• Homework I

By: Group 1 { 102082 - Simão Sanguinho
103252 - José Pereira

1)

- $H(y) = -\sum_{r \in y} p(r) \cdot \log p(r)$
- $IG(z|y) = H(z) - H(z|y)$, $H(z|y) = \sum_{r \in y} P(y=r) \cdot H(z|$

- $H(y_{out}|y_1>0.4) = -\left(\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right)\right) \simeq 1.557$

- $H(y_{out}|y_1>0.4, y_2) = \frac{3}{7}\left(-\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)\right) + \frac{2}{7}\left(-\left(0 + \frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right) + \frac{2}{7}\left(-\left(1\log_2(1) + 0 + 0\right)\right) =$
$= -\frac{3}{7}\log_2\left(\frac{1}{3}\right) - \frac{2}{7}\log_2\left(\frac{1}{2}\right) \simeq 0.964984$

- $H(y_{out}|y_1>0.4, y_3) = \frac{1}{7}\left(-\left(0 + 1\log_2(1) + 0\right)\right) + \frac{4}{7}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + 0 + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right) + \frac{2}{7}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right) + 0\right)\right) = -\frac{4}{7}\log_2\left(\frac{1}{2}\right) - \frac{2}{7}\log\left(\frac{1}{2}\right) \simeq 0.857143$

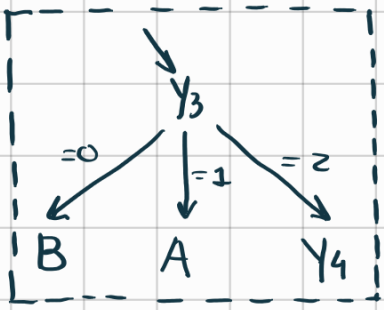
- $H(y_{out}|y_1>0.4, y_4) = \frac{2}{7}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + 0 + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right) + \frac{3}{7}\left(-\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right) + 0\right)\right) + \frac{2}{7}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + 0 + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right) =$
$= -\frac{4}{7}\log_2\left(\frac{1}{2}\right) - \frac{1}{7}\log_2\left(\frac{1}{3}\right) - \frac{2}{7}\log_2\left(\frac{2}{3}\right) \simeq 0.964984$

- $IG(y_{out}|y_1>0.4, y_2) = H(y_{out}|y_1>0.4) - H(y_{out}|y_1>0.4, y_2) = 1.557 - 0.96484 = 0.592$

- $IG(y_{out}|y_1>0.4, y_3) = H(y_{out}|y_1>0.4) - H(y_{out}|y_1>0.4, y_3) = 1.557 - 0.857143 = 0.700$

- $IG(y_{out}|y_1>0.4, y_4) = H(y_{out}|y_1>0.4) - H(y_{out}|y_1>0.4, y_4) = 1.557 - 0.964984 = 0.592$

We choose the feature that maximizes the information gain for each level. In this case, we will choose y_3.



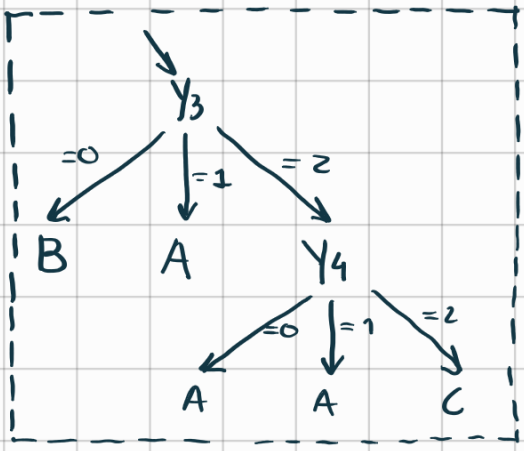
- For $y_3 = 0$, we have only one observation ({B}), therefore the node stops there.
- For $y_3 = 1$, we only have two observations ({A, B}), therefore the node stops there and we pick {A} - ascending alphabetic order.

- For $y_3 = 2$, we will need to calculate information gain again, in order to know if we pick y_2 or y_3.

- $IG(y_{out} | y_1 > 0.4, y_3 = 2, y_2) = \frac{1}{4}\left(-\left(0+0+1\log_2(1)\right)\right) + \frac{1}{4}\left(-\left(0+1\log_2(1)+0\right)\right) + \frac{1}{2}\left(-\left(1\log_2(1)+0+0\right)\right) = 0$

- $IG(y_{out} | y_1 > 0.4, y_3 = 2, y_4) = \frac{1}{2}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right)+0+\frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right) + \frac{1}{4}\left(-\left(1\log_2(1)+0+0\right)\right) + \frac{1}{4}\left(-\left(0+0+1\log_2(1)\right)\right) = \frac{1}{2}$

- Since the information gain using y_4 is higher, that's our pick for $y_3 = 2$.



- For $y_4 = 0$, we have two observations ($\{A, C\}$), therefore the node stops there and we pick $\{A\}$ - descending alphabetic order
- For $y_4 = 1$, we have one observation ($\{A\}$), therefore the node stops there.
- For $y_4 = 2$, we only have one observation ($\{C\}$), therefore the node stops there.

2)

PREDICTED \ REAL	A	B	C
A	4	1	1
B	0	2	0
C	0	1	3

3) $\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$, $\text{recall} = \frac{TP}{TP+FN}$, $\text{precision} = \frac{TP}{TP+FP}$

$\text{Recall}_A = \frac{4}{4+0+0} = 1$, $\text{Precision}_A = \frac{4}{4+1+1} = \frac{2}{3}$, $\text{F1-score}_A = 2 \times \frac{1 \times \frac{2}{3}}{1 + \frac{2}{3}} = \frac{4}{5}$

$\text{Recall}_B = \frac{2}{2+1+1} = \frac{1}{2}$, $\text{Precision}_B = \frac{2}{2+0+0} = 1$, $\text{F1-score}_B = 2 \times \frac{\frac{1}{2} \times 1}{\frac{1}{2} + 1} = \frac{2}{3}$

$\text{Recall}_C = \frac{3}{3+1+0} = \frac{3}{4}$, $\text{Precision}_C = \frac{3}{3+0+1} = \frac{3}{4}$, $\text{F1-score}_C = 2 \times \frac{\frac{3}{4} \times \frac{3}{4}}{\frac{3}{4} + \frac{3}{4}} = \frac{3}{4}$

The class with the lowest F1-score is B.

4) Spearman

Ordered $y_1 = [0.04, 0.06, 0.24, 0.32, 0.36, 0.44, 0.46, 0.52, 0.62, 0.68, 0.76, 0.9]$

ranks → 1 2 3 4 5 6 7 8 9 10 11 12

$y_1' = [3, 2, 1, 5, 4, 10, 12, 11, 7, 9, 6, 8]$

Ordered $y_2 = [0, 0, 0, 0, 0, 0, 1, 1, 1, 2, 2, 2]$

ranks → 3.5 3.5 3.5 3.5 3.5 3.5 8 8 8 11 11 11

$y_2' = [8, 11, 3.5, 3.5, 3.5, 11, 3.5, 11, 8, 3.5, 8, 3.5]$

y_1'	y_2'	$y_1'^2$	$y_2'^2$	$y_1' \times y_2'$
3	8	9	64	24
2	11	4	121	22
1	3.5	1	12.25	3.5
5	3.5	25	12.25	17.5
4	3.5	16	12.25	14
10	11	100	121	110
12	3.5	144	12.25	42
11	11	121	121	121
7	8	49	64	56
9	3.5	81	12.25	31.5
6	8	36	64	48
8	3.5	64	12.25	28
Σ	78	78	650	628.5	517.5

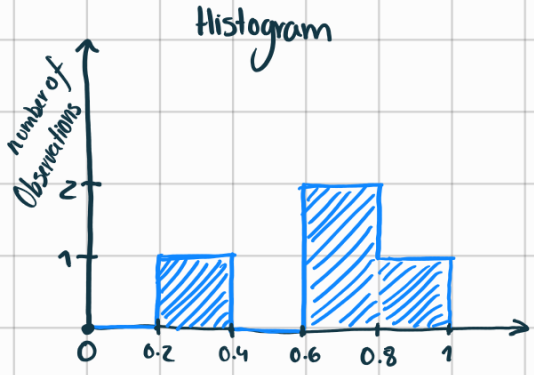
$$r = \frac{Cov(y_1, y_2)}{\sigma_{y_1} \times \sigma_{y_2}} = \frac{\Sigma y_1 y_2 - \frac{\Sigma y_1 \Sigma y_2}{n}}{\sqrt{\left(\Sigma y_1^2 - \frac{(\Sigma y_1)^2}{n}\right) \cdot \left(\Sigma y_2^2 - \frac{(\Sigma y_2)^2}{n}\right)}} = \frac{517.5 - \frac{78 \times 78}{12}}{\sqrt{\left(650 - \frac{78^2}{12}\right) \cdot \left(628.5 - \frac{78^2}{12}\right)}}$$

$$\simeq 0.07966$$

We conclude that y_1 and y_2, due to the low spearman value, have a low correlation.

5)

Class A:



$Bin_{A\,]0,0.2]} = \emptyset$

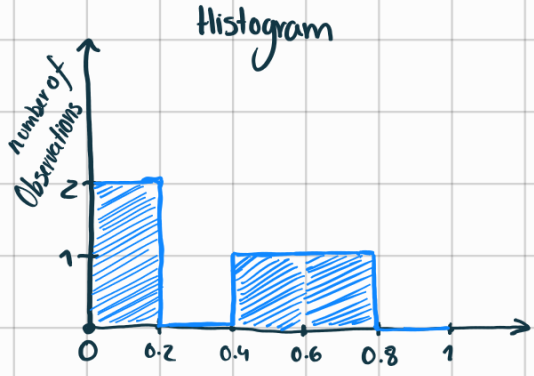
$Bin_{A\,]0.2,0.4]} = \{x_1\}$

$Bin_{A\,]0.4,0.6]} = \emptyset$

$Bin_{A\,]0.6,0.8]} = \{x_6, x_8\}$

$Bin_{A\,]0.8,1]} = \{x_7\}$

Class B:



$Bin_{B\,]0,0.2]} = \{x_2, x_3\}$

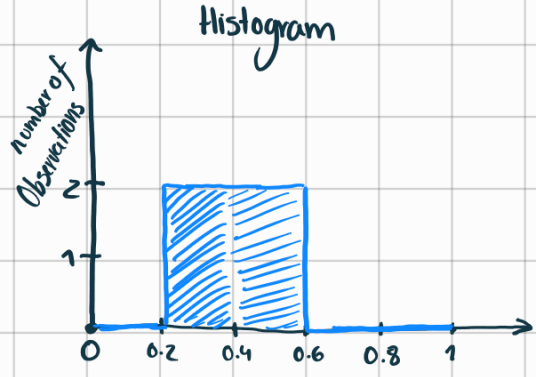
$Bin_{B\,]0.2,0.4]} = \emptyset$

$Bin_{B\,]0.4,0.6]} = \{x_9\}$

$Bin_{B\,]0.6,0.8]} = \{x_{10}\}$

$Bin_{B\,]0.8,1]} = \emptyset$

Class C:



$Bin_{C\,]0,0.2]} = \emptyset$

$Bin_{C\,]0.2,0.4]} = \{x_4, x_5\}$

$Bin_{C\,]0.4,0.6]} = \{x_{11}, x_{12}\}$

$Bin_{C\,]0.6,0.8]} = \emptyset$

$Bin_{C\,]0.8,1]} = \emptyset$

Challenge: Using the discriminant rules from these empirical distributions

In the range $]0,0.2]$ the most discriminant is Class B $\Leftrightarrow$ $Bin_{]0,0.2]} = B$

In the range $]0.2,0.4]$ the most discriminant is Class C $\Leftrightarrow$ $Bin_{]0.2,0.4]} = C$

In the range $]0.4,0.6]$ the most discriminant is Class C $\Leftrightarrow$ $Bin_{]0.4,0.6]} = C$

In the range $]0.6,0.8]$ the most discriminant is Class A $\Leftrightarrow$ $Bin_{]0.6,0.8]} = A$

In the range $]0.8,1]$ the most discriminant is Class A $\Leftrightarrow$ $Bin_{]0.8,1]} = A$