



# Aprendizagem - Homework 1

Pedro Curvo (ist1102716)

Salvador Torpes (ist1102474)

1º Semestre - 23/24

---

## Dataset

Consideramos o seguinte dataset D:

D	$y_1$	$y_2$	$y_3$	$y_4$	$y_{out}$
$x_1$	0.24	1	1	0	A
$x_2$	0.06	2	0	0	B
$x_3$	0.04	0	0	0	B
$x_4$	0.36	0	2	1	C
$x_5$	0.32	0	0	2	C
$x_6$	0.68	2	2	1	A
$x_7$	0.90	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

Tabela 1: Dataset D

# 1 Exercício 1.

De modo a corretamente completar a árvore de decisão, é necessário calcular o Information gain (IG) da variável de output  $y_{out}$  condicionada a cada uma das variáveis  $y_2$ ,  $y_3$  e  $y_4$ .

## 1.1 Escolha do 2º nó

Como queremos completar o ramo  $y_1 > 0.4$ , vamos apenas considerar as ocorrências em que  $y_1 > 0.4$  para calcular o IG.

**Information Gain de  $y_{out}$  condicionada a  $y_2$**

$$IG(y_{out}|y_2) = H(y_{out}) - H(y_{out}|y_2)$$

$$H(y_{out}) = \left( - \sum_{i=1}^3 p_{out_i} (\log_2 p_{out_i}) \right) = - \left( \frac{3}{7} \log_2 \left( \frac{3}{7} \right) + \frac{2}{7} \log_2 \left( \frac{2}{7} \right) + \frac{2}{7} \log_2 \left( \frac{2}{7} \right) \right) = 1.5567$$

$$H(y_{out}|y_2) = \sum_{i=0}^2 p_{y_2=i} H(y_{out}|y_2 = i)$$

Tabela dividida em 3 sub-tabelas, cada uma com os dados que verificam  $y_2 = 0$ ,  $y_2 = 1$  e  $y_2 = 2$ , respetivamente:

D	$y_2$	$y_{out}$
$x_7$	0	A
$x_{10}$	0	B
$x_{12}$	0	C

D	$y_2$	$y_{out}$
$x_9$	1	B
$x_{11}$	1	C

D	$y_2$	$y_{out}$
$x_6$	2	A
$x_8$	2	A

Tabela 2: Dataset D com  $y_2 = 0$    Tabela 3: Dataset D com  $y_2 = 1$    Tabela 4: Dataset D com  $y_2 = 2$

$$H(y_{out}|y_2 = 0) = - \left( \frac{1}{3} \log_2 \left( \frac{1}{3} \right) + \frac{1}{3} \log_2 \left( \frac{1}{3} \right) + \frac{1}{3} \log_2 \left( \frac{1}{3} \right) \right) = 1.58496$$

$$H(y_{out}|y_2 = 1) = - \left( \frac{1}{2} \log_2 \left( \frac{1}{2} \right) + \frac{1}{2} \log_2 \left( \frac{1}{2} \right) \right) = 1$$

$$H(y_{out}|y_2 = 2) = - (\log(1)) = 0$$

Assim, podemos calcular a entropia de  $y_{out}$  condicionada a  $y_2$ :

$$\begin{aligned} H(y_{out}|y_2) &= \frac{3}{7} H(y_{out}|y_2 = 0) + \frac{2}{7} H(y_{out}|y_2 = 1) + \frac{2}{7} H(y_{out}|y_2 = 2) = \\ &= \frac{3}{7} \times 1.58496 + \frac{2}{7} \times 1 + \frac{2}{7} \times 0 = 0.96498 \end{aligned}$$

Por fim, podemos calcular o Information Gain:

$$IG(y_{out}|y_2) = H(y_{out}) - H(y_{out}|y_2) = 1.5567 - 0.96498 = 0.59172$$

### Information Gain de $y_{out}$ condicionada a $y_3$

$$IG(y_{out}|y_3) = H(y_{out}) - H(y_{out}|y_3)$$

$$H(y_{out}|y_3) = \sum_{i=0}^2 p_{y_3=i} H(y_{out}|y_3 = i)$$

Tabela dividida em 3 sub-tabelas, cada uma com os dados que verificam  $y_3 = 0$ ,  $y_3 = 1$  e  $y_3 = 2$ , respectivamente:

D	$y_3$	$y_{out}$
$x_{10}$	0	B

D	$y_3$	$y_{out}$
$x_7$	1	A
$x_9$	1	B

D	$y_3$	$y_{out}$
$x_6$	2	A
$x_8$	2	A
$x_{11}$	2	C
$x_{12}$	2	C

Tabela 5: Dataset D com  $y_3 = 0$    Tabela 6: Dataset D com  $y_3 = 1$    Tabela 7: Dataset D com  $y_3 = 2$

$$H(y_{out}|y_3 = 0) = -(\log(1)) = 0$$

$$H(y_{out}|y_3 = 1) = -\left(\frac{1}{2} \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \log_2\left(\frac{1}{2}\right)\right) = 1$$

$$H(y_{out}|y_3 = 2) = -\left(\frac{2}{4} \log_2\left(\frac{2}{4}\right) + \frac{2}{4} \log_2\left(\frac{2}{4}\right)\right) = 1$$

Assim, podemos calcular a entropia de  $y_{out}$  condicionada a  $y_3$ :

$$\begin{aligned} H(y_{out}|y_3) &= \frac{1}{7} H(y_{out}|y_3 = 0) + \frac{2}{7} H(y_{out}|y_3 = 1) + \frac{4}{7} H(y_{out}|y_3 = 2) = \\ &= \frac{1}{7} \times 0 + \frac{2}{7} \times 1 + \frac{4}{7} \times 1 = 0.85714 \end{aligned}$$

Por fim, podemos calcular o Information Gain:

$$IG(y_{out}|y_3) = H(y_{out}) - H(y_{out}|y_3) = 1.5567 - 0.85714 = 0.69956$$

### Information Gain de $y_{out}$ condicionada a $y_4$

$$IG(y_{out}|y_4) = H(y_{out}) - H(y_{out}|y_4)$$

$$H(y_{out}|y_4) = \sum_{i=0}^2 p_{y_4=i} H(y_{out}|y_4 = i)$$

Tabela dividida em 3 sub-tabelas, cada uma com os dados que verificam  $y_4 = 0$ ,  $y_4 = 1$  e  $y_4 = 2$ , respectivamente:

D	$y_4$	$y_{out}$
$x_8$	0	A
$x_{12}$	0	C

D	$y_4$	$y_{out}$
$x_6$	1	A
$x_9$	1	B
$x_{10}$	1	B

D	$y_4$	$y_{out}$
$x_7$	2	A
$x_{11}$	2	C

Tabela 8: Dataset D com  $y_4 = 0$

Tabela 9: Dataset D com  $y_4 = 1$

Tabela 10: Dataset D com  $y_4 = 2$

$$H(y_{out}|y_4 = 0) = - \left( \frac{1}{2} \log_2 \left( \frac{1}{2} \right) + \frac{1}{2} \log_2 \left( \frac{1}{2} \right) \right) = 1$$

$$H(y_{out}|y_4 = 1) = - \left( \frac{2}{3} \log_2 \left( \frac{2}{3} \right) + \frac{1}{3} \log_2 \left( \frac{1}{3} \right) \right) = 0.918295$$

$$H(y_{out}|y_4 = 2) = - \left( \frac{1}{2} \log_2 \left( \frac{1}{2} \right) + \frac{1}{2} \log_2 \left( \frac{1}{2} \right) \right) = 1$$

Assim, podemos calcular a entropia de  $y_{out}$  condicionada a  $y_4$ :

$$\begin{aligned} H(y_{out}|y_4) &= \frac{2}{7} H(y_{out}|y_4 = 0) + \frac{3}{7} H(y_{out}|y_4 = 1) + \frac{2}{7} H(y_{out}|y_4 = 2) = \\ &= \frac{2}{7} \times 1 + \frac{3}{7} \times 0.918295 + \frac{2}{7} \times 1 = 0.96498 \end{aligned}$$

Por fim, podemos calcular o Information Gain:

$$IG(y_{out}|y_4) = H(y_{out}) - H(y_{out}|y_4) = 1.5567 - 0.96498 = 0.59172$$

**Comparação dos IG** Podemos confirmar, pelos cálculos acima, que:

$$IG(y_{out}|y_2) = IG(y_{out}|y_4) < IG(y_{out}|y_3)$$

Assim, a variável  $y_3$  é a que tem maior IG, pelo que é a variável que escolhemos para o 2º nó da árvore de decisão no ramo  $y_1 > 0.4$ . Este nó vai ter três ramos, um para cada valor possível de  $y_3$ : a ocorrência  $y_3 = 0$  tem apenas uma ocorrência e a ocorrência  $y_3 = 1$  tem apenas duas ocorrências, pelo que estes dois nós não são expandidos. Por outro lado,  $y_3 = 2$  tem 4 ocorrências, pelo que é o único nó que é expandido. Falta averiguar qual a variável que vai ser usada para expandir este nó.

## 1.2 Escolha do 3º nó

Queremos agora completar o ramo que verifica  $y_1 > 0.4$  e  $y_3 = 2$ . Para isso, vamos calcular o IG de  $y_{out}$  para  $y_2$  e  $y_4$  considerando apenas as ocorrências que verificam  $y_1 > 0.4$  e  $y_3 = 2$ :

**Information Gain de  $y_{out}$  condicionada a  $y_2$**

$$IG(y_{out}|y_2) = H(y_{out}) - H(y_{out}|y_2)$$

$$H(y_{out}) = \left( - \sum_{i=1}^3 p_{out_i} (\log_2 p_{out_i}) \right) = - \left( \frac{2}{4} \log_2 \left( \frac{2}{4} \right) + \frac{2}{4} \log_2 \left( \frac{2}{4} \right) \right) = 1$$

$$H(y_{out}|y_2) = \sum_{i=0}^2 p_{y_2=i} H(y_{out}|y_2 = i)$$

As entropias condicionadas de  $y_{out}$  para cada valor de  $y_2$  são:

$$H(y_{out}|y_2 = 0) = -(\log_2(1)) = 0$$

$$H(y_{out}|y_2 = 1) = -(\log_2(1)) = 0$$

$$H(y_{out}|y_2 = 2) = -(\log_2(1)) = 0$$

Assim, podemos calcular a entropia de  $y_{out}$  condicionada a  $y_2$ :

$$\begin{aligned} H(y_{out}|y_2) &= \frac{1}{4}H(y_{out}|y_2 = 0) + \frac{1}{4}H(y_{out}|y_2 = 1) + \frac{2}{4}H(y_{out}|y_2 = 2) = \\ &= \frac{1}{4} \times 0 + \frac{1}{4} \times 0 + \frac{2}{4} \times 0 = 0 \end{aligned}$$

Por fim, podemos calcular o Information Gain:

$$IG(y_{out}|y_2) = H(y_{out}) - H(y_{out}|y_2) = 1 - 0 = 1$$

**Information Gain de  $y_{out}$  condicionada a  $y_4$**

$$IG(y_{out}|y_4) = H(y_{out}) - H(y_{out}|y_4)$$

$$H(y_{out}|y_4) = \sum_{i=0}^2 p_{y_4=i} H(y_{out}|y_4 = i)$$

As entropias condicionadas de  $y_{out}$  para cada valor de  $y_4$  são:

$$H(y_{out}|y_4 = 0) = -\left(\frac{1}{2} \log_2\left(\frac{1}{2}\right) + \frac{1}{2} \log_2\left(\frac{1}{2}\right)\right) = 1$$

$$H(y_{out}|y_4 = 1) = -(\log_2(1)) = 0$$

$$H(y_{out}|y_4 = 2) = -(\log_2(1)) = 0$$

Assim, podemos calcular a entropia de  $y_{out}$  condicionada a  $y_4$ :

$$\begin{aligned} H(y_{out}|y_4) &= \frac{2}{4}H(y_{out}|y_4 = 0) + \frac{1}{4}H(y_{out}|y_4 = 1) + \frac{1}{4}H(y_{out}|y_4 = 2) = \\ &= \frac{2}{4} \times 1 + \frac{1}{4} \times 0 + \frac{1}{4} \times 0 = 0.5 \end{aligned}$$

Por fim, podemos calcular o Information Gain:

$$IG(y_{out}|y_4) = H(y_{out}) - H(y_{out}|y_4) = 1 - 0.5 = 0.5$$

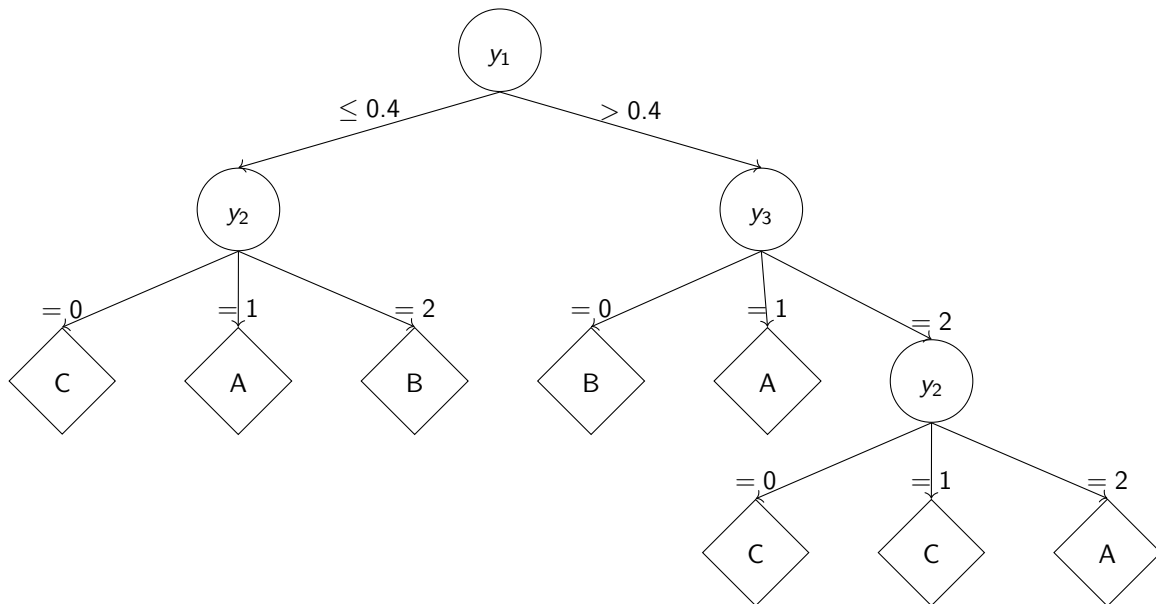
**Comparação dos IG** Podemos confirmar, pelos cálculos acima, que:

$$IG(y_{out}|y_2) > IG(y_{out}|y_4)$$

Assim, a variável  $y_2$  é a que tem maior IG, pelo que é a variável que escolhemos para o 3º nó da árvore de decisão no ramo  $y_1 > 0.4$  e  $y_3 = 2$ . Todos os nós desta árvore têm menos que 4 ocorrências, pelo que nenhum deles é expandido e termina a árvore de decisão.

### 1.3 Construção da árvore de decisão

Para completar a árvore, resta preencher os nós terminais com os valores de  $y_{out}$  que são mais prováveis em cada ramo. Em caso de empate, escolhemos por ordem alfabética. A árvore de decisão final é:



## 2 Exercício 2.

Com o objetivo de desenhar a matriz de confusão da árvore de decisão construída acima, começamos por calcular os valores previstos para o output,  $\hat{y}_{out}$ , para cada uma das ocorrências do dataset D:

D	$y_1$	$y_2$	$y_3$	$y_4$	$\hat{y}_{out}$	$y_{out}$
$x_1$	0.24	1	1	0	A	A
$x_2$	0.06	2	0	0	B	B
$x_3$	0.04	0	0	0	C	B
$x_4$	0.36	0	2	1	C	C
$x_5$	0.32	0	0	2	C	C
$x_6$	0.68	2	2	1	A	A
$x_7$	0.90	0	1	2	A	A
$x_8$	0.76	2	2	0	A	A
$x_9$	0.46	1	1	1	A	B
$x_{10}$	0.62	0	0	1	B	B
$x_{11}$	0.44	1	2	2	C	C
$x_{12}$	0.52	0	2	0	C	C

Tabela 11: Dataset D com  $\hat{y}_{out}$

Assim, desenhamos a **matriz de confusão**:

	Valores reais			
		A	B	C
	A	4	1	0
	B	0	2	0
	C	0	1	4

Tabela 12: Matriz de confusão

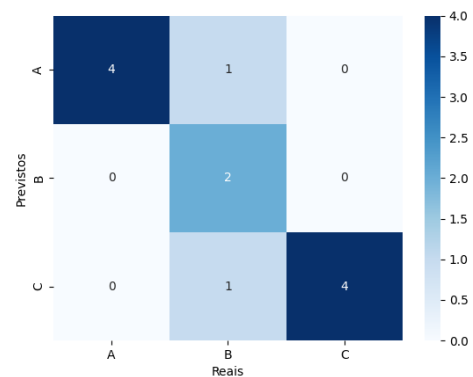


Figura 1: Matriz de confusão

### 3 Exercício 3.

Para calcular o  $F_1$ -score para cada uma das classes de  $y_{out}$ , começamos por calcular a precisão e o recall para cada uma delas:

A precisão é dada por:

$$P = \frac{TP}{TP + FP}$$

O recall é dado por:

$$R = \frac{TP}{TP + FN}$$

Assim, obtemos que:

$$P_A = \frac{4}{4 + 1 + 0} = \frac{4}{5}$$

$$P_B = \frac{2}{2 + 0 + 0} = 1$$

$$P_C = \frac{4}{4 + 1 + 0} = \frac{4}{5}$$

$$R_A = \frac{4}{4 + 0 + 0} = 1$$

$$R_B = \frac{2}{2 + 1 + 1} = \frac{1}{2}$$

$$R_C = \frac{4}{4 + 0 + 0} = 1$$

Por fim, o  $F_1$ -score é dado por:

$$F_1 = \frac{1}{0.5 \cdot \frac{1}{P} + 0.5 \cdot \frac{1}{R}}$$

Assim, podemos calcular o  $F_1$ -score para cada uma das classes:

$$F_1(A) = \frac{1}{0.5 \cdot \frac{5}{4} + 0.5 \cdot 1} = \frac{1}{\frac{5}{8} + \frac{1}{2}} = \frac{1}{\frac{5}{8} + \frac{4}{8}} = \frac{1}{\frac{9}{8}} = \frac{8}{9}$$

$$F_1(B) = \frac{1}{0.5 \cdot 1 + 0.5 \cdot 2} = \frac{1}{0.5 + 1} = \frac{1}{1.5} = \frac{2}{3}$$

$$F_1(C) = \frac{1}{0.5 \cdot \frac{5}{4} + 0.5 \cdot 1} = \frac{1}{\frac{5}{8} + \frac{1}{2}} = \frac{1}{\frac{5}{8} + \frac{4}{8}} = \frac{1}{\frac{9}{8}} = \frac{8}{9}$$

**Resposta:** Assim, podemos concluir que a classe com menor  $F_1$ -score é a classe B, com um  $F_1$ -score de  $\frac{2}{3}$ .



## 4 Exercício 4.

Para calcular o coeficiente de Spearman entre as variáveis  $y_1$  e  $y_2$ , começamos por calcular o rank de cada uma das variáveis:

D	$y_1$	$y_1$ rank	$y_2$	$y_2$ rank
$x_1$	0.24	3	1	8
$x_2$	0.06	2	2	11
$x_3$	0.04	1	0	3.5
$x_4$	0.36	5	0	3.5
$x_5$	0.32	4	0	3.5
$x_6$	0.68	10	2	11
$x_7$	0.90	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
$x_{12}$	0.52	8	0	3.5

Tabela 13: Dataset D com ranks

A fórmula para o coeficiente de Spearman é:

$$r_s = \frac{\text{cov}(\text{rank}(y_1), \text{rank}(y_2))}{\sqrt{\text{var}(\text{rank}(y_1)) \cdot \text{var}(\text{rank}(y_2))}} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n (\text{rank}(x_i) - \text{rank}(\bar{x}))^2\right) \cdot \left(\frac{1}{n} \sum_{i=1}^n (\text{rank}(y_i) - \text{rank}(\bar{y}))^2\right)}} = 0.079659$$

**Resposta:** Como o coeficiente de Spearman entre as duas variáveis é  $\ll 1$ , podemos concluir que as duas variáveis não estão correlacionadas.

## 5 Exercício 5.

Queremos desenhar os histogramas da variável  $y_1$  condicionados aos diferentes outcomes da variável  $y_{out}$ . Assim, é necessário calcular os valores dos bins para 3 histogramas diferentes. Em primeiro lugar, utilizamos os dados:

Queremos que cada histograma tenha 5 bins sendo a range total  $[0, 1]$ , logo, os bins possíveis são  $[0, 0.2]$ ,  $[0.2, 0.4]$ ,  $[0.4, 0.6]$ ,  $[0.6, 0.8]$  e  $[0.8, 1]$ .

Na seguinte tabela encontram-se os cálculos para a altura de cada bin em cada um dos três histogramas:

$y_{out}$	$y_1$
A	0.24
A	0.68
A	0.90
A	0.76
B	0.06
B	0.04
B	0.46
B	0.62
C	0.36
C	0.32
C	0.44
C	0.52

Tabela 14: Dataset D com  $y_1$  e  $y_{out}$

Bins	Contagens com $y_{out} = A$	Altura do bin para $y_{out} = A$	Contagens com $y_{out} = B$	Altura do bin para $y_{out} = B$	Contagens com $y_{out} = C$	Altura do bin para $y_{out} = C$	Classe predominante no bin
[0, 0.2]	0	0	2	$\frac{2}{4} = 0.5$	0	0	B
[0.2, 0.4]	1	$\frac{1}{4} = 0.25$	0	0	2	$\frac{2}{4} = 0.5$	C
[0.4, 0.6]	0	0	1	$\frac{1}{4} = 0.25$	2	$\frac{2}{4} = 0.5$	C
[0.6, 0.8]	2	$\frac{2}{4} = 0.5$	1	$\frac{1}{4} = 0.25$	0	0	A
[0.8, 1]	1	$\frac{1}{4} = 0.25$	0	0	0	0	A