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.

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

$$H(y_{\text{out}}|y_1 > 0.3) = -\left(\frac{3}{9}\log_2\left(\frac{3}{9}\right) + \frac{2}{9}\log_2\left(\frac{2}{9}\right) + \frac{4}{9}\log_2\left(\frac{4}{9}\right)\right) \approx 1,53049$$

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

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

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

$$P(y_2 = 0|y_1 > 0.3) = \frac{3}{9} \quad P(y_2 = 1|y_1 > 0.3) = \frac{2}{9} \quad P(y_2 = 2|y_1 > 0.3) = \frac{4}{9}$$

$$H(y_{\text{out}}|y_1 > 0.3, y_2 = 0) = -\left(\frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{1}{5}\log_2\left(\frac{1}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) \approx 1.37095$$
$$H(y_{\text{out}}|y_1 > 0.3, y_2 = 1) = -\left(\frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$
$$H(y_{\text{out}}|y_1 > 0.3, y_2 = 2) = -\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 0$$

$$H(y_{\text{out}}|y_1 > 0.3, y_2) = \frac{5}{9} \times 1.37095 + \frac{2}{9} \times 1 + \frac{2}{9} \times 0 \approx 0.98386$$

$$IG(y_{\text{out}}|y_1 > 0.3, y_2) = 1.53049 - 0.98386 = \mathbf{0.54663}$$

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

$$P(y_3 = 0|y_1 > 0.3) = \frac{2}{9} \quad P(y_3 = 1|y_1 > 0.3) = \frac{2}{9} \quad P(y_3 = 2|y_1 > 0.3) = \frac{5}{9}$$

$$E(y_{\text{out}}|y_1 > 0.3, y_3 = 0) = -\left(\frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$
$$E(y_{\text{out}}|y_1 > 0.3, y_3 = 1) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 1$$
$$E(y_{\text{out}}|y_1 > 0.3, y_3 = 2) = -\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{0}{5}\log_2\left(\frac{0}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) \approx 0.97095$$

$$E(y_{\text{out}}|y_1 > 0.3, y_3) = \frac{2}{9} \times 1 + \frac{2}{9} \times 1 + \frac{5}{9} \times 0.97095 \approx 0.98386$$

$$IG(y_{\text{out}}|y_1 > 0.3, y_3) = 1.53049 - 0.98386 = \mathbf{0.54663}$$

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

$$P(y_4 = 0|y_1 > 0.3) = \frac{2}{9} \quad P(y_4 = 1|y_1 > 0.3) = \frac{4}{9} \quad P(y_4 = 2|y_1 > 0.3) = \frac{3}{9}$$

$$E(y_{\text{out}}|y_1 > 0.3, y_4 = 0) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$
$$E(y_{\text{out}}|y_1 > 0.3, y_4 = 1) = -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{2}{4}\log_2\left(\frac{2}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) = 1.5$$
$$E(y_{\text{out}}|y_1 > 0.3, y_4 = 2) = -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) = 0.918295$$

$$E(y_{\text{out}}|y_1 > 0.3, y_4) = \frac{2}{9} \times 1 + \frac{4}{9} \times 1.5 + \frac{3}{9} \times 0.918296 \approx 1.19499$$

$$IG(y_{\text{out}}|y_1 > 0.3, y_4) = 1.53049 - 1.19499 = 0.3355$$

After calculating the information gains for each attribute, we can observe that both attributes y_2 and y_3 have the highest value of 0.54663.
Since we are faced with a tie, we choose y_2 as the next node, following the ascending alphabetical order (mentioned in point (ii) of the question summary). Considering that are at least four observations with $y_1 > 0.3$, we split the new node.

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

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

3. For more details on putting math into LaTeX documents you can see

4. We you get to the next problem, you can end the enumerate for the parts of the previous problem and then add another item.

   1. Use a nested enumerate environment to label the parts of the next problem.

   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