Aprendizagem 2023
Homework I — Group 003
(ist1107028, ist1107137)

Part I: Pen and paper

1. **Complete the given decision tree using Shannon entropy ($\log_2$) and considering that: (i) a minimum of 4 observations is required to split an internal node, and (ii) decisions by ascending alphabetic should be placed in case of ties.**

The entropy of y_{out} is given by:

$$\begin{aligned}
H(y_{\text{out}}|y_1 >= 0.3) &= -P(y_{\text{out}} = A|y_1 >= 0.3)\log_2(P(y_{\text{out}} = A|y_1 >= 0.3)) \\
H(y_{\text{out}}|y_1 >= 0.3) &= -P(y_{\text{out}} = B|y_1 >= 0.3)\log_2(P(y_{\text{out}} = B|y_1 >= 0.3)) \\
H(y_{\text{out}}|y_1 >= 0.3) &= -P(y_{\text{out}} = C|y_1 >= 0.3)\log_2(P(y_{\text{out}} = C|y_1 >= 0.3))
\end{aligned}$$

Replacing by the values is obtained:

$$H(y_{\text{out}}|y_1 >= 0.3) = -\left(\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right)\right) \approx 1.5567$$

We must now calculate $H(y_{\text{out}}|y_1 >= 0.3, y_x)$, in wich x will assume the values of 2, 3 and 4:

$$\begin{aligned}
H(y_{\text{out}}|y_1 >= 0.3, y_x) = {} & P(y_x = 0|y_1 >= 0.3)H(y_{\text{out}}|y_1 >= 0.3, y_x = 0) \\
& + P(y_x = 1|y_1 >= 0.3)H(y_{\text{out}}|y_1 >= 0.3, y_x = 1) \\
& + P(y_x = 2|y_1 >= 0.3)H(y_{\text{out}}|y_1 >= 0.3, y_x = 2)
\end{aligned} \tag{1}$$

The information gain of y_x is given by:

$$IG(y_{\text{out}}|y_1 >= 0.3, y_x) = H(y_{\text{out}}|y_1 >= 0.3) - H(y_{\text{out}}|y_1 >= 0.3, y_x) \tag{2}$$

$\boxed{\text{x=2:}}$

$$P(y_2 = 0|y_1 >= 0.3) = \frac{4}{7} \quad P(y_2 = 1|y_1 >= 0.3) = \frac{3}{7} \quad P(y_2 = 2|y_1 >= 0.3) = \frac{0}{7}$$

$$\begin{aligned}
H(y_{\text{out}}|y_1 > 0.3, y_2 = 0) &= -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) = 1.5 \\
H(y_{\text{out}}|y_1 > 0.3, y_2 = 1) &= -\left(\frac{2}{3}\log_2\left(\frac{2}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right)\right) \approx 0.9183
\end{aligned}$$

On account of the fact that $P(y_2 = 2|y_1 >= 0.3) = \frac{0}{7} = 0$, $H(y_{\text{out}}|y_1 > 0.3, y_2 = 2)$ will also be 0.

Thus, by inserting these values into equation (1), we get:

$$H(y_{\text{out}}|y_1 >= 0.3, y_2) = \frac{4}{7} \times 1.5 + \frac{3}{7} \times 0.9183 + \frac{0}{7} \times 0 = 1.2507$$

Lastly, we calculate the information gain as per (2):

$$IG(y_{\text{out}}|y_1 >= 0.3, y_2) = 1.5567 - 1.2507 = 0.306$$

$\boxed{\text{x=3:}}$

$$P(y_3 = 0|y_1 >= 0.3) = \frac{2}{7} \quad P(y_3 = 1|y_1 >= 0.3) = \frac{4}{7} \quad P(y_3 = 2|y_1 >= 0.3) = \frac{1}{7}$$

$$H(y_{\text{out}}|y_1 >= 0.3, y_3 = 0) = -\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 0$$
$$H(y_{\text{out}}|y_1 >= 0.3, y_3 = 1) = -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) = 1.5$$
$$H(y_{\text{out}}|y_1 >= 0.3, y_3 = 2) = -\left(\frac{0}{1}\log_2\left(\frac{0}{1}\right) + \frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) = 0$$

Thus, by inserting these values into equation (1), we get:

$$H(y_{\text{out}}|y_1 >= 0.3, y_3) = \frac{2}{7} \times 0 + \frac{4}{7} \times 1.5 + \frac{1}{7} \times 0 \approx 0.8571$$

Lastly, we calculate the information gain as per (2):

$$IG(y_{\text{out}}|y_1 >= 0.3, y_3) = 1.5567 - 0.8571 = \mathbf{0.6996}$$

$\boxed{\text{x=4:}}$

$$P(y_4 = 0|y_1 >= 0.3) = \frac{4}{7} \quad P(y_4 = 1|y_1 >= 0.3) = \frac{3}{7} \quad P(y_4 = 2|y_1 >= 0.3) = \frac{0}{7}$$

$$H(y_{\text{out}}|y_1 >= 0.3, y_4 = 0) = -\left(\frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{0}{4}\log_2\left(\frac{0}{4}\right)\right) = 1$$
$$H(y_{\text{out}}|y_1 >= 0.3, y_4 = 1) = -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) \approx 0.9183$$

Once again, $P(y_4 = 2|y_1 >= 0.3) = \frac{0}{7} = 0$. Therefore $H(y_{\text{out}}|y_1 > 0.3, y_4 = 2)$ will also equal 0.

Thus, by inserting these values into equation (1), we get:

$$H(y_{\text{out}}|y_1 >= 0.3, y_4) = \frac{4}{7} \times 1 + \frac{3}{7} \times 0.9183 + \frac{0}{7} \times 0 \approx 0.9650$$

Lastly, we calculate the information gain as per (2):

$$IG(y_{\text{out}}|y_1 >= 0.3, y_4) = 1.5567 - 0.9650 = 0.5917$$

After calculating the information gains for each attribute, we can observe that the attribute y_3 has the highest value of 0.6996. Accordingly, we chose it as the next node. Since there are at least four observations with $y_1 >= 0.3$, we split the new node.

If we fix $y_1 >= 0.3$ with $y_3 = 0$, $y_3 = 1$ and $y_3 = 2$, we will obtain 2, 4 and 1 observations, respectively. This gives us three new leaves for the branch in question. The node corresponding to $y_3 = 0$ will be class A, and the one corresponding to $y_3 = 2$ will be from class B, since these classes are the ones that appear most frequently for the respective conditions in the data set. For the $y_3 = 2$ branch with the $y_1 >= 0.3$ condition we get exactly four observations, as we said previously. That being the case, we must split the next node with the higher information gain.

$$\begin{aligned}
H(y_{\text{out}}|y_1 >= 0.3, y_3 = 1, y_2 = 0) &= -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) = 1.5 \\
H(y_{\text{out}}|y_1 >= 0.3, y_3 = 1, y_2 = 1) &= -\left(\frac{0}{0}\log_2\left(\frac{0}{0}\right) + \frac{0}{0}\log_2\left(\frac{0}{0}\right) + \frac{0}{0}\log_2\left(\frac{0}{0}\right)\right) = 0 \\
H(y_{\text{out}}|y_1 >= 0.3, y_3 = 1, y_2 = 2) &= -\left(\frac{0}{0}\log_2\left(\frac{0}{0}\right) + \frac{0}{0}\log_2\left(\frac{0}{0}\right) + \frac{0}{0}\log_2\left(\frac{0}{0}\right)\right) = 0 \\
H(y_{\text{out}}|y_1 >= 0.3, y_3 = 1, y_4 = 0) &= -\left(\frac{0}{1}\log_2\left(\frac{0}{1}\right) + \frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) = 0 \\
H(y_{\text{out}}|y_1 >= 0.3, y_3 = 1, y_4 = 1) &= -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) = 0.9183 \\
H(y_{\text{out}}|y_1 >= 0.3, y_3 = 1, y_4 = 2) &= -\left(\frac{0}{0}\log_2\left(\frac{0}{0}\right) + \frac{0}{0}\log_2\left(\frac{0}{0}\right) + \frac{0}{0}\log_2\left(\frac{0}{0}\right)\right) = 0
\end{aligned}$$

From other calculations above, we know that:

$$H(y_{\text{out}}|y_1 >= 0.3, y_3 = 1) = 1.5$$

Therefore, replacing the above values on equation (1), gives us:

$$\begin{aligned}
E(y_{\text{out}}|y_1 >= 0.3, y_3 = 1, y_2) &= \frac{4}{4} \times 1.5 + \frac{0}{4} \times 0 + \frac{0}{4} \times 0 = 1.5 \\
E(y_{\text{out}}|y_1 >= 0.3, y_3 = 1, y_4) &= \frac{1}{4} \times 0 + \frac{3}{4} \times 0.9183 + \frac{0}{4} \times 0 = 0.6887
\end{aligned}$$

Finally, we can calculate the information gain, as per (2):

$$IG(y_{\text{out}}|y_1 >= 0.3, y_3 = 1, y_2) = 1.5 - 1.5 = 0$$
$$IG(y_{\text{out}}|y_1 >= 0.3, y_3 = 1, y_4) = 1.5 - 0.6887 = \mathbf{0.8113}$$

Since y_4 is the node with the highest information gain of 0.8113, it is the node we chose to split. With the conditions of $y_1 >= 0.3$ and $y_3 = 1$, we can check that the conditions $y_4 = 0$, $y_4 = 1$ and $y_4 = 2$ all have less than 4 observations each. Therefore, we take the class with the highest frequency for each respective conditions on the dataset. Finally, we can construct the following decision tree:

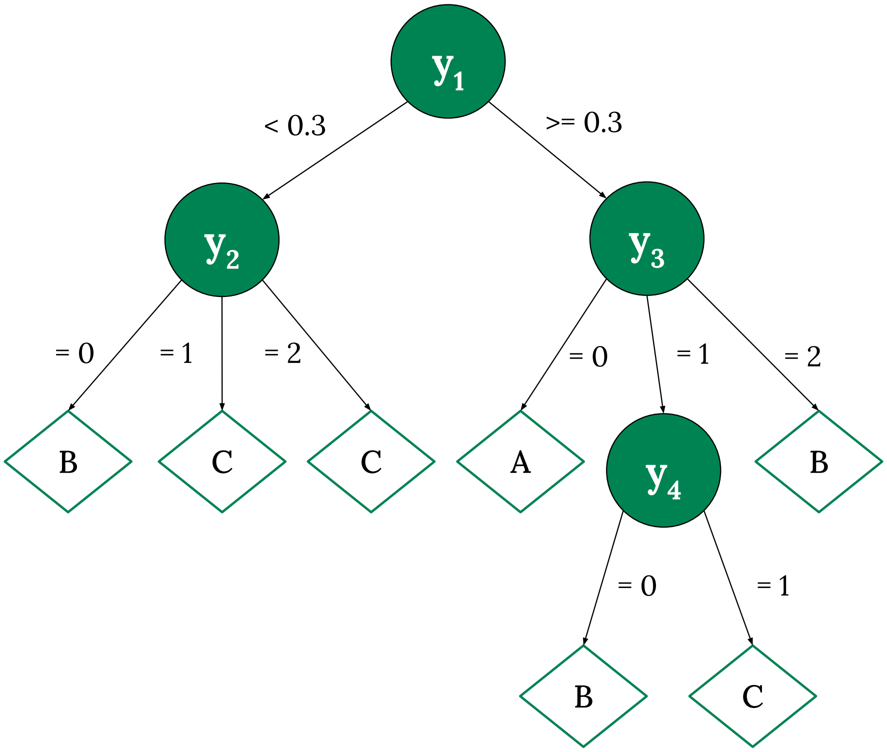


Figure 1: Decision Tree for exercise 1

2. **Draw the training confusion matrix for the learnt decision tree.**

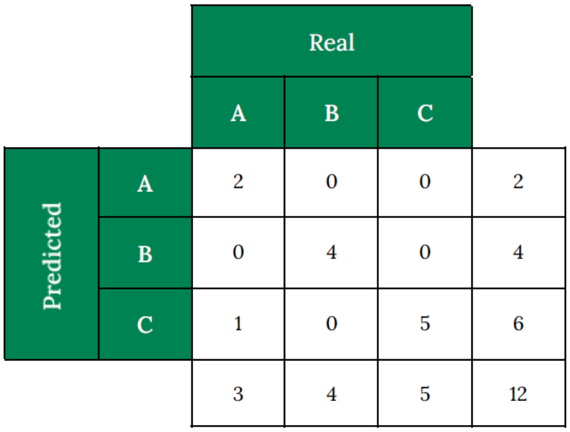
We can then predict the values for each observation through the decision tree drawn in the previous point.

To do this, we start at the initial node of the tree and follow its branches. When we reach the next node, the process is repeated for the succeeding variable and the following ones until a leaf is reached. The class present in this leaf will correspond to the predicted value for the respective D. The real value will be the one in the y_out column corresponding to that same D.

This results in the following values:

$$\begin{array}{rl} & \begin{array}{cccccccccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 & x_{10} & x_{11} & x_{12} \end{array} \\ \text{real} & = \left[\begin{array}{cccccccccccc} \text{C} & \text{B} & \text{C} & \text{B} & \text{C} & \text{B} & \text{A} & \text{A} & \text{C} & \text{C} & \text{A} & \text{B} \end{array}\right] \\ \text{predicted} & = \left[\begin{array}{cccccccccccc} \text{C} & \text{B} & \text{C} & \text{B} & \text{C} & \text{B} & \text{C} & \text{A} & \text{C} & \text{C} & \text{A} & \text{B} \end{array}\right] \end{array}$$

Finally, we set up the confusion matrix by counting the pairs of observations previously mentioned.



		Real			
		A	B	C	
Predicted	A	2	0	0	2
	B	0	4	0	4
	C	1	0	5	6
		3	4	5	12

Figure 2: Confusion matrix for exercise 2

3. **Identify which class has the lowest training F1 score.**

F1_{score} is given by the following equation:

$$\text{F1}_{\text{score}} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

And precision and recall are given by:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True PositivesFalse Positives}} \tag{4}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5}$$

We can calculate the precision by replacing the according values for each class in equation (4):

$$\text{Precision}_A = \frac{2}{2+0} = 1 \qquad \text{Precision}_B = \frac{4}{4+0} = 1 \qquad \text{Precision}_C = \frac{5}{5+1} = \frac{5}{6}$$

Same goes for recall, using equation (5):

$$\text{Recall}_A = \frac{2}{2+1} = \frac{2}{3} \qquad \text{Recall}_B = \frac{4}{4+0} = 1 \qquad \text{Recall}_C = \frac{5}{5+0} = 1$$

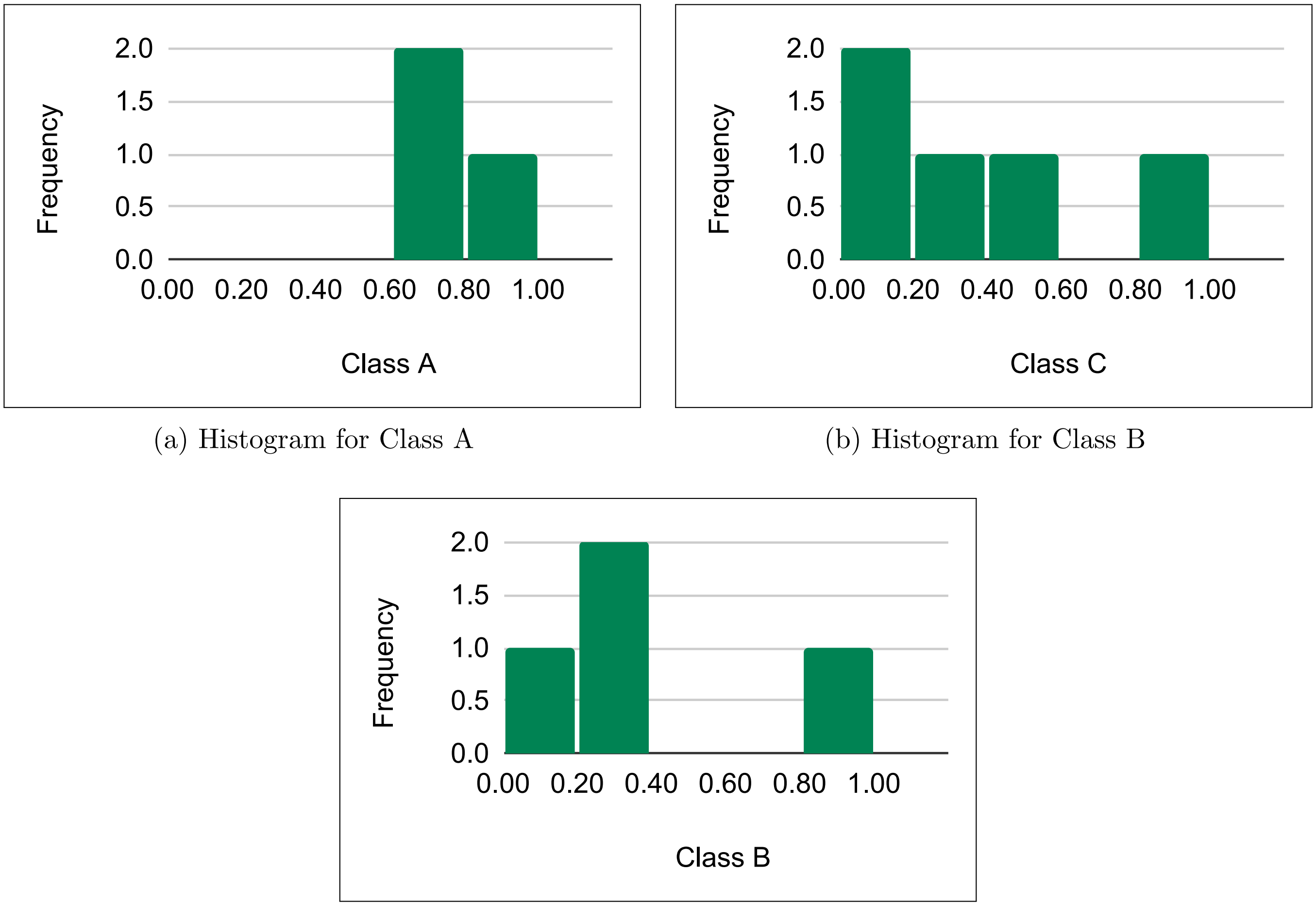
At Last, we determine each F1_{score} with (3):

$$\text{F1}_{\text{score}A} = 2 \times \frac{1 \times \frac{2}{3}}{1 + \frac{2}{3}} = 0.8000 \qquad \text{F1}_{\text{score}B} = 2 \times \frac{1 \times 1}{1+1} = 1 \qquad \text{F1}_{\text{score}C} = 2 \times \frac{\frac{5}{6} \times 1}{\frac{5}{6} + 1} \approx 0.9091$$

The **class with the lowest training F1 score is A**, with a score of 0.8000.

4. **Draw the class-conditional relative histograms of y1 using 5 equally spaced bins in [0,1]. Find the n-ary root split using the discriminant rules from these empirical distributions.**

    1. First step for this problem was creating a histogram for every class present, which are A, B and C. For each, when a value of y_1 falls into a bin, we add one to the value of that bin. Each bin has a size of 0.2, since we are using 5 equal spaced bins in the interval [0,1].



(a) Histogram for Class A

(b) Histogram for Class B

(c) Histogram for Class C

Figure 3: Class-conditional relative histograms of y_1 using 5 equal spaced bins in [0,1]

2. For a quick and broad overview of how to create documents in LaTeX see

Part II: Programming

3. Solution to the programming questions here.

End note: do not forget to also submit your Jupyter notebook