Machine Learning

Quiz - 1 - Decision Trees

- Kinshuk Vasisht
- Roll Number : 19
- M.Sc. Computer Science

Q3. b) Instances :

f_1	f_2	L
10	3	1
6	3	1
2	7	0
9	6	1
4	6	0

Given splits : $f1 > 2$

$f1 > 4$

$f2 > 3$

$f2 > 6$

Now, Information Gain $(\Delta_{info}) = \phi_{parent} - \phi_{split}$

(Node entropy for parent

− Split entropy based on decision).

$$\Rightarrow \cancel{\Delta_{info} = -\sum_{i=1}^{2} p_m^i \log_2(p_m^i) + \sum_{i=1} \frac{N_{mj}^i}{N_m^i} \sum}$$

$$\Delta_{info} = \underbrace{-\sum_{i=1}^{2} p_m^i \log_2(p_m^i)}_{Node} + \underbrace{\sum_{j=1}^{2} \frac{N_m^j}{N_m} \sum_{i=1}^{2} p_{mj}^i \log p_{mj}^i}_{Split}$$

where p_m^i = probability of class i at node m

p_{mj}^i = probability of class i at node m, branch j

Now, $\phi_{parent} = -\left[\frac{2}{5}\log\left(\frac{2}{5}\right) + \frac{3}{5}\log\left(\frac{3}{5}\right)\right]$

1) Now, split entropy for $f_1 > 2$

	L = 0	L = 1
$f_1 > 2$	1	2
$f_1 \leq 2$	1	1

$$\therefore \phi_{f_1>2} = -\frac{3}{5}\sum_i p_{m1}^i \log p_{m1}^i - \frac{2}{5}\sum p_{m2}^i \log p_{m2}^i$$

$$\Rightarrow -\frac{3}{5}\left[\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}\right]-\frac{2}{5}\left[\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right]$$

$$= +\frac{3}{5}\left[\frac{\log_2 3}{3}+\frac{2\log_2 1\cdot5}{3}\right]+\frac{2}{5}\times 1 =$$

2) Now, split entropy for $f_1 > 4$

	$L=0$	$L=1$
$f_1>4$	0	3
$f_1\leq 4$	2	0

$$\therefore \phi_{f_1>4} = -\frac{3}{5}\sum_i p^i_{m1}\log_2(p^i_{m1}) - \frac{2}{5}\sum_i p^i_{m2}\log_2(p^i_{m2})$$

$$\Rightarrow \phi_{f_1>4} = -\frac{3}{5}\left[\frac{0}{3}\log_2\frac{0}{3}+\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right]-\frac{2}{5}\left[\frac{2}{2}\log_2\frac{2}{2}+\frac{0}{2}\log_2\frac{0}{2}\right]$$

$$= (3/5)\cdot 0+(2/5)\cdot 0 = \underline{0}$$

3) Now, split entropy for $f_2 > 3$:

	$L=0$	$L=1$
$f_2>3$	1	1
$f_2\leq 3$	1	2

$$\therefore \phi_{f_2>3} = -\frac{2}{5}\sum_i p^i_{m1}\log_2(p^i_{m1}) - \frac{3}{5}\sum_i p^i_{m2}\log_2(p^i_{m2})$$

$$\Rightarrow \phi_{f_2>3} = -\frac{2}{5}\left[\frac{1}{2}\lg\left(\frac{1}{2}\right)+\frac{1}{2}\lg\left(\frac{1}{2}\right)\right]-\frac{3}{5}\left[\frac{1}{3}\lg\left(\frac{1}{3}\right)+\frac{2}{3}\lg\left(\frac{2}{3}\right)\right]$$

$$= +\frac{2}{5}\times 1 + \frac{3}{5}\left[\frac{\lg(3)}{3}+\frac{2\lg(1\cdot5)}{3}\right] =$$

4) Now, split entropy for $f_2 > 6$:

	$L=0$	$L=1$
$f_2>6$	1	0
$f_2\leq 6$	1	3

$$\phi_{f_2>6} = -1/5\sum_i p^i_{m1}\lg(p^i_{m1}) - 4/5\sum_i p^i_{m2}\lg(p^i_{m2})$$

$$\Rightarrow \phi_{f_2>6} = -\frac{1}{5}\left[\frac{1}{1}\lg\left(\frac{1}{1}\right)+\frac{0}{1}\lg\left(\frac{0}{1}\right)\right]-\frac{4}{5}\left[\frac{1}{4}\lg\left(\frac{1}{4}\right)+\frac{3}{4}\lg\left(\frac{3}{4}\right)\right]$$

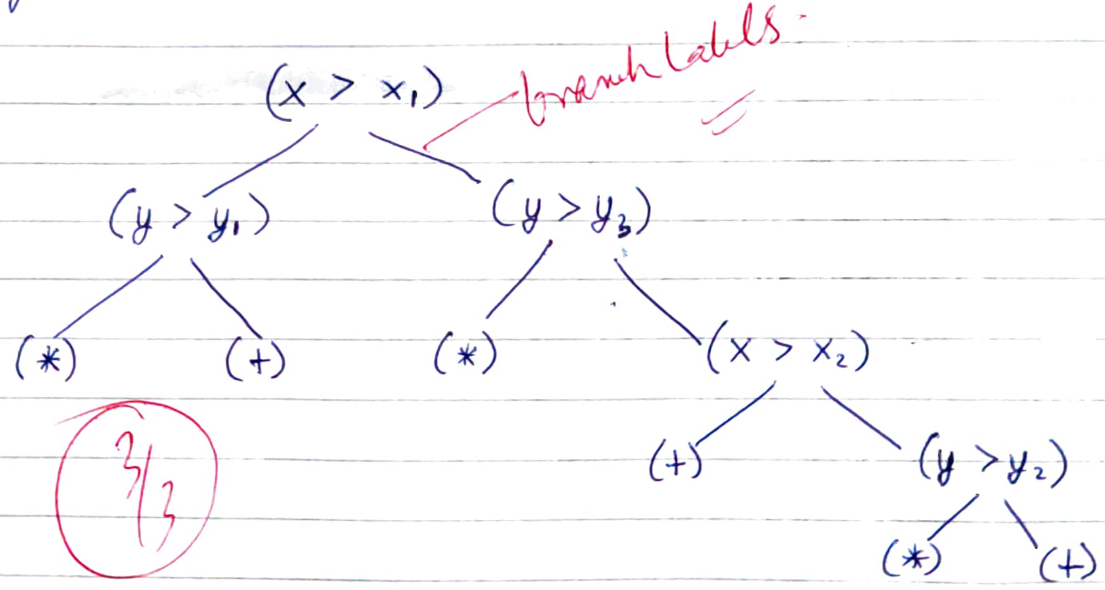
$$= \frac{1}{5}\times 0 + \frac{4}{5}\left[\frac{2}{4}+\left(-\frac{3}{4}\lg\left(\frac{4}{3}\right)\right)\right] =$$ 4/4

4/4

Now, split entropy for $f_1 > 4$ is the lowest (0). Corresponding to the same, the gain obtained by subtracting split entropy from the node entropy of the parent is the highest.

$\therefore$ Best Split : $f_1 > 4$ $(f_1 > 4)$

Q5 Every node of the ~~the~~ decision tree makes a split over an attribute different from that of its parent, based on the best split criteria (obtaining lowest heterogeneity). Following the given splits, the primary split is made over $x > x_1$, then on $\cancel{y_1} < y$ & $y > y_3$ on left & right subtrees. So we have — :



3/3