

# Homework 1

Gomealo Herneses 103401  
Tomás Arédi 103239

1) Podemos ver que a árvore de decisão já foi começada. Tivemos trabalho c/ o dataset  $>0,4$ . Ficamos com:

	$y_1$	$y_2$	$y_3$	$y_4$	$y_{out}$
$x_6$	0,68	2	2	1	A
$x_7$	0,9	0	1	2	A
$x_8$	0,76	2	2	0	A
$x_9$	0,46	1	1	1	B
$x_{10}$	0,62	0	0	1	B
$x_{11}$	0,44	1	2	2	C
$x_{12}$	0,52	0	2	0	C

Para decidir qual a variável que iremos colocar temos que calcular o ganho para  $\{y_2, y_3, y_4\}$ . Tendo em conta que:

$$IG(y_{out} | y_i) = H(y_{out}) - H(y_{out} | y_i)$$

onde  $H \equiv$  entropia e é dada por  $H(y) = - \sum_{v \in y} \log_2(P_v) \cdot P_v$

→ Começemos por calcular  $H(y_{out})$ :

$$H(y_{out}) = \sum_{v \in y_{out}} \log_2(P_v) \cdot P_v = - \frac{3}{7} \log_2\left(\frac{3}{7}\right) - \frac{2}{7} \log_2\left(\frac{2}{7}\right) - \frac{2}{7} \log_2\left(\frac{2}{7}\right) \approx 1,56$$

→ Calcular  $H(y_{out} | y_i)$  para  $i = \{2, 3, 4\}$ :

$$H(y_{out} | y_2) = P_0 \cdot H(y_{out} | y_2=0) + P_1 \cdot H(y_{out} | y_2=1) + P_2 \cdot H(y_{out} | y_2=2)$$

$$= - \frac{3}{7} \cdot \left[ \frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) + \frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) + \frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) \right] = \frac{2}{7} \cdot \left[ \frac{1}{2} \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) \right]$$

$$- \frac{2}{7} \cdot [\log_2(1)] \approx 0,9650$$

$$\begin{aligned}
 H(Y_{\text{out}} | Y_3) &= P_0 \cdot H(Y_{\text{out}} | Y_3=0) + P_1 \cdot H(Y_{\text{out}} | Y_3=1) + \\
 &\quad + P_2 \cdot H(Y_{\text{out}} | Y_3=2) = \\
 &= -\frac{1}{7} \cdot \cancel{\log_2(1)}^0 - \frac{2}{7} \cdot \left[ \frac{1}{2} \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \log_2\left(\frac{1}{2}\right) \right] - \\
 &- \frac{4}{7} \cdot \left[ \frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) \right] \approx 0,8571
 \end{aligned}$$

$$\begin{aligned}
 H(Y_{\text{out}} | Y_4) &= P_0 \cdot H(Y_{\text{out}} | Y_4=0) + P_1 \cdot H(Y_{\text{out}} | Y_4=1) + P_2 \cdot H(Y_{\text{out}} | Y_4=2) \\
 &= -\frac{2}{7} \cdot \left[ \frac{1}{2} \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \cdot \log_2\left(\frac{1}{2}\right) \right] - \frac{3}{7} \cdot \left[ \frac{1}{3} \log_2\left(\frac{1}{3}\right) + \frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right) \right] - \\
 &- \frac{2}{7} \cdot \left[ \frac{1}{2} \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \log_2\left(\frac{1}{2}\right) \right] \approx 0,9650
 \end{aligned}$$

→ Calculate IG:

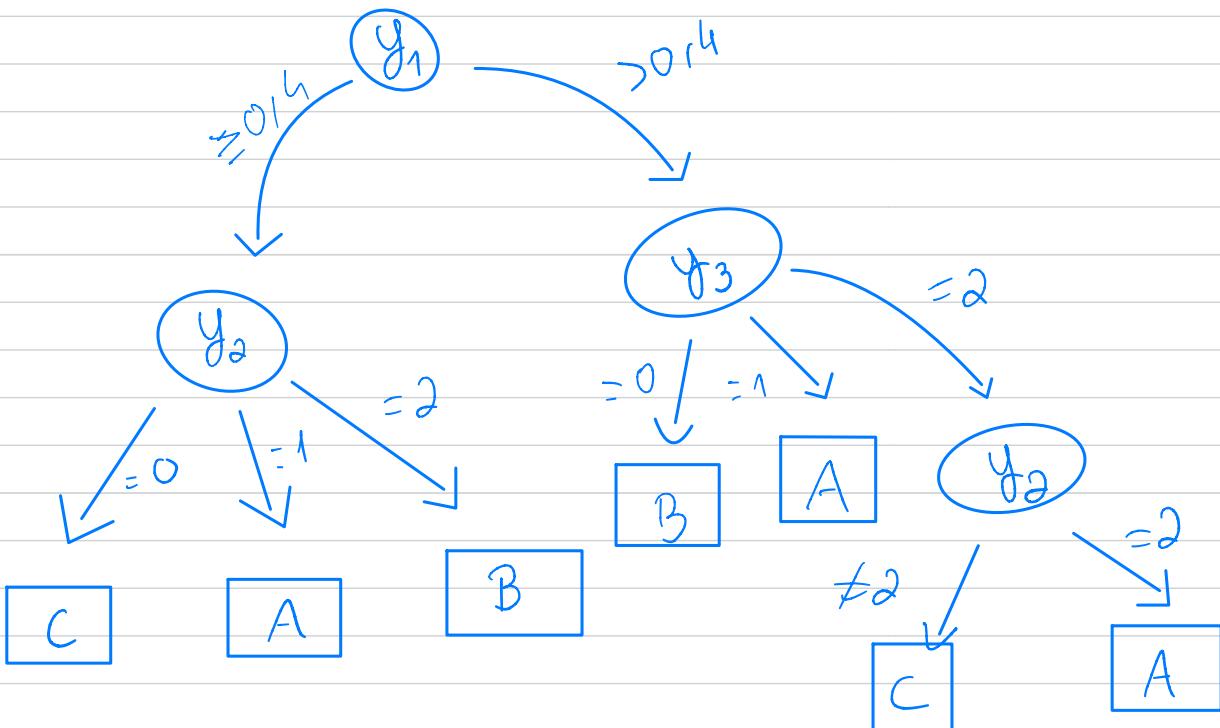
$$IG(Y_{\text{out}} | Y_2) = H(Y_{\text{out}}) - H(Y_{\text{out}} | Y_2) \approx 0,5917$$

$$IG(Y_{\text{out}} | Y_3) = H(Y_{\text{out}}) - H(Y_{\text{out}} | Y_3) \approx 0,6995$$

$$IG(Y_{\text{out}} | Y_4) = H(Y_{\text{out}}) - H(Y_{\text{out}} | Y_4) \approx 0,5917$$

Logo, já temos o próximo passo da árvore de decisão porque:

$$IG(y_{\text{out}} | y_3) > IG(y_{\text{out}} | y_2) = IG(y_{\text{out}} | y_4)$$



Conseguimos ver pelo dataset que quando  $y_3 = 0 \Rightarrow y_{\text{out}} = B$ , logo, podemos adicionar à árvore. Conseguimos também ver que quando  $y_3 = 1 \Rightarrow y_{\text{out}} = A \vee y_{\text{out}} = B$ . Por ser um empate utilizamos a regra ii) do enunciado e escolhemos A. Quando  $y_3 = 2$  temos 4 observações portanto seguimos a regra do enunciado i) e abrimos um novo nó. Só que agora ficámos com este dataset possível:

	$y_2$	$y_4$	$y_{\text{out}}$
$x_6$	2	1	A
$x_8$	2	0	A
$x_{11}$	1	2	C
$x_{12}$	0	0	C

Vamos ter de calcular de novo o ganho. Temos então:

$$\begin{aligned}
 I(G(Y_{\text{out}} | Y_2)) &= H(Y_{\text{out}}) - H(Y_{\text{out}} | Y_2) = \\
 &= -\frac{1}{2} \log_2\left(\frac{1}{2}\right) - \frac{1}{2} \log_2\left(\frac{1}{2}\right) + \cancel{\frac{1}{4} \cdot \log_2(1)} + \cancel{\frac{1}{4} \log_2(1)} + \cancel{\frac{1}{2} \cdot \log_2(1)} \\
 &= 1
 \end{aligned}$$

$$\begin{aligned}
 I(G(Y_{\text{out}} | Y_4)) &= H(Y_{\text{out}}) - H(Y_{\text{out}} | Y_4) = \\
 &= -\frac{1}{2} \log_2\left(\frac{1}{2}\right) - \frac{1}{2} \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \cdot \left[ \frac{1}{2} \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \log_2\left(\frac{1}{2}\right) \right] + \\
 &\quad + \cancel{\frac{1}{4} \cdot \log_2(1)} + \cancel{\frac{1}{4} \cdot \log_2(1)} = 0,5
 \end{aligned}$$

Como,  $I(G(Y_{\text{out}} | Y_2)) > I(G(Y_{\text{out}} | Y_4))$  colocamos  $Y_2$  na árvore de decisão. Olhando para o dataset é fácil de ver que se  $Y_2 = 0 \vee Y_2 = 1 \Rightarrow Y_{\text{out}} = C$  e que se  $Y_2 = 2 \Rightarrow Y_{\text{out}} = A$ .

Concluímos assim a árvore de decisão!

2)

Real

Legenda:

	A	B	C
A	TA	FA	FA
B	FB	TB	FB
C	FC	FC	TC

TA / TB / TC - True A / B / C  
 FA / FB / FC - False A / B / C

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

(4)

3)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{F-measure} = \frac{(\beta^2 + 1) \cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

↓ como vamos utilizar  $F_1$ , entao  $\beta=1$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision A} = \frac{4}{5} = 80\% \quad \text{Recall A} = \frac{4}{4} = 100\%.$$

$$\text{Precision B} = \frac{2}{2} = 100\%. \quad \text{Recall B} = \frac{2}{4} = 50\%.$$

$$\text{Precision C} = \frac{4}{5} = 80\%. \quad \text{Recall C} = \frac{4}{4} = 100\%.$$

$$F_1(A) = 2 \cdot \frac{0,8 \cdot 1}{1,8} = 0,8889 = 88,89\%.$$

$$F_1(B) = 2 \cdot \frac{1 \cdot 0,5}{1,5} = 0,6667 = 66,67\%.$$

$$F_1(C) = 2 \cdot \frac{0,8 \cdot 1}{1,8} = 0,8889 = 88,89\%.$$

A classe c/  $F_1$  menor é a classe B.

4)

	$y_1$	rank $y_1$	$y_2$	rank $y_2$	
$x_1$	0,24	3	1	8	$\frac{1+2+3+4+5+6}{6} = 3,5$
$x_2$	0,06	2	2	11	
$x_3$	0,04	1	0	3,5	
$x_4$	0,36	5	0	3,5	$\frac{7+8+9}{3} = 8$
$x_5$	0,32	4	0	3,5	
$x_6$	0,08	10	2	11	$\frac{10+11+12}{3} = 11$
$x_7$	0,9	12	0	3,5	
$x_8$	0,76	11	2	11	
$x_9$	0,46	7	1	8	
$x_{10}$	0,62	9	0	3,5	
$x_{11}$	0,44	6	1	8	
$x_{12}$	0,11	8	0	3,5	

$$\text{rank } \bar{y}_1 = \text{rank } y_1 = \bar{y}_1 = \frac{\sum_{i=1}^{12} x_i}{12} = 6,5$$

Coeficiente de Spearman = Pearson (rank  $y_1$ ; rank  $y_2$ )

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \quad \text{e} \quad \text{Cov}(y_1, y_2) = \frac{\sum (y_1 - \bar{y}_1) \cdot (y_2 - \bar{y}_2)}{n-1}$$

$$\text{Pearson}(y_1, y_2) = \frac{\text{Cov}(y_1, y_2)}{\sigma_{y_1} \cdot \sigma_{y_2}}$$

Utilizamos  $n-1$  pq. apenas temos acesso a uma amostra e não à população inteira.

de qualquer das formas, aplicando a fórmula de Pearson isso é irrelevante. Vamos por que?

(6)

$$\begin{aligned}
 \text{Pearson}(y_1, y_2) &= \frac{\sum (y_1 - \bar{y})(y_2 - \bar{y})}{n-1} = \\
 &= \frac{\sqrt{\sum (y_1 - \bar{y})^2} \cdot \sqrt{\sum (y_2 - \bar{y})^2}}{\sqrt{\sum (y_1 - \bar{y})^2} \cdot \sqrt{\sum (y_2 - \bar{y})^2}} = \\
 &= \frac{\sum (y_1 - \bar{y}) \cdot (y_2 - \bar{y})}{\sqrt{n-1} \cdot \sqrt{\sum (y_1 - \bar{y})^2} \cdot \sqrt{\sum (y_2 - \bar{y})^2}} = \\
 &= \frac{\sum (y_1 - \bar{y}) \cdot (y_2 - \bar{y})}{\sqrt{\sum (y_1 - \bar{y})^2} \cdot \sqrt{\sum (y_2 - \bar{y})^2}} \quad \rightarrow \text{ind. d.} \\
 &\quad n-1 \text{ auf } n
 \end{aligned}$$

$$\text{Pearson}(\text{rank } y_1, \text{rank } y_2) = \frac{\sum (\text{rank } y_1 - \bar{y}) \cdot (\text{rank } y_2 - \bar{y})}{\sqrt{\sum (\text{rank } y_1 - \bar{y})^2} \cdot \sqrt{\sum (\text{rank } y_2 - \bar{y})^2}}$$

$$\sum (\text{rank } y_1 - \bar{y}) \cdot (\text{rank } y_2 - \bar{y}) = 10,5$$

$$\sqrt{\sum (\text{rank } y_1 - \bar{y})^2} \approx 11,9583$$

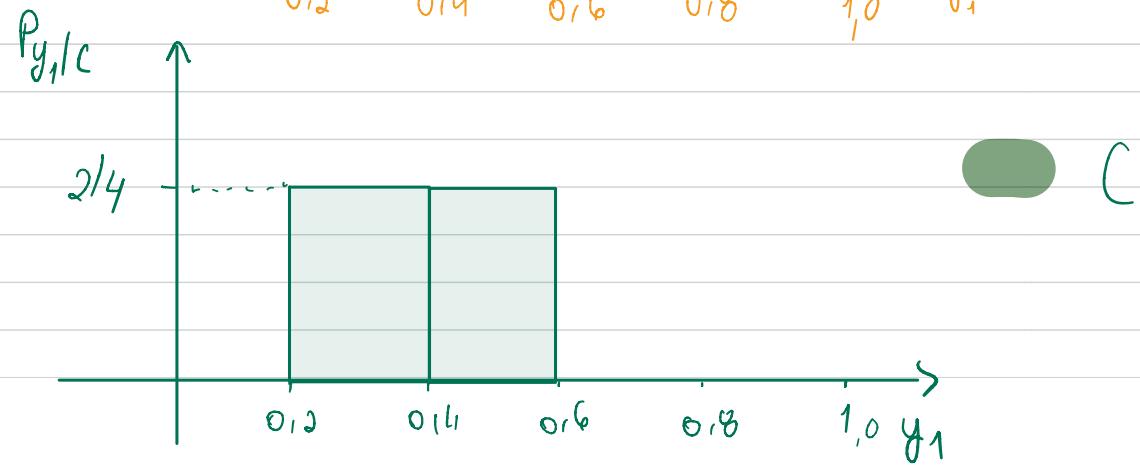
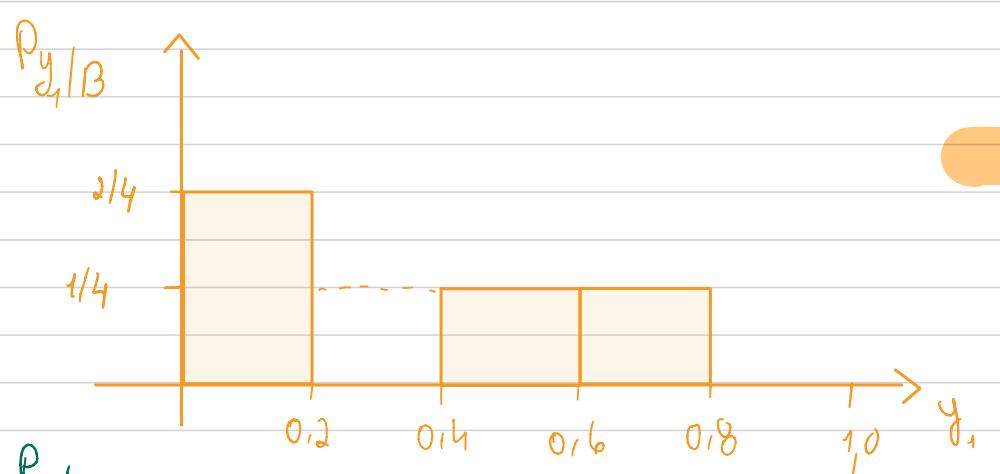
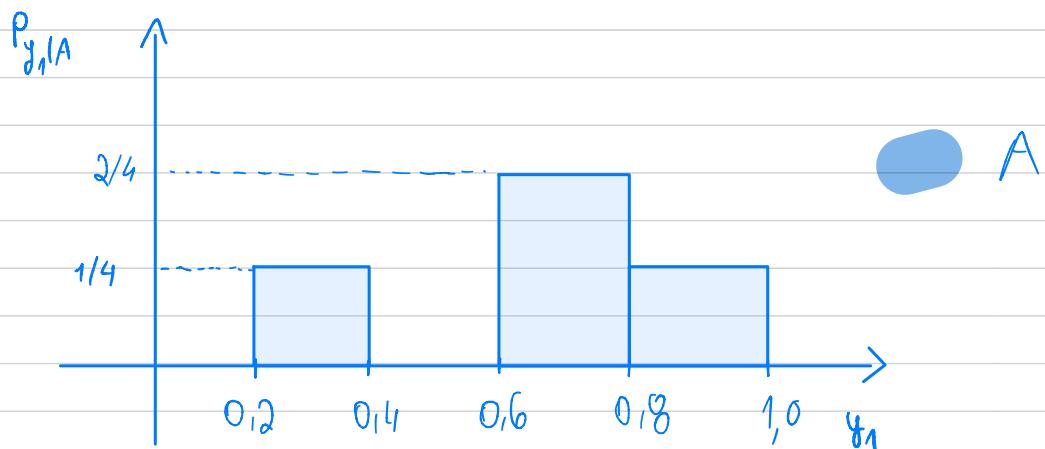
$$\sqrt{\sum (\text{rank } y_2 - \bar{y})^2} \approx 11,0227$$

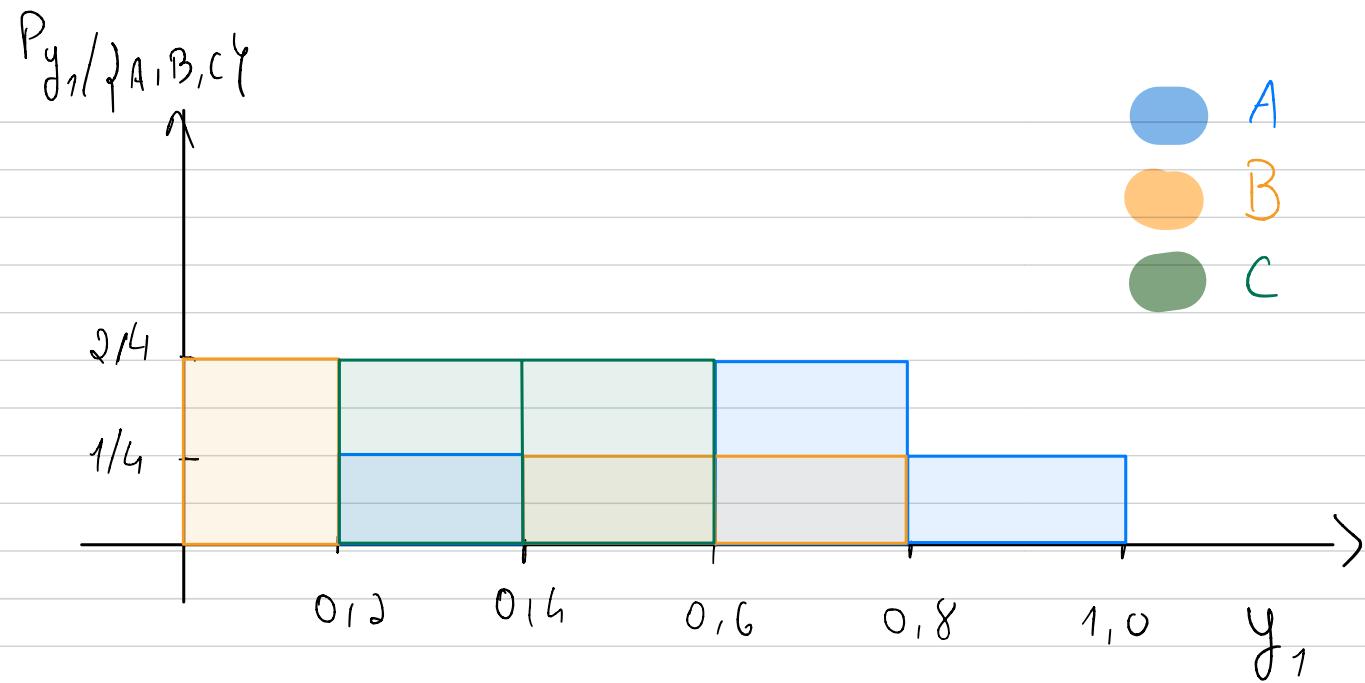
$$\text{Pearson}(\text{rank } y_1, \text{rank } y_2) \approx \frac{10,5}{11,9583 \cdot 11,0227} \approx$$

$$\approx 6,079659$$

Portanto, Coeficiente de Spearman ( $y_1, y_2$ )  $\approx$   
 $\approx 0,079659$

5)





Encontrámos 2 root splits:  $y_1 = 0,2$  e  $y_1 = 0,6$

Decidimos que estas duas são as root-splits uma vez que para  $y_1 < 0,2$  a classe B tem maior probabilidade, para  $0,2 \leq y_1 < 0,6$  a classe C tem maior probabilidade e para  $y_1 \geq 0,6$  a classe A tem maior probabilidade. Assim ficariamos com uma árvore de decisão deste género

