Final Exam!

→ Decision Tree learning :-

Instance	a_1	a_2	a_3	Target class.
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

How to compute :- Entropy ; Info gain ; Gini Index ; Splitting attribute.

(1) Entropy :- $(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$

where 'pi' is the probability of i^{th} class in this case.

Here '+' : 4/9
'−' : 5/9.

$$(S) = -\frac{4}{9}\log_2(4/9) - \frac{5}{9}\log_2(5/9)$$

$$\boxed{(S) = 0.9911}$$

(2) What is the inf gain of a_1 wrt to Training Ex?

a_1 has 2 possibilies 'T' or 'F'.

first we find out entropy of 'T' examples and entropy of 'F' examples.

$$S = -\sum_{i=1} p_i \log_2(p_i)$$

$$\times\left[\because \quad S(T) = \Sigma\right]\times$$

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

Totally there are 4 true (T) values for a_1.

Now in 'T' there are 3 '+' values and 1 '-' value, hence explains 3/4 and 1/4.

$$\therefore S(T) = 0.811$$

$$S(F) = -\frac{1}{5}\log_2(1/5) - \frac{4}{5}\log_2(4/5)$$

$$S(F) = 0.722$$

$$Gain(a_1) = Entropy(S) - \sum_{v \in \{T,F\}} \frac{|S_v|}{|S|} Entropy(S_v)$$

Entropy of whole dataset.

proportionate of 'T' and 'F' in the whole dataset.

$$= Entropy(S) - \left(\frac{4}{9}\right)(S_T) - \left(\frac{5}{9}\right)(S_F)$$

total 'T' out of 9 in a_1.

total 'F' out of 9 in a_1

$$= 0.9911 - \frac{4}{9}(0.811) - \frac{5}{9}(0.722)$$

$$\boxed{Gain(a_1) = 0.23.}$$

(3) Information gain of a_2 = ?.

$$S(T) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.971$$

$$S(F) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$$

$$= 1.$$

$$\text{gain}(a_2) = 0.9911 - \frac{5}{9}(0.971) - \frac{4}{9}(1)$$

$$= 0.0072.$$

(4) Compute the Gini Index of attribute a_1.

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2.$$

$$\text{Gini}(T) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$$

$$\text{Gini}(F) = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.32$$

$$\text{GiniIndex}(a_1) = \sum_{v \in \{T,F\}} \frac{|S_v|}{|S|} \text{Gini}(S_v)$$

$$= \frac{4}{9}(0.375) + \frac{5}{9}(0.32)$$

$$\text{GiniIndex} = 0.3444$$

Gini Index of a_2 :-

$$\text{Gini}(T) = -\frac{2}{5}\left(\frac{2}{5}\right)^2 - \frac{3}{5}\left(\frac{3}{5}\right)^2$$

$$= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2$$

$$= 0.48.$$

$$\text{Gini}(F) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{3}{4}\right)^2$$

$$= 0.5$$

$$\text{Gini Index} = \frac{5}{9}(0.48) + \frac{4}{9}(0.5)$$

$$\boxed{0.48888}$$

(5) Which is the best splitting attribute between a_1 and a_2?

$\text{gain}(a_1) = 0.229$ (higher info gain produces higher split)

$\text{gain}(a_2) = 0.00072$

$\therefore a_1$ is best split attribute

GiniIndex $(a_1) = 0.3444$

$(a_2) = 0.4889$.

Smaller the GiniIndex, it produces better split.

$\therefore a_1$ has better split.

(b) a_1 / a_2 are discrete valued attributes.

a_3 are continuous attributes.

arrange a_3 in increasing order.

a_3	TC	split point	Entropy	Info gain
1.0	+	$\frac{3.0+1.0}{2} = 2.0$	0.8484	
3.0	−	$\frac{4.0+3.0}{2} = 3.5$	0.988.	
4.0	+	$\frac{5.0+4.0}{2} = 4.5$		
5.0	−	5.5		
5.0	−	$\frac{5.0+6.0}{2} = 5.5$		
6.0	+	$\frac{6.0+7.0}{2} = 6.5$		
7.0	−	7.5		
7.0	+	$\frac{8.0+7.0}{2} = 7.5$		
8.0	−			

S = 0.9911 (as we know).

split = 2.0 (chuk how many Ex 2.0 ↑ how many after)

$E = -\frac{1}{9}\log_2\left(\frac{1}{9}\right) - \frac{8}{9}\log_2(8/9) = 0.8484$

Check how many Ex befor & after 2.0. Then chek within that how may + for each & − for each

split point = 2.0

Entropy: $\frac{1}{9}\left[\frac{-1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{9}\right] +$

$\frac{8}{9}\left[\frac{-3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8}\right]$

$= 0.8484$

split point = 3.5

Entropy $= \frac{2}{9}\left[\frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] +$

$\frac{7}{9}\left[\frac{-3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right]$

$= 0.222 + 0.766 = 0.988.$

Continue this way for all split points.

Gain = whole entropy − individual one.

Split 2.0 → $= 0.9911 - 0.8484 = 0.1427.$

Best split point after calculating gain.