Priyash Gupta 20178072

Data Mining
Assignment-3

(1) (a) No
(b) Yes
(c) Yes, test may trigger more than 1 rule
(d) No, every instance will trigger at least one rule.

(2) (a) Rule Accuracy
$R_1 \rightarrow 80\%$ $R_2 \rightarrow 75\%$.
$R_3 \rightarrow 52.6\%$.
Hence, $\therefore$ R_1 has most.

(b) FOIL'S INFO GAIN
$P_0 \rightarrow 100$ $n_0 \rightarrow 400$.

$R_1 \rightarrow$ 4 +ve, 1 -ve

$$\rightarrow 4 \times \left(\log_2 \frac{4}{5} - \log_2 \frac{100}{500} \right) = 8.$$

$R_2 \rightarrow$ 30 +ve 10 -ve.

$$= 30 \times \left(\log_2 \left(\frac{30}{40} \right) - \log_2 \left(\frac{100}{500} \right) \right) = 57.2$$

$R_3 \rightarrow$ 100 +ve 90 -ve.

$$\rightarrow 100 \left(\log_2 \left(\frac{100}{190} \right) - \log_2 \left(\frac{100}{500} \right) \right) = 139.6$$

$\therefore R_1 < R_2 < R_3$.

(c) Likelihood Ratio

I { for R_1, expected freq. +ve = $\frac{5 \times 100}{500} = 1$.
for R_1, expected freq. -ve = $\frac{5 \times 400}{500} = 4$.

II { for R_2, " " +ve = $\frac{40 \times 100}{500} = 80$.
" " " " -ve = $\frac{40 \times 400}{500} = 32$.

(I) → $2 \times \left(4 \log_2 4 + 1 \log_2 \left(\frac{1}{4}\right)\right) = 12$

(II) → $2 \times \left(30 \log_2 \frac{30}{110} + 10 \log_2 \frac{10}{32}\right) = 80.85$

Similarly for R_3 → 143.09.

∴ $R_1 < R_2 < R_3$.

(d) Laplace Measure:

$R_1 = 71.93\%$ | $R_2 = 73.81\%$ | $R_3 = 52\%$.

∴ R_2 → best, R_3 → worst.

(e) M-estimate (K = 2, point @ 0.2).

$R_1 = 62.86\%$ | $R_2 = 73.38\%$ | $R_3 = 52.3\%$

R_2 → best, R_3 → worst.

(3) (a) $P(S|UG) = 0.15$ | $P(S|G) = 0.23$ | $P(G) = 0.2$ |

$P(UG) = 0.8$. $\quad P(UG) = 0.8$.

$$P(G|S) = \frac{0.23 \times 0.2}{0.15 \times 0.8 + 0.23 \times 0.2} = 0.277$$ (from Bayes' Thm.)

(b) Undergraduate ($\because P(UG) > P(G)$)

(c) Undergraduate ($\because P(UG|S) > P(G|S)$).

(d) $P(D|UG) = 0.1$ | $P(D|G) = 0.3$)

$$P(D) = P(UG) \cdot P(D|UG) + P(G) \cdot P(D|G)$$
$$= 0.8 \times 0.1 + 0.2 \times 0.3 = 0.14.$$

$$P(S) = P(S|UG) \cdot P(UG) + P(S|G) \cdot P(G)$$
$$= 0.19 \times 0.8 + 0.23 \times 0.2 = 0.166.$$

$$P(DS|G) = 0.3 \times 0.23 = 0.069.$$
$$P(DS|UG) = 0.1 \times 0.19 = 0.015.$$

$$P(G|DS) = \frac{0.069 \times 0.2}{P(DS)} = \frac{0.0138}{x}$$

$$P(UG|DS) = \frac{0.015 \times 0.8}{P(DS)} = \frac{0.012}{x}$$

$\therefore P(G|DS) > P(UG|DS)$

$\therefore$ most likely → undergraduate.

④ (a) $P(A=1|-) = 2/5 = 0.4$ $P(A=1|+) = 3/5 = 0.6$

$P(B=1|-) = 2/5 = 0.4$ $P(B=1|+) = 1/5 = 0.2$

$P(C=1|-) = 1$ $P(C=1|+) = 4/5 = 0.8$

$P(A=0|-) = 3/5 = 0.6$ $P(A=0|+) = 2/5 = 0.4$

$P(B=0|-) = 3/5 = 0.6$ $P(B=0|+) = 4/5 = 0.8$

$P(C=0|-) = 0$ $P(C=0|+) = 1/5 = 0.2$

(b) Let $P(A=0, B=1, C=0) = K$.

$= P(+ | A=0, B=1, C=0)$

$= \frac{P(A=0, B=1, C=0|+) \times P(+)}{K}$

$= \frac{P(A=0|+)\ P(B=1|+)\ P(C=0|+) \times P(+)}{K}$

$= \frac{0.008}{K}$

$P(- | A=0, B=1, C=0)$

$= \frac{P(A=0, B=1, C=0|-) \times P(-)}{K}$

$= \frac{P(A=0|-) \times P(B=1|-) \times P(C=0|-) \times P(-)}{K}$

$= \frac{0}{K} = 0.$

∴ Class label: +.

(c) m estimate with $p = 0.5$, $m = 4$.

$P(A = 0 | +) = 4/9$ $\quad P(B = 1 | -) = 4/9$

$P(A = 0 | -) = 5/9$ $\quad P(C = 0 | +) = 3/9$

$P(B = 1 | +) = 3/9$ $\quad P(C = 0 | -) = 2/9$.

(d) $P(A = 0, B = 1, C = 0) = K$.

$P(+ | A = 0, B = 1, C = 0) = 0.0247 / K$

$P(- | A = 0, B = 1, C = 0) = 0.0549 / K$.

class label $\Rightarrow -$.

(5) (a) Yes

(b) Yes

(c) No

(d) No.