Aprendizagem 2024
Homework I – Group 016
(ist1106022, ist1106720)

## **Part I**: Pen and paper

**1.)**

- Data Subset for $y_1 \geq 0.3$

  - Class A $\longrightarrow$ 3 $(x_7, x_8, x_{11})$
  - Class B $\longrightarrow$ 2 $(x_6, x_{12})$
  - Class C $\longrightarrow$ 2 $(x_9, x_{10})$

| $D$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_{out}$ |
|---|---|---|---|---|---|
| $x_6$ | 0.30 | 0 | 1 | 0 | $B$ |
| $x_7$ | 0.76 | 0 | 1 | 1 | $A$ |
| $x_8$ | 0.86 | 1 | 0 | 0 | $A$ |
| $x_9$ | 0.93 | 0 | 1 | 1 | $C$ |
| $x_{10}$ | 0.47 | 0 | 1 | 1 | $C$ |
| $x_{11}$ | 0.73 | 1 | 0 | 0 | $A$ |
| $x_{12}$ | 0.89 | 1 | 2 | 0 | $B$ |

- Entropy Calculation for $H(y_{\text{out}})$

$$H(y_{\text{out}}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{2}{7}\log_2\frac{2}{7} - \frac{2}{7}\log_2\frac{2}{7} = 1.557$$

- Calculating $H(y_{\text{out}}|y_2)$

$$H(y_{\text{out}}|y_2) = \frac{4}{7} \times \left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{3}{7} \times \left(-\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3}\right) = 1.251$$

- Information Gain for $y_2$

$$IG(y_2) = H(y_{\text{out}}) - H(y_{\text{out}}|y_2) = 1.557 - 1.251 = 0.306$$

- Calculating $H(y_{\text{out}}|y_3)$

$$H(y_{\text{out}}|y_3) = \frac{2}{7} \times (-1\log_2 1) + \frac{4}{7} \times \left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{1}{7} \times (-1\log_2 1) = 0.857$$

- Information Gain for $y_3$

$$IG(y_3) = H(y_{\text{out}}) - H(y_{\text{out}}|y_3) = 1.557 - 0.857 = 0.7$$

- Calculating $H(y_{\text{out}}|y_4)$

$$H(y_{\text{out}}|y_4) = \frac{4}{7} \times \left(-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{3}{7} \times \left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) = 0.965$$

- Information Gain for $y_4$

$$IG(y_4) = H(y_{\text{out}}) - H(y_{\text{out}}|y_4) = 1.557 - 0.965 = 0.592$$

- Information Gain Evaluation
  Since feature $y_3$ has the greatest information gain, it is selected as the splitting criterion.

- $y_3 = 0 \rightarrow$ Class A
- $y_3 = 1 \rightarrow$ Class A/B/C
- $y_3 = 2 \rightarrow$ Class B

- Data Subset for $y_1 \geq 0.3$ and $y_3 = 1$
  - Class A $\rightarrow 1$ $(x_7)$
  - Class B $\rightarrow 1$ $(x_6)$
  - Class C $\rightarrow 2$ $(x_9, x_{10})$

| $D$ | $y_1$ | $y_2$ | $y_3$ | $y_4$ | $y_{\text{out}}$ |
|---|---|---|---|---|---|
| $x_6$ | 0.30 | 0 | 1 | 0 | $B$ |
| $x_7$ | 0.76 | 0 | 1 | 1 | $A$ |
| $x_9$ | 0.93 | 0 | 1 | 1 | $C$ |
| $x_{10}$ | 0.47 | 0 | 1 | 1 | $C$ |

- Entropy Calculation for $H(y_{\text{out}})$

$$H(y_{\text{out}}) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2} = 1.500$$

- Calculating $H(y_{\text{out}}|y_2)$

$$H(y_{\text{out}}|y_2) = 1 \times \left(-\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2}\right) = 1.500$$

- Information Gain for $y_2$

$$IG(y_2) = H(y_{\text{out}}) - H(y_{\text{out}}|y_2) = 1.500 - 1.500 = 0$$

- Calculating $H(y_{\text{out}}|y_4)$

$$H(y_{\text{out}}|y_4) = \frac{1}{4} \times (-1\log_2 1) + \frac{3}{4} \times \left(-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right) = 0.689$$

- Information Gain for $y_4$

$$IG(y_4) = H(y_{\text{out}}) - H(y_{\text{out}}|y_4) = 1.500 - 0.689 = 0.811$$

- Information Gain Evaluation
  Since feature $y_4$ has the greatest information gain, it is selected as the splitting criterion.
  - $y_4 = 0 \rightarrow$ Class B
  - $y_3 = 1 \rightarrow$ Class A/C

- Decision Tree
  Since it is not possible to create any more subsets with a minimum of 4 observations, we can build the tree, taking into account that any ties are resolved by the majority class.

**2.)**

- Calculating Confusion Matrix entries
  $[\text{predicted}, \text{true}] = \text{value}$
  - $[A, A] = 2$ $[A, B] = 0$ $[A, C] = 0$
  - $[B, A] = 0$ $[B, B] = 2 + 1 + 1 = 4$ $[B, C] = 0$
  - $[C, A] = 1$ $[C, B] = 0$ $[C, C] = 1 + 2 + 2 = 5$

- Filling in the Matrix

| Predicted \True | $A$ | $B$ | $C$ |
|---|---|---|---|
| $A$ | 2 | 0 | 0 |
| $B$ | 0 | 4 | 0 |
| $C$ | 1 | 0 | 5 |

**3.)**

- Class A Calculations

$$\text{Precision}_A = \frac{TP}{TP + FP} = \frac{2}{2 + 0} = 1.0$$

$$\text{Recall}_A = \frac{TP}{TP + FN} = \frac{2}{2 + 1} = \frac{2}{3}$$

$$F1_A = 2 \times \frac{\text{Precision}_A \times \text{Recall}_A}{\text{Precision}_A + \text{Recall}_A} = 2 \times \frac{1.0 \times \frac{2}{3}}{1.0 + \frac{2}{3}} = \frac{4}{5}$$

- Class B Calculations

$$\text{Precision}_B = \frac{TP}{TP + FP} = \frac{4}{4 + 0} = 1.0$$

$$\text{Recall}_B = \frac{TP}{TP + FN} = \frac{4}{4 + 0} = 1.0$$

$$F1_B = 2 \times \frac{\text{Precision}_B \times \text{Recall}_B}{\text{Precision}_B + \text{Recall}_B} = 2 \times \frac{1.0 \times 1.0}{1.0 + 1.0} = 1.0$$

- Class C Calculations

$$\text{Precision}_C = \frac{TP}{TP + FP} = \frac{5}{5 + 1} = \frac{5}{6}$$

$$\text{Recall}_C = \frac{TP}{TP + FN} = \frac{5}{5 + 0} = 1.0$$

$$F1_C = 2 \times \frac{\text{Precision}_C \times \text{Recall}_C}{\text{Precision}_C + \text{Recall}_C} = 2 \times \frac{\frac{5}{6} \times 1.0}{\frac{5}{6} + 1.0} = \frac{10}{11}$$

- Conclusion
  The Class with the lowest training F1 score is **Class A**.

**4.)**

- Bin Creation
  - Bin 1: $[0, 0.2[$
  - Bin 2: $[0.2, 0.4[$
  - Bin 3: $[0.4, 0.6[$
  - Bin 4: $[0.6, 0.8[$
  - Bin 5: $[0.8, 1]$
- Placement of $y_1$ values in respective bins
  - Bin 1: Class A = 0, Class B = 1 (0.06), Class C = 2 (0.16, 0.01)

- Bin 2: Class A = 0, Class B = 2 (0.21, 0.30), Class C = 1 (0.22)
- Bin 3: Class A = 0 , Class B = 0, Class C = 1 (0.47)
- Bin 4: Class A = 2 (0.73, 0.76), Class B = 0, Class C = 0
- Bin 5: Class A = 1 (0.86), Class B = 1 (0.89), Class C = 1 (0.93)

- Relative Frequencies Calculation

| Bin | Class A | Class B | Class C |
|---|---|---|---|
| Bin 1 | 0 | 0.25 | 0.40 |
| Bin 2 | 0 | 0.50 | 0.20 |
| Bin 3 | 0 | 0 | 0.20 |
| Bin 4 | 0.667 | 0 | 0 |
| Bin 5 | 0.333 | 0.25 | 0.20 |

- Histogram
- $n$-ary Root Split Calculation

  To find the $n$-ary root split, the dominant class must be singled out for each bin:
  - Bin 1: Dominant Class C (0.40)
  - Bin 2: Dominant Class B (0.50)
  - Bin 3: Dominant Class C (0.20)
  - Bin 4: Dominant Class A (0.667)
  - Bin 5: Class A is also present (0.333)
- Conclusion

  The $n$-ary root split can be defined based on the dominant classes:
  - The first split could occur at Bin 1 with Class C.
  - The second split could occur at Bin 2 with Class B.
  - Subsequent splits would be based on the observed distributions in the remaining bins, particularly focusing on Class A's dominance in Bin 4.

  The decision tree's splits will optimize for maximum information gain based on these empirical distributions.

## **Part II**: Programming

**1.)** To compare the performance of a kNN classifier with $k = 5$ and a naive Bayes classifier, a 5-fold stratified cross-validation was performed on the on a heart disease dataset:

**a.)** aa

**b.)** bb

**c.)** cc

**2.)** Empty

**3.)** Empty

**4.)** Empty