Lec-23 04/03/21 → statistical measures (contd.)
→ Construction of DT

(2) Gain ratio :

→ C4.5, a successor of ID3, uses an extension to info. gain known as gain ratio.
→ It overcomes the bias factor encountered in info. gain
→ It applies a kind of normalization to info gain using split information.

$$\text{Gain Ratio} = \frac{\text{Gain}(A)}{\text{Split info}(A)}$$

$$\text{Split info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \log_2\left(\frac{|D_j|}{|D|}\right)$$

This value represents the potential information generated by splitting the training data set D, into v partitions, corresponding to v outcomes of a test attribute A.

* The attribute with the maximum gain ratio is selected as the splitting attribute.

Gain (Age) : 0.246 bits
Gain (Income) 0.029 bits
Gain (student) 0.151 bits
Gain (CR) = 0.048 "

split info (income) : low (04), med (06), high (04)

$$\text{split info (income)} = -\frac{4}{14}\log_2^{(4/14)} - \frac{6}{14}\log_2^{(6/14)} - \frac{4}{14}\log_2^{(4/14)}$$

$$= 1.556$$

$$\text{Gain ratio (income)} = \frac{0.029}{(1.556)} = 0.0188$$

(3) Gini Index :

→ This measure is used in CART.

→ It considers a binary split for each attribute.

→ The Gini Index measures the impurity of D

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} P_i^2$$

where P_i is the probability that a tuple in D belongs to class C_i & is estimated by $\frac{|C_i, D|}{|D|}$

$$\text{Gini}_A(D) = \frac{|D_1|}{|D|}\text{Gini}(D_1) + \frac{|D_2|}{|D|}\text{Gini}(D_2)$$

$$\Delta\text{Gini}(A) = \text{Gini}(D) - \text{Gini}_A(D)$$

The reduction in impurity that would be incurred by a binary split on a discrete or continuous valued attribute A.

→ The attribute that maximizes the redn. in impurity is selected as the splitting attribute.

Prob: Gini (D) = $1 - (9/14)^2 - (5/14)^2$

buy car Yes (09) buy car NO (04)

$= 0.459$

Gini (Income)

Income (low (04) med (06) high (04))

$\Rightarrow$ { low, med } (10) { high } (04) — categ (1)

{ med, high } { low } cat (2)

{ high, low } { medium } cat (3)

Gini { (low, med), (high) }

10 4

$\frac{10}{14}$ Gini (D1) + $\frac{4}{14}$ Gini (D2)

(low, medium)

Yes NO

07 03

high

Yes NO

02 02

$= \frac{10}{14}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) +$

$\frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$

$= 0.443$ ✓

Gini cat(2) { low, high } { med } = 0.458

Gini cat(3) { med, high } { low } = 0.45.

As Gini is to be minimized, therefore category (1) has the lowest Gini, which is the preferrable binary split.

Problem to construct DT. considering Entropy as the statistical measure.

	Name	hair	Height	weight	lotion	sunburn(res)
1	A	Blonde	avg	light	NO	pos
2	B	Blonde	tall	avg	Yes	neg
3	C	Brown	short	avg	Yes	neg
4	D	Blonde	short	avg.	NO	pos
5	E	Red	avg	heavy	NO	pos
6	F	Brown	tall	heavy	NO	neg
7	G	Brown	avg	heavy	NO	neg
8	H	Blonde	short	light	Yes	neg

Entropy (hair)
- blonde → 2pos, 2 neg
- brown → 3 neg
- red → 1 pos

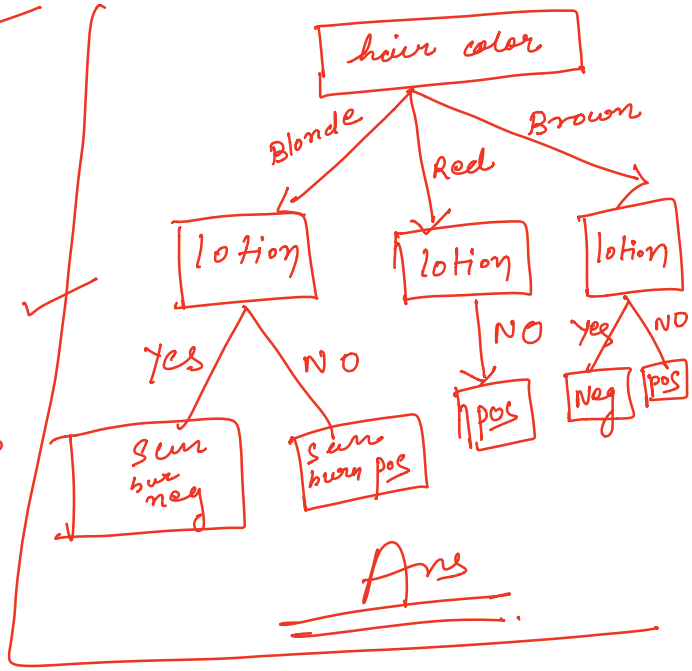
Entropy : $\frac{4}{8}\left(-\frac{2}{4}\log_2^{2/4} - 2/4\log_2^{2/4}\right) +$

$\frac{3}{8}\left(-0 - \frac{3}{3}\log_2^{3/3}\right) +$

$\frac{1}{8}\left(-\frac{1}{1}\log_2^{1/1} - 0\right)$

$= 0.5$

Entropy (hair):	0.5
Ent (height) :	0.69
Ent (weight) :	0.94
Ent (lotion) :	0.61



Ans

hair color, Blonde — ?

hair color, red — ?

hair color, brown — ?

< Blonde, height >

E (Blonde, height):

tall (1) (neg) | avg (01) (pos) | short (02) < pos(01) / neg(01)

$$\frac{1}{4}\left(-\frac{1}{1}\log_2^{1/1} - 0\right) + \frac{1}{4}\left(-\frac{1}{1}\log_2^{1/1} - 0\right) + \frac{2}{4}\left(-1/2\log_2^{1/2} - \frac{1}{2}\log_2^{1/2}\right)$$

$$= 0.54$$

Entropy (Blonde, wt)

$$= \frac{2}{4}\left(-1/2\log_2^{1/2} - 1/2\log_2^{1/2}\right) + \frac{2}{4}\left(-1/2\log_2^{1/2} - 1/2\log_2^{1/2}\right)$$

$$+ \frac{0}{4}\left(-0 - 0\right)$$

$$= 1$$

Entropy (blonde, lotion)

$E = \frac{2}{4}\left(-2/2 \log_2 \frac{2}{2} - 0\right) + \frac{2}{4}\left(-0 - \frac{2}{2}\log_2 \frac{2}{2}\right)$

$= 0$

Entropy (Blonde, hr) : 0.5

Entropy (Blonde, wt) = 1

Entropy (Blonde, lotion) = 0 ← ✓