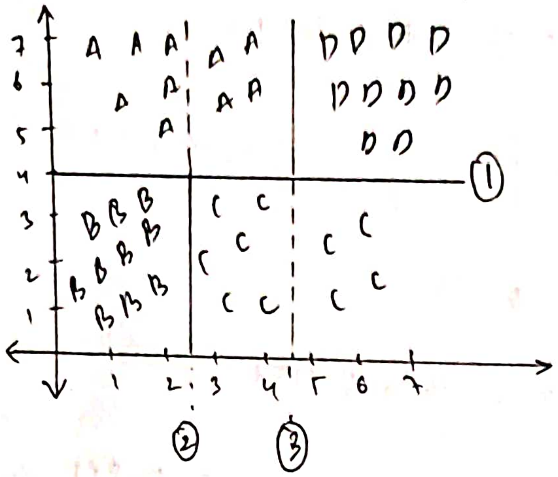
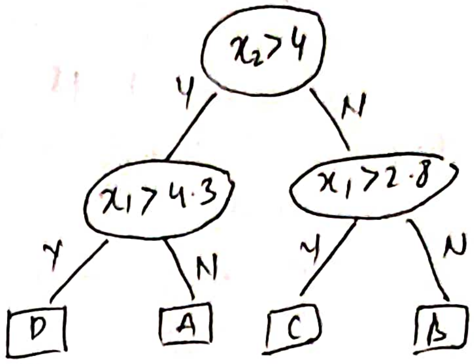
Name : Soumya. Jahagirdar
Roll no: 2021701015

QUIZ

1. What are the information gains corresponding to splits (1), (2) & (3).





Entropy

$$I(D) = -\sum_{i=1}^{c} P_i \log P_i$$

Information Gain

$$\Delta I = I(D) - \frac{n_L}{n} I(D_l) - \frac{n_r}{n} I(D_R)$$

$$I(D) = -\left[\frac{10}{40}\log_2\frac{10}{40} + \frac{10}{40}\log_2\frac{10}{40} + \frac{10}{40}\log_2\frac{10}{40} + \frac{10}{40}\log_2\frac{10}{40}\right]$$

$$= -\left[4\times\left[\frac{10}{40}\log_2\frac{10}{40}\right]\right] = -\left[4\times 0.25\frac{(\log 0.25)}{(\log 2)}\right] : 2.$$

$\therefore I(D) = 2$

For split(1)

$$I(D_l) = -\left[\frac{10}{20}\log_2\frac{10}{20} + \frac{10}{20}\log_2\frac{10}{20}\right] = -\left(2\times\frac{10}{20}\log_2\frac{10}{20}\right) = 1.$$

$$I(D_r) = -\left[\frac{10}{20}\log_2\frac{10}{20} + \frac{10}{20}\log_2\frac{10}{20}\right] = \left[2\times\frac{10}{20}\log_2\frac{10}{20}\right] = 1$$

$$\Delta I_{split1} = 2 - \frac{20}{40}\times 1 - \frac{20}{40}\times 1 = 2 - 0.5 - 0.5 = 1$$

For split 2 :

$$I(D_l) = -\left[\frac{10}{24}\log_2\frac{10}{24} + \frac{10}{24}\log_2\frac{10}{24} + \frac{4}{24}\log_2\frac{10}{24}\right]$$

$$= -[-0.5264 + 0.5264 + 0.429]$$

$$= 1.48$$

$$I(D_r) = -\left[\frac{10}{16}\log_2\frac{10}{16} + \frac{6}{16}\log_2\frac{6}{16}\right] = -\left[0.625 \log 0.625 + 0.375 \log 0.375\right]$$

$$= -[-0.423 - 0.530] = 0.953$$

$$\Delta I_{split_2} = 2 - \frac{24}{240} * 1.48 - \frac{16}{40} * 0.953$$

$$= 2 - 0.6 * 1.48 - 0.4 * 0.953 = 2 - 0.887 - 0.3812$$

$$= 2 - 1.2692 = 0.7308$$

For split 3 :

$$I(D_l) = -\left[\frac{10}{14}\log_2\frac{10}{14} + \frac{4}{14}\log_2\frac{4}{14}\right] = -[-0.347 - 0.516] = 0.8634$$

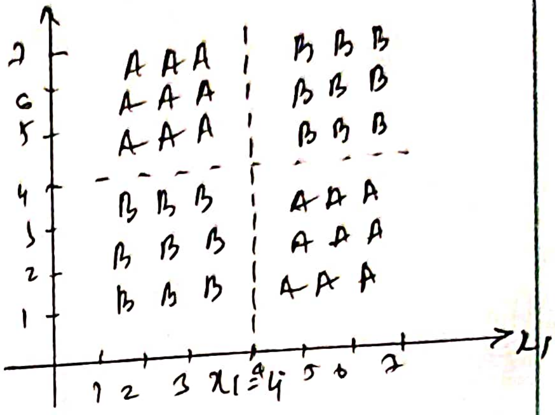
$$I(D_r) = -\left[\frac{10}{26}\log_2\frac{10}{26} + \frac{10}{26}\log_2\frac{10}{26} + \frac{6}{26}\log_2\frac{6}{26}\right] = -[-1.06 - 0.487] = +1.54$$

$$\Delta I_{splits3} = 2 - \frac{14}{40} \times 0.8634 - \frac{26}{40} \times 1.54 = 2 - 0.3021 - 1.$$

$$= 0.6979$$

$\therefore$ For the Split 1 at $x_2 > 4$ we obtain the maximum information Gain ΔI :

2) what is the problem with finding the split ? How can we overcome this problem ?



ans: Consider the I for entire data.

$$I(D) = -2 \times \frac{18}{36} \log_2\left(\frac{18}{36}\right)$$

$$= -2 \times \frac{18}{36} \times (-1)$$

$$= 1.$$

For split at $x_1 = 4$. the values of info will be

$I(D_L) = -\left[2 \times \frac{9}{18} \log \frac{9}{18}\right] = 1$ similarly $I(D_R) = 1$

i.e Info Gain : $1 - \frac{18}{36} \times 1 - \frac{18}{36} \times 1 = 1 - 0.5 - 0.5 = 0$.

As there is no information Gain, we cannot find anappropriate split

As the data is not linearly separable, one of the solutions as mentioned in the textbook which states that we should not allow splits that are parallel to the feature axis.

And probably even if we doallow, the splits would be something like $x_1 < 4$ and $x_2 < 4$, which ~~creatornot~~ is not possible. with normal decision boundaries.