Data Mining

Date:

211-6260 Zain Al Abidin

Q1

(a) $P(x = true) = 0.5$ $P(x = false) = 0.5$

$$E(X) = -\sum_{i=0}^{c-1} P(x_i) \log_2 P(x_i)$$

$$P\{+0.5, -0.5\} = E(X) = -\frac{0.5}{1}\log_2\left(\frac{0.5}{1}\right) - \frac{0.5}{1}\log_2\left(\frac{0.5}{1}\right)$$

$$= (-0.5)(-1)(-0.5)(-1)$$

$$= 1$$

(b) $P(X = true) = 1$ $P(X = false) = 0$

$$E(X) = \frac{-1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2(0)$$

$$= -1(0) + 0$$

$$\Rightarrow 0$$

(c) $P(x = true) = 0.2$ $P(x = false) = 0.8$

$$E(X) = -\frac{0.2}{1}\log_2\left(\frac{0.2}{1}\right) - 0.8\log_2\left(\frac{0.8}{1}\right)$$

$$= 0.7219$$

Date: _____

Q2

Pointed	Nail	bolt	total
No	3	3	6
Yes	1	1	2

(a) Entropy $= -\sum_{i=0}^{c-1} p_i(t) \log_2 p_i(t)$

$$E(\text{Pointed} = \text{no}) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right)$$

$$= 1$$

$$E(\text{Pointed} = \text{yes}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$= 1$$

Threaded	Nail	bolt	total
Yes	3	3	6
no	1	1	2

$$E(\text{threaded} = \text{yes}) = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) \Rightarrow 1$$

$$E(\text{threaded} = \text{no}) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) \Rightarrow 1$$

Width	Nail	bolt	total
Slim	2	0	2
Medium	1	2	3
fat	1	2	3

Date: _____

$$\mathcal{E}(\text{width} = \text{slim}) = \frac{-2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right) \Rightarrow 0.$$

$$\mathcal{E}(\text{width} = \text{medium}) = \frac{-1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) \Rightarrow 0.918$$

$$\mathcal{E}(\text{width} = \text{fat}) = \frac{-1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) \Rightarrow 0.918$$

Entropy of Class:

$$\mathcal{E}(\text{class}) = \frac{-4}{8}\log_2\left(\frac{4}{8}\right) - \frac{4}{8}\log_2\left(\frac{4}{8}\right) \Rightarrow 1.$$

(6) Gini-Index

$$(\text{Pointed} = \text{no}) = 1 - (3/6)^2 - (3/6)^2 = 1/2$$
$$(\text{Pointed} = \text{yes}) = 1 - (1/2)^2 - (1/2)^2 = 1/2$$
$$(\text{Threaded} = \text{yes}) = 1 - (3/6)^2 - (3/6)^2 = 1/2$$
$$(\text{threaded} = \text{no}) = 1 - (1/2)^2 - (1/2)^2 = 1/2$$
$$(\text{width} = \text{slim}) = 1 - (2/2)^2 - (0/2)^2 = 0$$
$$(\text{width} = \text{medium}) = 1 - (1/3)^2 - (2/3)^2 = 0.44$$
$$(\text{width} = \text{fat}) = 1 - (1/3)^2 - (2/3)^2 = 0.44.$$

Date:

(c) Information Gain

Info Gain (Pointed) = $1 - (1 \times 6/8) - (1 \times 2/8) \Rightarrow 0.$

(threaded) = $1 - (1 \times 6/8) - (1 \times 2/8) \Rightarrow 0.$

(width) = $1 - (0 \times 2/8) - (0.92 \times 3/8) - (0.92 \times 3/8) \Rightarrow 0.312$

Therefore we can select width because it has the maximum information gain.

(d) Gain Ratio

Split information (Pointed) = $-6/8 \log_2(6/8) - 2/8 \log_2(2/8)$
$\Rightarrow 0.811.$

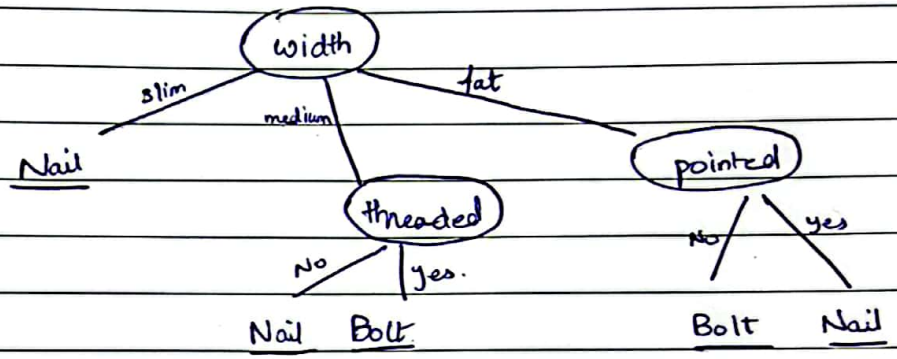
Split information (threaded) = $-6/8 \log_2(6/8) - 2/8 \log_2(2/8).$
$\Rightarrow 0.811.$

Split information (width) = $-2/8 \log_2(2/8) - 3/8 \log_2(3/8) - 3/8 \log_2(3/8).$
$\Rightarrow 1.561$

Gain Ratio (pointed) $\Rightarrow 0/0.811 \Rightarrow 0.$
Gain Ratio (threaded) $\Rightarrow 0/0.811 \Rightarrow 0$
Gain Ratio (width) $\Rightarrow 0.3115/1.561 \Rightarrow 0.199$

Therefore the best split feature is width.

Date: _____



Q3

(a) $P(A=t) = 0.03 + 0.12 + 0.17 + 0.18$

$\Rightarrow 0.5$

(b) $P(B=f) = 0.17 + 0.18 + 0.24 + 0.11$

$\Rightarrow 0.7$

(c) $P(C=t) = 0.03 + 0.17 + 0.03 + 0.24$

$\Rightarrow 0.47$

(d) $P(B=t \mid C=t) = \frac{0.03+0.03}{0.47} \Rightarrow 0.128$

(e) $P(A=f \mid c=t) = \frac{0.03+0.24}{0.47} \Rightarrow 0.574$

(f) $P(A=t \mid c=f) = \frac{0.12+0.18}{0.53} \Rightarrow 0.566$

Date:

(g) $P(A=f, c=t) = 0.03 + 0.24 \Rightarrow \underline{0.27}$

(h) $P(A=t, c=t) = 0.03 + 0.17 \Rightarrow \underline{0.20}$

(i) $P(A=t, B=f) = 0.17 + 0.18 \Rightarrow \underline{0.35}$