## 1. Decision trees:

a) · Fake news item: $\overset{{}^{\prime\prime}C}{FN}$ : Yes = 1, No = 0.

· $\underbrace{\text{Origin}}_{A_1}$: One-hot encoding

|  | b | s | n |
|---|---|---|---|
| Blog: | 1 | 0 | 0 |
| Soc. Net.: | 0 | 1 | 0 |
| Newspaper: | 0 | 0 | 1 |

· Exessive use of capital letters: $\overset{A_2}{\overbrace{CL}}$: T = 1, F = 0.

Information gain:

$$IG(C|A_i) = H(C) - H(C|A_i)$$

· $H(C) = -\left(\frac{1}{2}\log_2(1/2) + \frac{1}{2}\log_2(1/2)\right) = -\log_2 1/2 = 1.$

· $H(C|A_1) = P(A_1 = \text{"Blog"})P(C|A_1 = \text{"Blog"}) + \dots$ $\circledast$

$$\begin{cases} H(C|A_1 = \text{"Blog"}) = -\frac{\cancel{1}}{\cancel{2}}\log_2\frac{1}{2}\cdot\cancel{2} = 1 \\ H(C|A_1 = \text{"S.N"}) = 0 \\ H(C|A_1 = \text{"Newspaper"}) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = \underbrace{0{,}9182958341}_{\textcircled{A}} \end{cases}$$

$\circledast$ $H(C|A_1) = \frac{1}{3}\cdot 1 + 0 + \frac{1}{2}\textcircled{A} \simeq \boxed{0{,}795}$

$$\left.\begin{cases} H(C|A_2 = T) = 1 \\ H(C|A_2 = F) = 1 \end{cases}\right\} H(C|A_2) = 1\cdot\frac{2}{3} + 1\cdot\frac{1}{3} = 1$$

$\Rightarrow IG(C|A_1) = 1 - 0{,}795 = \underline{0{,}205}$

$IG(C|A_2) = 1 - 1 = 0$

First question would be: Origin

HW4-1

Notation: $A / B \Leftrightarrow A = 1, B = 0$

| Id | $a_1$<br>Cap.letters | $a_2$<br>blog | $a_3$<br>soc.net | $a_4$<br>newspaper | $a_5$<br>politics/sport | $b$<br>FN |
|---|---|---|---|---|---|---|
| $v_1$: 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| $v_2$: 2 | 1 | 0 | 1 | 0 | 0 | 1 |
| $v_3$: 3 | 1 | 0 | 0 | 1 | 1 | 0 |
| $v_4$: 4 | 1 | 0 | 0 | 1 | 0 | 0 |
| $v_5$: 5 | 0 | 1 | 0 | 0 | 0 | 0 |
| $v_i$: 6 | 0 | 0 | 0 | 1 | 1 | 1 |

I use $\|\cdot\|_1$ for distance, using rows as vectors in $\mathbb{R}^5$:

$v_i = (a_1, a_2, a_3, a_4, a_5)$

$k = 3$

Data: $(1, 0, 0, 1, 0)$

Distance: $d(v, v_i)$:

| | |
|---|---|
| 1: | 3 |
| 2: | 2 |
| 3: | 1 |
| 4: | 0 |
| 5: | 3 |
| 6: | 2 |

Preference: lowest id in case of tie.

3 nearest neighbors: 4, 3, 2
$\downarrow \ \downarrow \ \downarrow$
0 0 1

Prediction: $0 = \boxed{\text{No}}$

HW4-2

## 2 Decision-trees.

$\boxed{si : 1, \ no : 0.}$ $H(C) = 1$

a)

$$\begin{cases} H(C \mid \text{Age-range} = 1) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) = 0{,}9183 \\ H(C \mid \text{Age-range} = 2) = \nearrow \quad = 0{,}9183 \\ H(C \mid \text{Age-range} = 3) = -\frac{1}{2}\log_2(1/2) - \frac{1}{2}\log_2(1/2) = -\log_2(1/2) = 1 \end{cases}$$

$$\Rightarrow H(C \mid \text{Age-range}) = 0{,}9183\left(\frac{3}{8} + \frac{3}{8}\right) + 1 \cdot \frac{2}{8} = 0{,}938725$$

$$\Rightarrow \underline{IG(\text{Age-range}) = 0{,}061275}$$

$$\left.\begin{cases} H(C \mid \text{Kids} = 1) = 1 \\ H(C \mid \text{kids} = 0) = 1 \end{cases}\right\} \begin{array}{l} H(C \mid \text{Kids}) = 1 \\ \underline{IG(\text{Kids}) = 0} \end{array}$$

$$\begin{cases} H(C \mid \text{Changed}=1) = 0 \\ H(C \mid \text{Changed}=0) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right) = 0{,}72193 \end{cases}$$

$$\Rightarrow H(C \mid \text{Changed}) = 0{,}72193 \cdot \frac{5}{8} = 0{,}45121$$

$$\Rightarrow \underline{IG(\text{Changed}) = 0{,}5488}$$

Root: Has changed companies before?

b) Vector: $(\underset{\overset{\uparrow}{C_2 = \{1,2\}}}{\text{Age range}}, \underset{\overset{\uparrow}{C_2}}{\text{Has Kids}}, \underset{\overset{\uparrow}{C_2}}{\text{Has changed c.}})$.

$\boxed{\text{Distance : } \|\cdot\|_1}$

$v_1 = (1,0,0) \in \text{NO}$

$v_2 = (1,1,0) \in \text{NO}$

$v_3 = (1,1,1) \in \text{Sí}$

$v_4 = (2,0,1) \in \text{Sí}$

$v_5 = (2,0,0) \in \text{NO}$

$v_6 = (2,1,1) \in \text{Sí}$

$v_7 = (3,1,0) \in \text{NO}$

$v_8 = (3,0,0) \in \text{Sí}$

$v = (2,1,0)$

$$d(v, v_i) := \begin{cases} 1: 2 \\ 2: 1 \\ 3: 2 \\ 4: 2 \\ 5: 1 \\ 6: 1 \\ 7: 2 \\ 8: 2 \end{cases}$$

$$k = 3 \Rightarrow 3\text{ nearest: } \begin{cases} 2 \rightarrow \text{NO} \\ 5 \rightarrow \text{NO} \\ 6 \rightarrow \text{Sí} \end{cases}$$

$\boxed{\text{Prediction: NO}}$

HW4. 4