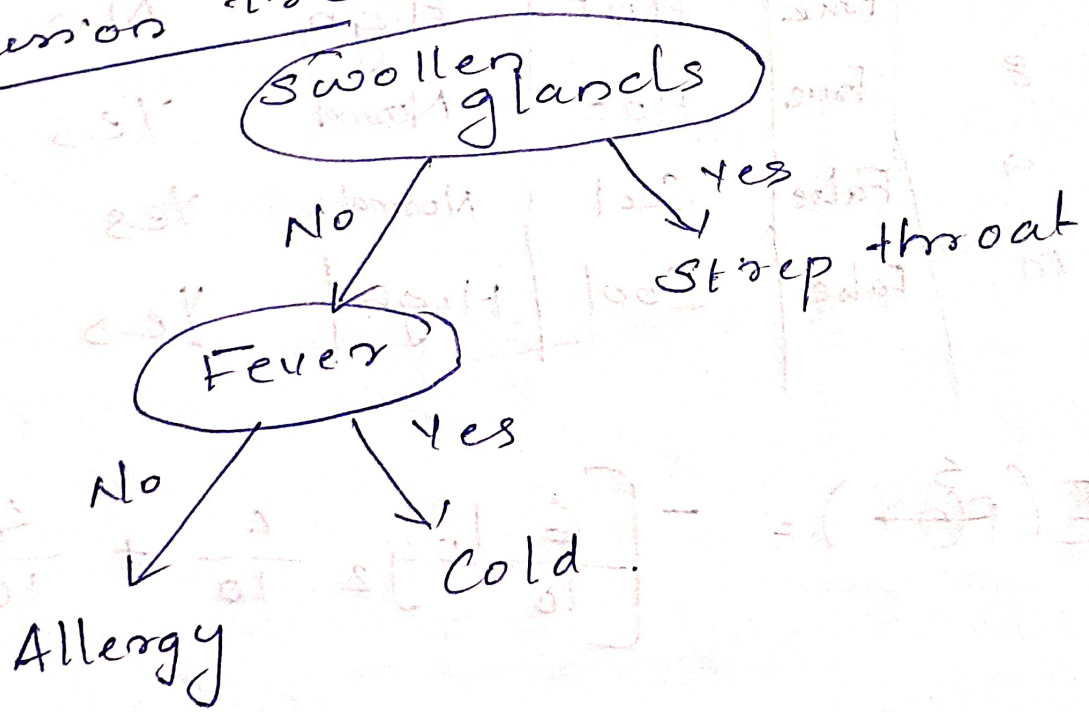
$$\text{Gain} = I(P(n) - E(A)$$

$$= 1.562 - 1.52$$

$$\text{Gain}(S.T) = 0.05$$

Attribute	Gain
S.T	0.05
Fever	0.72
Swollen glands	0.88 ← highest value.
Congestion	0.45
Headache	0.05

Decision tree



Q2

	a1	a2	a3	classification
1	True	Hot	High	No
2	True	Hot	High	No
3	False	Hot	High	yes
4	False	Cool	Normal	Yes
5	False	cool	Normal	Yes
6	True	Cool	High	No
7	True	Hot	High	No
8	True	Hot	Normal	yes
9	False	Cool	Normal	Yes
10	False	Cool	High	Yes

$$E(S) = -\left[\frac{6}{10}\log_2\frac{6}{10} + \frac{4}{10}\log_2\left(\frac{4}{10}\right)\right]$$

$$= 0.9709$$

Entropy : a1

a → True P:4 N:1

a → False P:5.

$$I(true) = -\left[\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right]$$

$$= 0.7219$$

$$I_{neg}(false) = 0$$

$E(a1) = 0.5 \times 0.7219$

$= 0.3609$

$Gain(a1) = 0.9709 - 0.3609$

$= 0.609$

A2

a2 < Hot : P : 3 N : 2.

a2 < cool P : 4. N : 1

$I(Hot) = -\left[\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\frac{2}{5}\right)$

$= 0.9709$

$I(cool) = -\left[\frac{4}{5}\log_2\left(\frac{4}{5}\right) + \frac{1}{5}\log_2\frac{1}{5}\right]$

$= 0.7219.$

$Gain(a2) = 0.1245.$

a3

a3 < High P : 2 N : 4

a3 < normal P : 4

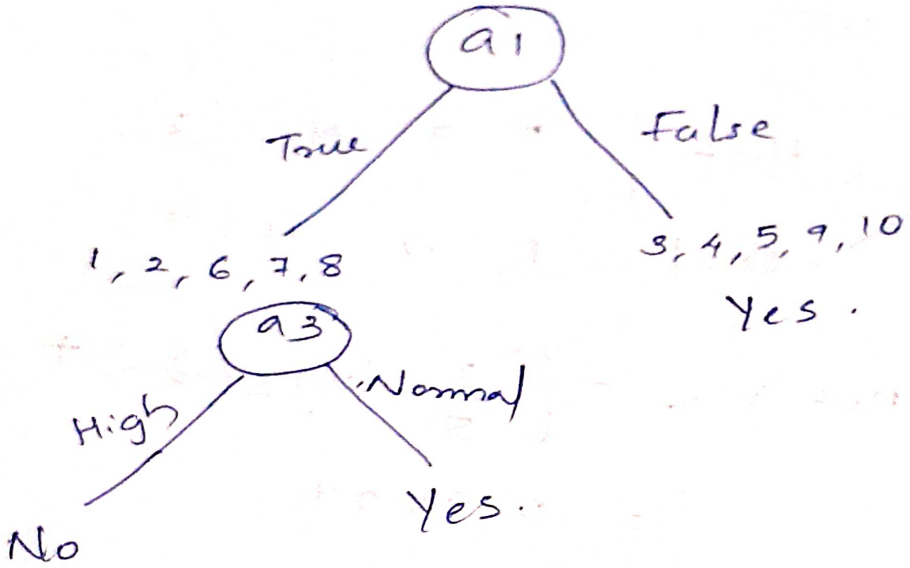
$I(High) = 0.9183$

$I(Normal) = = 0$

$Gain(a3) = 0.4199$

Attr	Gain
a1	0.6099 (Max)
a2	0.1245
a3	0.4199

	Gain
a2	0.3219
a3	0.7219 (Max)



a1	a2	a3	Class.
True	Hot	High	No
True	Hot	High	No
True	Cool	High	No
True	Hot	High	No
True	Hot	Normal	Yes

$$I(D) = -\left[\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right] = 0.7219$$

a2

a2 < Hot (P: 1 N: 3)
     < cool (N: 1)

Entropy (Hot) = 0.8112

Entropy (cool) = 0

Entropy (a2) = $\frac{4}{5}$ x 0.8112

Gain = 0.9709 - 0.8112 = 0.1597

a3

a3 < High
     < Normal

Entropy (High) = 0

Entropy (Normal) = 0

Gain = 0.9709