



I. Pen-and-paper

$$1) \quad H(M) = \sum_{i=1}^n -p(m_i) \log_2(p(m_i)) \quad IG(y_j) = H(z) - H(z|y_j)$$

$$H(z|y_j) = \sum_{i=1}^k \frac{|x_i|}{|x|} H(z|x_i)$$

$y_1 \geq 0,3$

D	y_1	y_2	y_3	y_4	y_{out}
x_6	0,3	0	1	0	B
x_7	0,76	0	1	1	A
x_8	0,86	1	0	0	A
x_9	0,93	0	1	1	C
x_{10}	0,47	0	1	1	C
x_{11}	0,73	1	0	0	A
x_{12}	0,89	1	2	0	B

$$\bullet H(y_{out} | y_1 \geq 0,3) =$$

$$= -\frac{3}{7} \cdot \log_2\left(\frac{3}{7}\right) - \frac{2}{7} \cdot \log_2\left(\frac{2}{7}\right) - \frac{2}{7} \cdot \log_2\left(\frac{2}{7}\right) \approx$$

$$\bullet H(y_{out} | y_1 \geq 0,3, y_3) =$$

$$\approx 1,251$$

$$= \frac{2}{7} \cdot H(y_{out} | y_1 \geq 0,3, y_3 = 0) + \frac{4}{7} H(y_{out} | y_1 \geq 0,3, y_3 = 1) + \frac{1}{7} H(y_{out} | y_1 \geq 0,3, y_3 = 2) =$$

$$= \frac{2}{7} \cdot 0 + \frac{4}{7} \left(-\frac{1}{4} \log_2\left(\frac{1}{4}\right) - \frac{1}{4} \log_2\left(\frac{1}{4}\right) - \frac{2}{4} \log_2\left(\frac{2}{4}\right) \right) + \frac{1}{7} \cdot 0 \approx 0,857$$

$$\bullet H(y_{out} | y_1 \geq 0,3, y_4) = \frac{4}{7} \cdot H(y_{out} | y_1 \geq 0,3, y_4 = 0) + \frac{3}{7} \cdot H(y_{out} | y_1 \geq 0,3, y_4 = 1) =$$

$$= \frac{4}{7} \cdot (1) + \frac{3}{7} \cdot \left(-\frac{1}{3} \log_2\left(\frac{1}{3}\right) - \frac{2}{3} \log_2\left(\frac{2}{3}\right) \right) \approx 0,965$$

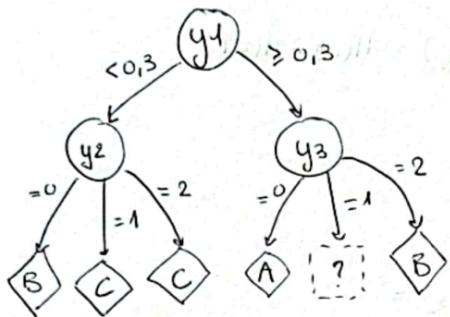
$$\bullet IG(y_{out} | y_1 \geq 0,3, y_i) = H(y_{out} | y_1 \geq 0,3) - H(y_{out} | y_1 \geq 0,3, y_i)$$

$$\bullet IG(y_{out} | y_1 \geq 0,3, y_2) = 1,557 - 1,251 \approx 0,306$$

$$\bullet IG(y_{out} | y_1 \geq 0,3, y_3) = 1,557 - 0,857 = 0,700$$

$$\bullet IG(y_{out} | y_1 \geq 0,3, y_4) = 1,557 - 0,965 = 0,592$$

$0,7 > 0,592 > 0,306$. Dendo assim, excluimos y_3 como nó já que leva ao maior ganho de informação.



- Para $y_3 = 0$ e $y_1 \geq 0,3$ verificam-se apenas observações da classe A para y_{out} pelo que será esse o nó folha
- Para $y_1 \geq 0,3$ e $y_3 = 2$ verificam-se apenas observações da classe B para y_{out} pelo que será esse o nó folha

- Para $y_1 \geq 0,3$ e $y_3 = 1$ temos quatro observações, o número mínimo necessário para expandir o nó.

$$y_1 \geq 0,3 ; y_3 = 1$$

D	y_1	y_2	y_3	y_4	y_{out}
x_6	0,3	0	1	0	B
x_7	0,76	0	1	1	A
x_9	0,93	0	1	1	C
x_{10}	0,47	0	1	1	C

$$H(y_{out} | y_1 \geq 0,3, y_3 = 1) =$$

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

$$= 1,5$$

$$H(y_{out} | y_1 \geq 0,3, y_3 = 1, y_2) =$$

$$= \frac{4}{4} \cdot H(y_{out} | y_1 \geq 0,3, y_3 = 1, y_2 = 0) =$$

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

$$= 1,5$$

$$H(y_{out} | y_1 \geq 0,3, y_3 = 1, y_4) = \frac{1}{4} \cdot H(y_{out} | y_1 \geq 0,3, y_3 = 1, y_4 = 0) + \frac{3}{4} \cdot H(y_{out} | y_1 \geq 0,3, y_3 = 1, y_4 = 1)$$

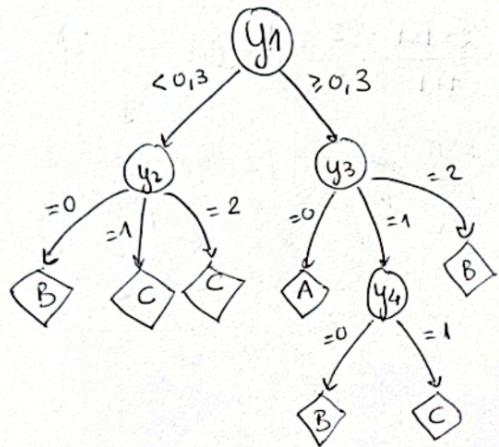
$$= \frac{1}{4} \times 0 + \frac{3}{4} \times \left(-\frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) - \frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right) \right) \approx 0,689$$

$$IG(y_{out} | y_1 \geq 0,3, y_3 = 1, y_i) = H(y_{out} | y_1 \geq 0,3, y_3 = 1) - H(y_{out} | y_1 \geq 0,3, y_3 = 1, y_i)$$

$$IG(y_{out} | y_1 \geq 0,3, y_3 = 1, y_2) = 1,5 - 1,5 = 0$$

$$IG(y_{out} | y_1 \geq 0,3, y_3 = 1, y_4) = 1,5 - 0,689 = 0,811$$

$0,811 > 0$. Escolhemos y_4 como nó já que leva ao maior ganho de informação.



- Para $y_1 \geq 0,3$, $y_3 = 1$ e $y_4 = 0$ verificam-se apenas observações da classe B pelo que será esse o nó folha
- Para $y_1 \geq 0,3$, $y_3 = 1$ e $y_4 = 1$ temos apenas três observações pelo que não iremos expandir o nó. Uma vez que se verificam mais ocorrências da classe C (duas de classe C contra uma de classe A), será esse o valor do nó folha

2) $\frac{y_{out}}{\hat{y}_{out}}$

x_i	y_{out}	\hat{y}_{out}
x_1	C	C
x_2	B	B
x_3	C	C
x_4	B	B
x_5	C	C
x_6	B	B
x_7	A	C
x_8	A	A
x_9	C	C
x_{10}	C	C
x_{11}	A	A
x_{12}	B	B

previsto

real

	A	B	C
A	2	0	0
B	0	4	0
C	1	0	5

3)

$$\text{precision}_A = \frac{TP}{TP + FP}$$

$$\text{recall}_A = \frac{TP}{TP + FN}$$

$$F_1 \text{ score} = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} =$$

$$= \frac{2 \times \text{Precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$\bullet \text{precision}_A = \frac{2}{2+0} = 1$$

$$\bullet \text{recall}_A = \frac{2}{2+1} = \frac{2}{3}$$

$$\bullet \text{precision}_B = \frac{4}{4+0} = 1$$

$$\bullet \text{recall}_B = \frac{4}{4+0} = 1$$

$$\bullet \text{precision}_C = \frac{5}{5+1} = \frac{5}{6}$$

$$\bullet \text{recall}_C = \frac{5}{5+0} = 1$$

$$\bullet F_{1A} = \frac{2 \times 1 \times \frac{2}{3}}{1 + \frac{2}{3}} = \frac{\frac{4}{3}}{\frac{5}{3}} = \frac{4}{5}$$

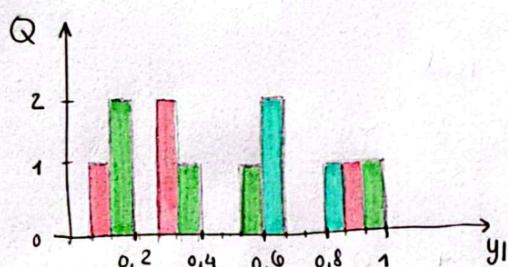
$$\bullet F_{1B} = \frac{2 \times 1 \times 1}{1 + 1} = \frac{2}{2} = 1$$

$$\bullet F_{1C} = \frac{2 \times \frac{5}{6} \times 1}{\frac{5}{6} + 1} = \frac{\frac{5}{3}}{\frac{11}{6}} = \frac{10}{11}$$

$$\frac{4}{5} < \frac{10}{11} < 1 \rightarrow F_{1A} < F_{1C} < F_{1B}$$

∴ A classe A tem o menor valor de F_1 .

4)



Divisão da raiz, usando as regras discriminantes:

$]0; 0,4[\rightarrow \text{classes B, C}$

$]0,4; 0,6[\rightarrow \text{classe C}$

$]0,6; 1[\rightarrow \text{classe A}$

- No intervalo $]0; 0,4[$ observam-se as classes B e C em igual quantidade
- Em $]0,4; 0,6[$ verifica-se apenas uma ocorrência da classe C
- Os valores da classe A concentram-se no intervalo $]0,6; 1[$

A divisão da raiz define-se, então, de acordo com os pontos acima.