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.

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

$$H(y_{\text{out}}|y_1 > 0.3) = -\left(\frac{3}{6}\log_2\left(\frac{3}{6}\right) + \frac{1}{6}\log_2\left(\frac{1}{6}\right) + \frac{2}{6}\log_2\left(\frac{2}{6}\right)\right) \approx 1.4591$$

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

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

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

$$P(y_2 = 0|y_1 > 0.3) = \frac{3}{6} \quad P(y_2 = 1|y_1 > 0.3) = \frac{3}{6} \quad P(y_2 = 2|y_1 > 0.3) = \frac{0}{6}$$

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

ou então colocar "Como não existem B's nas condições avaliadas, a entropia é zero"

$$H(y_{\text{out}}|y_1 > 0.3, y_2) = \frac{3}{6} \times 0.91830 + \frac{3}{6} \times 0.9183 + \frac{0}{6} \times 0 \approx 0.9183$$

$$IG(y_{\text{out}}|y_1 > 0.3, y_2) = 1.4591 - 0.9183 = 0.5408$$

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

$$P(y_3 = 0|y_1 > 0.3) = \frac{2}{6} \quad P(y_3 = 1|y_1 > 0.3) = \frac{3}{6} \quad P(y_3 = 2|y_1 > 0.3) = \frac{1}{6}$$

$$E(y_{\text{out}}|y_1 > 0.3, y_3 = 0) = -\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right) + \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) = 0$$
$$E(y_{\text{out}}|y_1 > 0.3, y_3 = 1) = -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{0}{3}\log_2\left(\frac{0}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) \approx 0.91830$$
$$E(y_{\text{out}}|y_1 > 0.3, y_3 = 2) = -\left(\frac{0}{1}\log_2\left(\frac{0}{1}\right) + \frac{1}{1}\log_2\left(\frac{1}{1}\right) + \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) = 0$$

$$E(y_{\text{out}}|y_1 > 0.3, y_3) = \frac{2}{6} \times 0 + \frac{3}{6} \times 0.9183 + \frac{1}{6} \times 0 \approx 0.4592$$

$$IG(y_{\text{out}}|y_1 > 0.3, y_3) = 1.4591 - 0.4592 = \mathbf{0.9999}$$

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

$$P(y_4 = 0|y_1 > 0.3) = \frac{3}{6} \quad P(y_4 = 1|y_1 > 0.3) = \frac{3}{6} \quad P(y_4 = 2|y_1 > 0.3) = \frac{0}{6}$$

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

mesma situação de $x = 2$

$$E(y_{\text{out}}|y_1 > 0.3, y_4) = \frac{3}{6} \times 0.9183 + \frac{3}{6} \times 0.9183 + \frac{0}{6} \times 0 \approx 0.9183$$

$$IG(y_{\text{out}}|y_1 > 0.3, y_4) = 1.4591 - 0.9183 = 0.5408$$

After calculating the information gains for each attribute, we can observe that the attribute y_3 has the highest value of 0.9999. 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, 3 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, the one with $y_3 = 1$ will be class C, and the remaining, $y_3 = 2$ will be from class B, since these classes are the ones that appear most frequently for the respective conditions in the data set.

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

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

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