1) Entropy and decision trees

1.1) Entropy

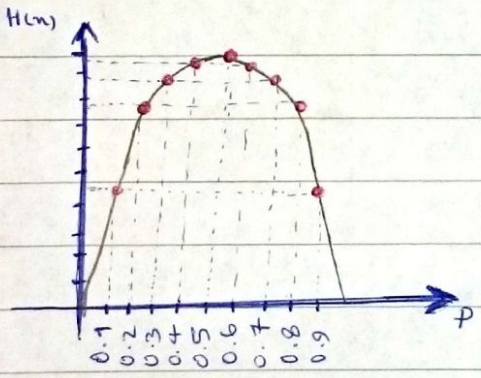
1) $p = 0.1 \Rightarrow H(x) = 0.1 \times \log_2\left(\frac{1}{0.1}\right) + (1-0.1)\log_2\left(\frac{1}{1-0.1}\right) = 0.46$

$p = 0.2 \Rightarrow H(x) = 0.72$ $\quad$ $p = 0.3 \Rightarrow H(x) = 0.88$

$p = 0.4 \Rightarrow H(x) = 0.97$ $\quad$ $p = 0.5 \Rightarrow H(x) = 1$

$p = 0.6 \Rightarrow H(x) = 0.97$ $\quad$ $p = 0.7 \Rightarrow H(x) = 0.88$

$p = 0.8 \Rightarrow H(x) = 0.72$ $\quad$ $p = 0.9 \Rightarrow H(x) = 0.46$



- When the probabilities for taking both values are equal, the disorder is larg and correspondingly the entropy value is large; both are in the highest state. On the other hand, when X takes one of the values in most of the times, the disorder is low and the value of p will be close to 0 or 1, and the entropy value would be Low.

2) – $p_n = \frac{1}{N}$ $\qquad H(x) = \sum_{n=0}^{N-1} p_n \log_2\left(\frac{1}{p_n}\right)$

$$H(x) = \sum_{n=0}^{N-1} \frac{1}{N} \log_2(N) = \log_2(N) \quad (2)$$

– $f : \log_2(\,)$

$\lambda_n : p_n$

$u_n : \frac{1}{p_n}$

(1)

$$H(x) = \sum_{n=0}^{N-1} p_n \log_2\left(\frac{1}{p_n}\right)$$

$$H(x) = \sum_{n=0}^{N-1} \lambda_n \; f \; u_n$$

$$H(x) \overset{(1)}{=} \sum_{n=0}^{N-1} \lambda_n f(u_n)$$

– $f\left(\sum_{n=0}^{N-1} \lambda_n u_n\right) \overset{(1)}{:} \log_2\left(\sum_{n=0}^{N-1} p_n \frac{1}{p_n}\right) = \log_2\left(\sum_{n=0}^{N-1} 1\right) = \log_2(N)$

Jensen's inequality: $\sum_{n=0}^{N-1} \lambda_n f(u_n) \leq f\left(\sum_{n=0}^{N-1} \lambda_n u_n\right)$

$\Downarrow$ (1)

$$H(x) \leq \log_2(N)$$

(1) and (2) $\Longrightarrow$ entropy of x is maximal if and only if x follows a uniform distribution.

1.2) Relation between entropy and information

1.2.1) Arresting a criminal

1) $p = \frac{1}{5 \times 80}$ $q = \frac{5 \times 80 - 1}{5 \times 80}$ $H(x) = p \log_2(\frac{1}{p}) + q \log_2(\frac{1}{q})$

$= 0.0252$

2) $p = \frac{1}{80}$ $q = \frac{80 - 1}{80}$ $H(x) = 0.0269$

3) $p = \frac{1}{30}$ $q = \frac{30 - 1}{30}$ $H(x) = 0.2108$

4) The information that officers recieve increase the uncertainty; as officer A is more certain about the innocence of every randomely chosen sampel. The officer B is less certain than officer A, and officer C is less certain than officer B. Thus, in case of officer A we have more information, and for officer B we have less information regarding A, but more regarding C. Therefore, in the case of officer A we have less entropy, and in case of officer B we have more entropy than case A but less than case C.

1.2.2) Guessing Cards

1) $P(X=x) = \frac{1}{52}$ for all x

$$H(x) = \sum_{n=0}^{52-1} \frac{1}{52} \log_2(52) = \log_2(52) = 5.7$$

2) $P(\text{Ey}) = \frac{4}{52}$

$$P(y) = \begin{cases} \frac{4}{52} & y = \text{true} \\ \frac{48}{52} & y = \text{false} \end{cases}$$

~~$P(X|Y=y) = \frac{1}{4}$~~

3) $$p(x|y=\text{true}) = \begin{cases} \frac{1}{4} & \text{for ace } x\text{s} \\ 0 & \text{for all not ace } x\text{s} \end{cases} \qquad p(x|y=\text{false}) = \begin{cases} \frac{1}{48} & \text{for all not ace } x\text{s} \\ 0 & \text{for ace } x\text{s} \end{cases}$$

$$H(x|y=\text{true}) = \sum_{i=0}^{52-1} p(x=x_i|y=\text{true}) \log_2\left(\frac{1}{p(x=x_i|y=\text{true})}\right)$$

$$= 4\left(\frac{1}{4}\log_2(4)\right) = 2$$

$$H(x|y=\text{false}) = \sum_{i=0}^{52-1} p(x=x_i|y=\text{false}) \log_2\left(\frac{1}{p(x=x_i|y=\text{false})}\right)$$

$$= 48\left(\frac{1}{48}\log_2(48)\right) = 5.58$$

4) if E_y happens then we have much information about the identity of the card (X), so $H(X|Y=true)$ is small; however, the probability of event Y happening is also small.

If the event Y (E_y) does not happen, then gussing the card would be much harder ([redacted] $H(X|Y=false)$ is large [redacted]), and the probability of E_y not happening is also large.

Therefor, knowing about E_y does not help considerably, which means it does not give us much information, so correspondingly we observe that $H(X|Y)$ is large.

large small small large large

$$H(X|Y) = P(Y=true) \cdot H(X|Y=true) + P(Y=false) \cdot H(X|Y=false)$$

5) $H(X|Y) = \frac{4}{52} \times 2 + \frac{48}{52} \times 5.58 = 5.3$ $\quad H(X) = 5.7$

$H(X|Y)$ is not much less than $H(X)$. It seems that knowing about E_y does not considerably affect the entropy of X.

6) $IG(X|y) = H(x) - H(x|y)$

$= 5.7 - 5.3 = 0.4$

$IG(x|y)$ should be large. When $IG(x|y)$ is large it means that knowing about y gives us much information that by knowing it the entropy of x would be much smaller.

7) Know about "is the top card spades" is more informative.

E_2: the top card is spades $\quad p(E_2) = \frac{1}{4}$

$H(X|Z) = \frac{1}{4}(\log_2(13)) + \frac{3}{4}(\log_2(39)) = 4.88$

$IG(X|Z) = H(X) - H(X|Z) = 5.7 - 4.88 = 0.82$

$IG(X|Z) > IG(X|Y)$

Knowing about the color is more informative.

1.3) Decision trees

1) $p(A) = \frac{5}{10}$ $p(B) = \frac{5}{10}$ $H(class) = \frac{5}{10}\log_2(\frac{10}{5}) + \frac{5}{10}\log_2(\frac{10}{5}) = 1$

$H(class|x) = p(x=1)\, H(class|x=1) + p(x=0)\, H(class|x=0)$

$= \frac{4}{10}\left[\frac{1}{4}\log_2(4) + \frac{3}{4}\log_2(\frac{4}{3})\right] + \frac{6}{10}\left[\frac{4}{6}\log_2(\frac{6}{4}) + \frac{2}{6}\log_2(\frac{6}{2})\right] = 0.87$

$H(class|y) = \frac{5}{10}\left[\frac{3}{5}\log_2(\frac{5}{3}) + \frac{2}{5}\log_2(\frac{5}{2})\right] + \frac{5}{10}\left[\frac{2}{5}\log_2(\frac{5}{2}) + \frac{3}{5}\log_2(\frac{5}{3})\right] = 0.97$

$H(class|z) = \frac{6}{10}\left[\frac{3}{6}\log_2(\frac{6}{3}) + \frac{3}{6}\log_2(\frac{6}{3})\right] + \frac{4}{10}\left[\frac{2}{4}\log_2(\frac{4}{2}) + \frac{2}{4}\log_2(\frac{4}{2})\right] = 1$

$IG(class|x) = H(class) - H(class|x) = 1 - 0.87 = 0.125$ ✓

$IG(class|y) = 1 - 0.97 = 0.029$

$IG(class|z) = 1 - 1 = 0$

- The first node uses X for classification.

X	Y	Z	Class
1	0	0	A
1	0	1	B
1	1	1	B
1	1	1	B

X	Y	Z	Class
0	0	0	B
0	1	1	A
0	0	1	A
0	1	1	A
0	1	0	A
0	0	0	B

$$H(\text{class}|Y, X=1) = \frac{2}{4}\left[0 + \frac{2}{2}\log_2(1)\right] + \frac{2}{4}\left[\frac{1}{2}\log_2(2) + \frac{1}{2}\log_2(2)\right]$$

$$= \frac{2}{4}\times 0 + \frac{2}{4}\times 1 = \frac{2}{4} = 0.5$$

$$H(\text{class}|Z, X=1) = \frac{3}{4}\left[1\log_2(1)\right] + \frac{1}{4}\left[1\log_2(1)\right] = 0$$

~~$IG(\text{class}|Y, X=1)$~~

$$H(\text{class}|X=1) = \frac{1}{4}\log_2(4) + \frac{3}{4}\log_2\left(\frac{4}{3}\right) = 0.81$$

$$IG(\text{class}|Y, X=1) = 0.81 - 0.5 = 0.31$$

$$IG(\text{class}|Z, X=1) = 0.81 - 0 = 0.81$$ ✓

- The next node when $X=1$ should use **Z** for classification

$$H(\text{class}|Y, X=0) = \frac{3}{6}\left[\frac{1}{3}\log_2(3) + \frac{2}{3}\log_2\left(\frac{3}{2}\right)\right] + \frac{3}{6}\left[\frac{3}{3}\log_2\left(\frac{3}{3}\right) + 0\right] = 0.45$$

$$H(\text{class}|Z, X=0) = \frac{3}{6}\left[\frac{1}{3}\log_2(3) + \frac{2}{3}\log_2(3/2)\right] + 3/6\left[1\log_2(1)\right] = 0.45$$

$$H(\text{class}|X=0) = \frac{4}{6}\log_2\left(\frac{6}{4}\right) + \frac{2}{6}\log_2\left(\frac{6}{2}\right) = 0.91$$

$$IG(\text{class}|Y, X=0) = 0.91 - 0.45 = 0.46$$

$$IG(\text{class}|Z, X=0) = 0.91 - 0.45 = 0.46$$

- both features in this step have the same value. we choose one randomely, e.g., Z.

$X=1, Z=1$

X	Y	Z	class
1	0	1	B
1	1	1	B
1	1	1	B

$H(\text{class} \mid X=1, Z=1) = 0$

$H(\text{class} \mid Y, X=1, Z=1) = \frac{2}{3} \times 0 + \frac{1}{3} \times 0 = 0$

$IG(\text{class} \mid Y, X=1, Z=1) = 0$

$X=1, Z=0$

X	Y	Z	class
1	0	0	A

$H(\text{class} \mid X=1, Z=0) = 0$

$H(\text{class} \mid Y, X=1, Z=0) = 0$

$IG(\text{class} \mid Y, X=1, Z=0) = 0$

- there would be no point in using Y here, since the entropy of the class in this setting befor and after using y would remain the same.

$X=0, Z=1$

X	Y	Z	class
0	1	1	A
0	0	1	A
0	1	1	A

$H(\text{class} \mid X=0, Z=1) = 0$

$H(\text{class} \mid Y, X=0, Z=1) = \frac{2}{3} \times 0 + \frac{1}{3} \times 0 = 0$

$IG(\text{class} \mid Y, X=0, Z=1) = 0$

$X=0, Z=0$

X	Y	Z	class
0	0	0	B
0	1	0	A
0	0	0	B

$H(\text{class} \mid X=0, Z=0) = \frac{1}{3} \log_2(3) + \frac{2}{3} \log_2\left(\frac{3}{2}\right) = 0.91$

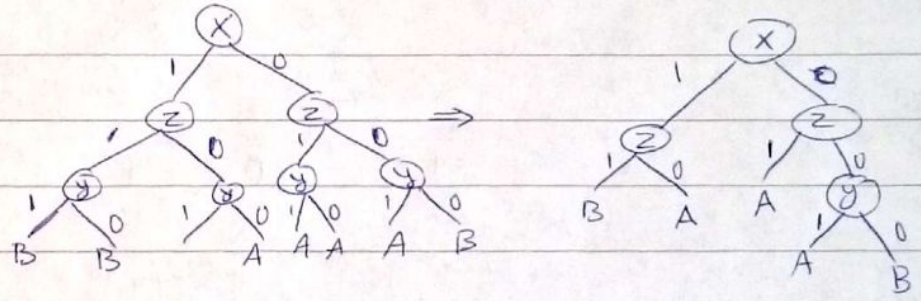
$H(\text{class} \mid Y, X=0, Z=0) = \frac{1}{3} \times 0 + \frac{2}{3} \times 0 = 0$

$IG(\text{class} \mid Y, X=0, Z=0) = 0.91$

- here Y helps, so we choose Y.

- So in the first level we use X, in the second level Z and in the third level, since we want to grow a full tree, we use Y.

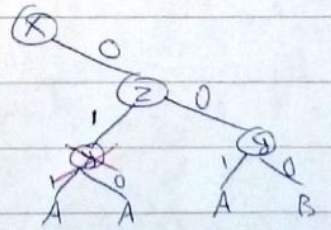
- Full decision tree:



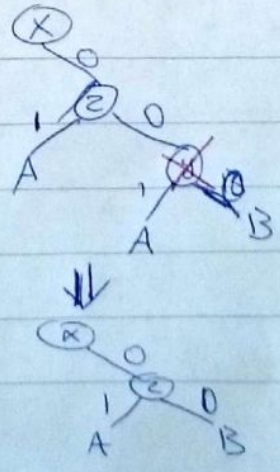
2) the only correctly classified sample in the pruning dataset is:

0 0 0 B (Accuracy: 20%)

- if we omit node Y when X=0 and Z=1 accuracy will remain the same.



- if we omit Y when X=0 and Z=0 Accuracy will be higher.

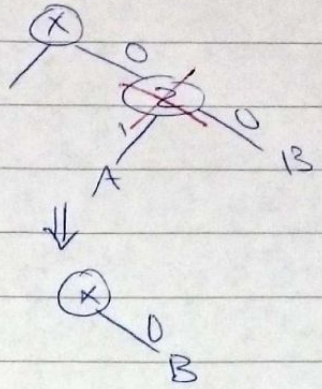


correctly classified samples:

0 0 0 B

0 1 0 B (Accuracy: 40%)

- if we omit z when x=0



then the accuracy will be even higher.

Correctly classified samples:

0	1	1	B
0	1	0	B
0	0	1	B
0	0	0	B

(Accuracy: 80%)

So according to the pruning dataset, the tree will be like this: