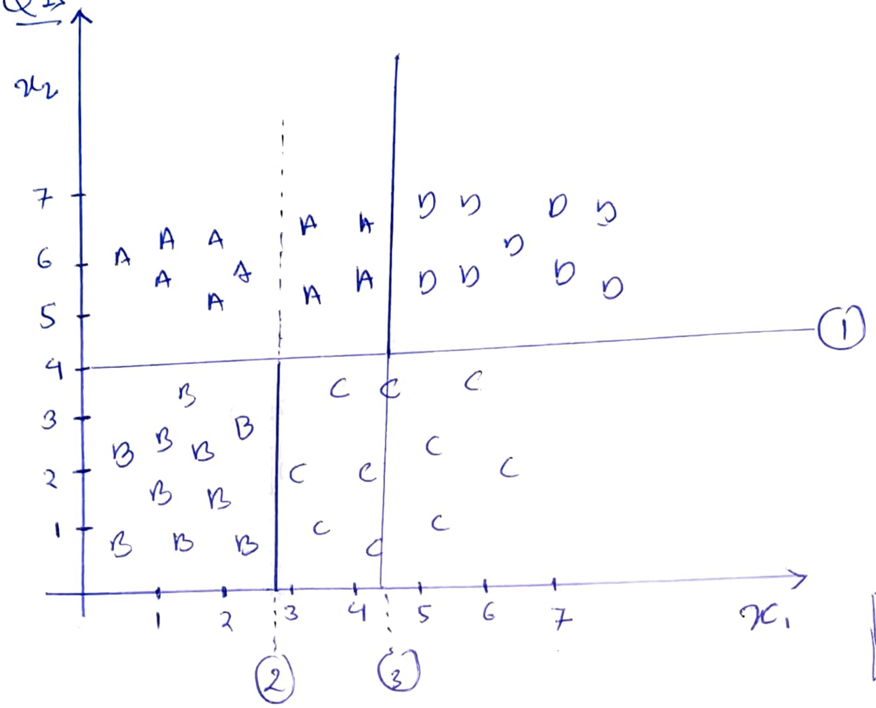
SMAI - Quiz (on Decision Trees)

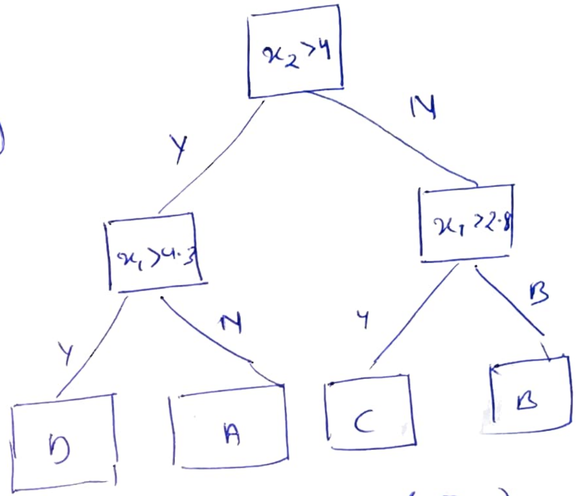
Name : Aditya Kumar Singh

ID : 2021701010

Q.1

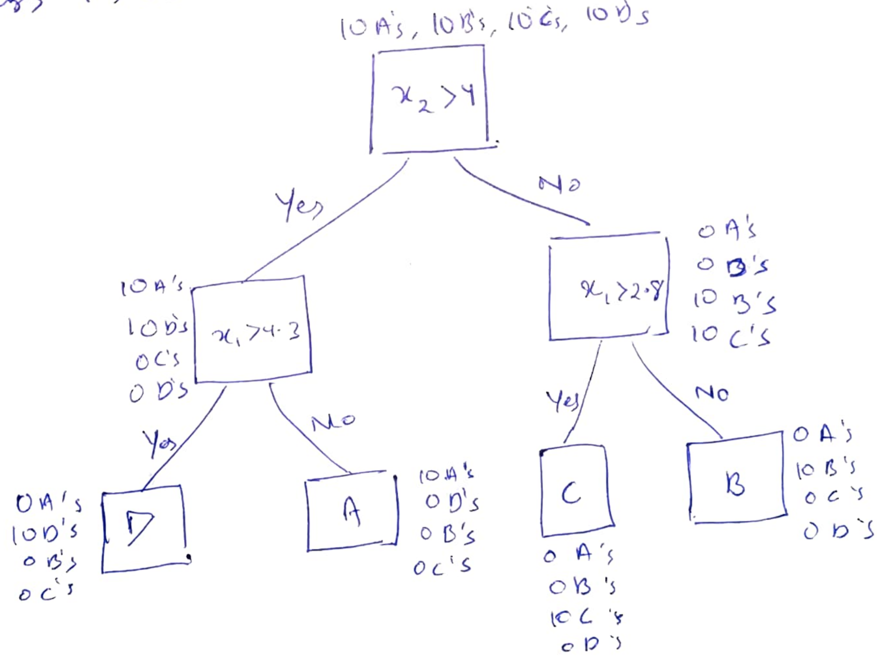


Given Decision Tree structure :



Calculate Information Gain, Δi, for each split (i.e, (1), (2) & (3)).

Ans Let's first denote the no. of examples at each node with all the class information.



So, the Entropy at root node (or the node with $x_2>4$ decision) $= -\sum_{i=1}^{4} p_i \log(p_i)$

p_i = i^{th} class probability

$$H(root) = H(x_2 > 4) := \text{Entropy for Decision } x_2 > 4$$

$$= -\sum_{i=1}^{4} P_i \log P_i$$

$$= -\sum_{i=1}^{4} \frac{1}{4} \log_2 \frac{1}{4} \quad \left(\text{as } P_A = P_B = P_C = P_D = \frac{10}{40} = \frac{1}{4}\right)$$

$$= -4 \cdot \frac{1}{4}\left(\log_2 1 - \log_2 2^2\right) = 2 \quad \text{--- (1)}$$

Similarly:

$$H(x_1 > 4.3) = H(③) = -\sum_{i=1}^{4} P_i \log P_i$$

$$= -\left[\frac{10}{20}\log_2\frac{10}{20} + \frac{10}{20}\log_2\frac{10}{20}\right]$$

$$= 1 \quad \text{--- (2)} \quad \left(\text{where } P_A = P_D = \frac{10}{20+0+0} = \frac{1}{2}\right.$$

$$P_A = \frac{\#A}{\#A + \#B + \#C + \#D}$$

$$P_D = \frac{\#D}{\#A + \#B + \#C + \#D}$$

or we can simply say, to calculate the entropy we just look at the data points that fall in $\boxed{x_1 > 4.3}$ node after splitting at root node.

$$H(x_1 > 2.8) = H(②)$$

$$= -\left[\cancelto{0}{P_A \log P_A} + \cancelto{0}{P_B \log P_B} + P_C \log P_C + P_D \log P_D\right]$$

$$= -\left[0 + 0 + \frac{10}{20+0}\log\frac{10}{20+0} + \frac{10}{20+0}\log\frac{10}{20+0}\right] = -2 \cdot \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1 \quad \text{--- (3)}$$

$$H(D) = -\left[\cancelto{0}{P_A \log P_A} + \cancelto{0}{P_B \log P_B} + \cancelto{0}{P_C \log P_C} + P_D \log P_D\right]$$

$$= -\left[0 + 0 + 0 + \frac{10}{10+0}\log_2\frac{10}{10}\right] = 0 \quad \text{--- (4)}$$

Similarly: $H(A) = H(C) = H(B) = 0$ (as all they contain only data-points of single class i.e., all are pure nodes) → Hence no mixing of classes ⇒ entropy (confusion) is '0'.

---

let's calculate information gain Δ_i at each node:

$$\Delta(x_2 > 4) = IG(x_2 > 4) = H(x_2 > 4) - \left[P(D_1 = x_2 > 2.8)\, H(x_1 > 2.8) + P(D_3 = x_1 > 4.3)\, H(x_1 > 4.3)\right]$$

$$= 2 - \left[\frac{20}{40} \cdot 1 + \frac{20}{40} \cdot 1\right] \quad (\text{from } (1), (2) \,\&\, (3))$$

$$\Delta(x_2>4) = \underset{\substack{\downarrow \\ \text{Before split}}}{2} - \underset{\substack{\searrow \\ \text{After split}}}{1} = 1$$ (where $\frac{20}{40}$ = probability of datasets at each child nodes of root node.) — ⑤

Similarly

$\Delta(③) = \Delta(x_1>4.3)$ = Entropy before split at (D_3 = Decision 3 $= x_1>4.3$) − Entropy after split at (D_2 ~~= Decision 2 = x_1~~ Both D & A node)

$$= 1 - \left[\frac{10}{20}\cdot 0 + \frac{10}{20}\cdot 0\right] = 1 \text{ — ⑥}$$

$\frac{10}{20}$ ↓ P(datasets at D); 0 → H(D); $\frac{10}{20}$ → P(datasets at A)

$$\Delta(②) = \Delta(x_1>2.8) = \underbrace{H(②)}_{\text{Entropy before split}} - \underbrace{\left[\frac{10}{20}\cdot H(C) + \frac{10}{20}\cdot H(B)\right]}_{\text{Entropy after split}}$$

$$= 1 - \left[0\cdot\frac{10}{20} + 0\cdot 10/20\right] = 1 \text{ — ⑦}$$

$$\Delta(D) = \Delta(A) = \Delta(C) = \Delta(B) = 0 \text{ — ⑧}$$

($\because$ $H(D) = H(A) = H(C) = H(B) = 0$ & all they have are "null" as child nodes where anyway $p_A = p_B = p_C = p_D = 0$ as no split is happening at these leaf nodes. e.g

D: 10 D → 0 D [null], 0 D [null]

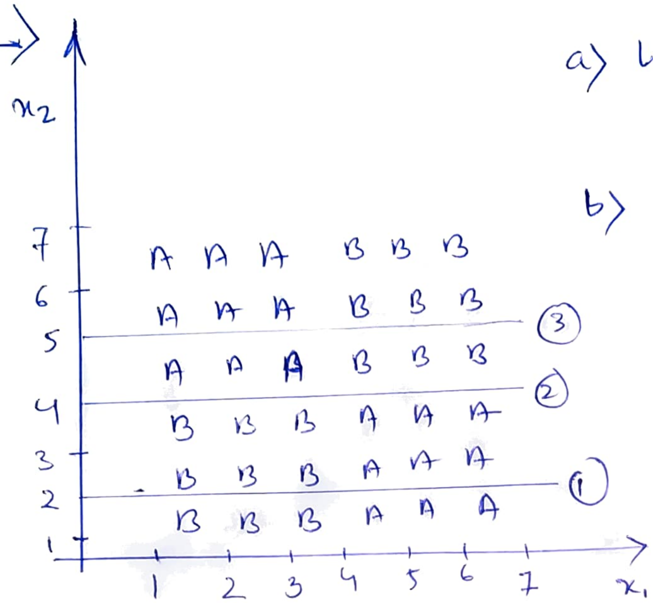
$\Rightarrow \Delta(D) = H(D) - \sum[P(\text{sub-split})\, H(\text{sub-split})]$

$= 0 - 0$

$= 0$.

Nodes	H = Entropy	Δ = IG = Information Gain
① = $x_2>4$	2	1
③ = $x_1>4.3$	1	1
② = $x_1>2.8$	1	1
D	0	0
A	0	0
C	0	0
B	0	0

Q-3)



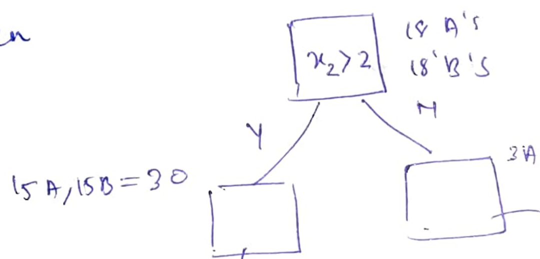
a) What is the problem with this data.

b) How can we overcome this?

Ans

① Say we start our root node with $x_2 > 2$ (i.e. classifier ①)

then



$H(x_2>2) = -\left[\frac{18}{36}\log_2\frac{18}{36} + \frac{18}{36}\log_2\frac{18}{36}\right]$

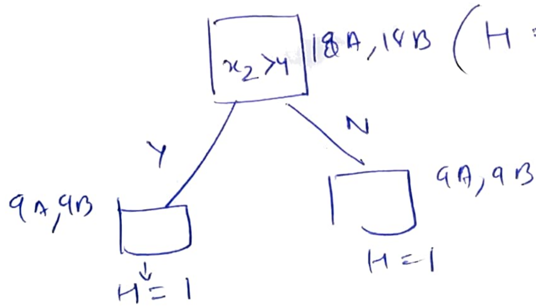
$= 1$

$\rightarrow H(\ ) = -\left(\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{3}{6}\log_2\frac{3}{6}\right) = 1$

$H(\ ) = -\left(\underbrace{\frac{15}{30}\log_2\frac{15}{30}}_{A} + \underbrace{\frac{15}{30}\log_2\frac{15}{30}}_{B}\right) = 1$

~~$\Delta(x_2)$~~ $\Rightarrow \Delta(x_2>2) = 1 - \left(\frac{30}{36}\cdot 1 + \frac{6}{36}\cdot 1\right) = 0.$

② Similarly, ~~for~~ if we have taken decision $(x_2>4)$, then



18A, 18B $(H = 1)$, $\Delta = 0$

& so on.

So what we observed, that for "any" horizontal (not supposedly inbetn them) split, we're passing through datapoints rather getting "**even-split**" i.e., each class have equal number of candidates at child nodes.

5

→ And since it's two class-classification, this even-split results in '0' Entropy (i.e., for any node/Decision having data with ~~containing~~ equal no. of labels from each class (of course, for binary/two classes)) ~~like~~ → results in 'zero' entropy) → e.g. at child node of $x_2 > 4$, 9A, 9B ⇒ even splits of data at this node into two labels ⇒ H( ) = 0 & so on.

→ Not only horizontal, for all sorts of vertical (parallel to x_2) decision boundaries would produce the same result. This is due to the fact that;

1) We're constrained to choose decision boundary parallel to axes.

2) Dataset are also well-aligned with axes.

3) Further they are "diagonally" fused/mixed up → which ensures, results will be same for both vertical & horizontal decision boundaries to start with.

→ Hence, the root node decision no matter what decision boundary (parallel to axes) splits the data into two halves for which the "proportion" of classes in both remain the same → which accounts for "Entropy after split" = 1

→ Now since, we have equal no. of both class to begin with, the Entropy of root node, will always be = 1.

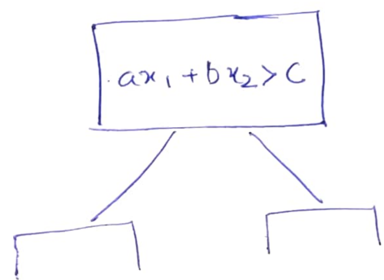
⇒ Information Gain = '0' (always for any vertical/Horizontal decision) → further splitting/construction of trees would STOP. [∵ Info-gain varies bet^n [0,1], with '0' nothing can be done].

How to overcome this? (b).

* Now the whole problem lies with the fact that both ~~are~~ our data ~~are~~ & decision boundaries to be constructed are axes-aligned.

* If somehow, we start with a "tilted" decision boundary

at root node (i.e. ~~combination of~~ function of combination of both features) then the ~~case~~ chance of getting equal "proportion" of labels in ~~both~~ the split is 'zero'. In other words, 'no' even split can take place which implies the ~~IG =~~ ~~Informati~~ Entropy won't be 1 ⇒ $IG \neq 0$ (won't be 0).



$a \neq 0$
$b \neq 0$
$c \in \mathbb{R}$

⇒ Now with above structure & with possible optimisation of 'a', 'b' & 'c', we can obtain such a decision boundary that outputs ~~max~~ min entropy (= less fusion among classes = less confusion) ~~which~~ which is indirectly 'max' information gain.

⇒ Once that set, we can start of our journey for constructing DT further from root node.
(e.g. C4.5 DT algo invokes such possibilities).