Aprendizagem - HomeWork 1

Pedro Curvo (ist1102716)

Salvador Torpes (ist1102474)

1º Semestre - 23/24

---

1 Dataset

Considering 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

2 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:

2.1 Information Gain de y_{out} condicionada a y_2

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

$$H(y_{out}) = \left(-\sum_{i=1}^{3} p_{out_i}(\log_2 p_{out_i})\right) = -\left(\frac{4}{12}\log_2\left(\frac{4}{12}\right) + \frac{4}{12}\log_2\left(\frac{4}{12}\right) + \frac{4}{12}\log_2\left(\frac{4}{12}\right)\right) = 1.58496$$

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

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_3	0	B
x_4	0	C
x_5	0	C
x_7	0	A
x_{10}	0	B
x_{12}	0	C

Tabela 2: Dataset D com $y_2 = 0$

D	y_2	y_{out}
x_1	1	A
x_9	1	B
x_{11}	1	C

Tabela 3: Dataset D com $y_2 = 1$

D	y_2	y_{out}
x_1	1	A
x_9	1	B
x_{11}	1	C

Tabela 4: Dataset D com $y_2 = 1$

$$H(y_{out}|y_2 = 0) = -\left(\frac{1}{6}\log_2\left(\frac{1}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1.45915$$

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

$$H(y_{out}|y_2 = 2) = -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) = 0.9183$$

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

$$H(y_{out}|y_2) = \frac{6}{12}H(y_{out}|y_2 = 0) + \frac{3}{12}H(y_{out}|y_2 = 1) + \frac{3}{12}H(y_{out}|y_2 = 2) =$$
$$= \frac{6}{12} \times 1.45915 + \frac{3}{12} \times 1.58496 + \frac{3}{12} \times 0.9183 = 1.35538$$

Por fim, podemos calcular o Information Gain:

$$IG(y_{out}|y_2) = H(y_{out}) - H(y_{out}|y_2) = 1.58496 - 1.35538 = 0.22958$$

2.2 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 $y_3 = 2$, respetivamente:

D	y_3	y_{out}
x_2	0	B
x_3	0	B
x_5	0	C
x_{10}	0	B

Tabela 5: Dataset D com $y_3 = 0$

D	y_3	y_{out}
x_1	1	A
x_7	1	A
x_9	1	B

Tabela 6: Dataset D com $y_3 = 1$

D	y_3	y_{out}
x_4	2	C
x_6	2	A
x_8	2	A
x_{11}	2	C
x_{12}	2	C

Tabela 7: Dataset D com $y_3 = 2$

$$H(y_{out}|y_3=0) = -\left(\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) = 0.81128$$

$$H(y_{out}|y_3=1) = -\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) = 0.9183$$

$$H(y_{out}|y_3=2) = -\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) = 0.97095$$

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

$$H(y_{out}|y_3) = \frac{4}{12}H(y_{out}|y_3=0) + \frac{3}{12}H(y_{out}|y_3=1) + \frac{5}{12}H(y_{out}|y_3=2) =$$
$$= \frac{4}{12} \times 0.81128 + \frac{3}{12} \times 0.9183 + \frac{5}{12} \times 0.97095 = 0.90456$$

Por fim, podemos calcular o Information Gain:

$$IG(y_{out}|y_3) = H(y_{out}) - H(y_{out}|y_3) = 1.58496 - 0.90456 = 0.6804$$

2.3 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 $y_4 = 2$, respetivamente:

D	y_4	y_{out}
x_1	0	A
x_2	0	B
x_3	0	B
x_8	0	A
x_{12}	0	C

Tabela 8: Dataset D com $y_4 = 0$

D	y_4	y_{out}
x_4	1	C
x_6	1	A
x_9	1	B
x_{10}	1	B

Tabela 9: Dataset D com $y_4 = 1$

D	y_4	y_{out}
x_5	2	C
x_7	2	A
x_{11}	2	C

Tabela 10: Dataset D com $y_4 = 2$

$$H(y_{out}|y_4=0) = -\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right) = 1.52193$$

$$H(y_{out}|y_4=1) = -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) = 1.5$$

$$H(y_{out}|y_4=2) = -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) = 0.9183$$

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

$$H(y_{out}|y_4) = \frac{5}{12}H(y_{out}|y_4=0) + \frac{4}{12}H(y_{out}|y_4=1) + \frac{3}{12}H(y_{out}|y_4=2) =$$
$$= \frac{5}{12} \times 1.52193 + \frac{4}{12} \times 1.5 + \frac{3}{12} \times 0.9183 = 1.3637$$

Por fim, podemos calcular o Information Gain:

$$IG(y_{out}|y_4) = H(y_{out}) - H(y_{out}|y_4) = 1.58496 - 1.3637 = 0.22126$$

2.4 Construção da árvore de decisão

Ordenando os IG por ordem decrescente obtemos:

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

Assim, o nó com $y_1 > 0.4$ corresponde a y_3. A variável y_3 tem 3 possíveis valores, pelo que a árvore de decisão terá 3 ramos: como estamos condicionados a $y_1 > 0.4$, temos as seguintes ocorrências em cada ramo:

$$\#(y_3 = 0|y_1 > 0.4) = 1$$
$$\#(y_3 = 1|y_1 > 0.4) = 2$$
$$\#(y_3 = 2|y_1 > 0.4) = 4$$

Assim, apenas o nó $y_3 = 2$ tem pelo menos 4 ocorrências, logo, é o único que é expandido para a variável y_2.

Nenhum dos ramos da variável y_2 tem pelo menos 4 ocorrências, pelo que nenhum deles é expandido para a variável y_4 e termina a árvore de decisão.

A árvore de decisão final é: