Sai Mahesh Mudavat

11527206

CSCE 5380

Section-002

②

Actual value

| | | |
|---|---|---|
| Predicted value | TP | FP |
| | FN | FN |

(a)

Predicted Value

| | | |
|---|---|---|
| Actual value | TP | FN |
| | FP | TN |

Table

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $P(y=+\|C_1)$ | 0.15 | 0.2 | 0.25 | 0.37 | 0.41 | 0.55 | 0.65 | 0.8 | 0.92 | 0.99 |
| $P(y=+\|C_2)$ | 0.33 | 0.22 | 0.1 | 0.41 | 0.68 | 0.59 | 0.72 | 0.3 | 0.64 | 0.95 |
| | − | − | + | − | + | − | − | + | + | + |

(i) $C_1 \geq 0.7$

Predicted

| | | |
|---|---|---|
| Actual | TP<br>0.8, 0.92, 0.99 | FN<br>0.25, 0.41 |
| | FP<br>− | FN<br>0.15, 0.2, 0.37<br>0.55, 0.65 |

$$\begin{bmatrix} 0.8, 0.92, 0.99 & 0.25, 0.41 \\ - & 0.15, 0.2, 0.37, 0.55, 0.65 \end{bmatrix}$$

$P(Y=1/C_i) \geq 0.625$

~~Actual Values~~

| ~~Pred~~ Actual | Predicted | |
|---|---|---|
| | TP: 0.68, 0.64, 0.75, 0.95 | FN: 0.72 |
| | FP: 0.1 | TN: 0.33, 0.22, 0.59, 0.41 |

(b) F-measure:

$$\frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TN + FP + TP + FN}$$

Accuracy $(C_1)$ =

$$\frac{(0.82+0.92+0.99) + 0.15+0.2+0.37+0.55+0.65}{0.82+0.92+0.99+0.15+0.2+0.37+0.55+0.65+0.25+0.41}$$

$= 0.87$

Accuracy $(C_2)$

$$= \frac{(0.68+0.64+0.75+0.95)(+0.33+0.22+0.59+0.41)}{0.68+0.64+0.75+0.95+0.72+0.1}$$

$= 0.86$

$F_1$ measure: $\cancel{2 * \left(\frac{2.7}{3.37}\right.}$

$(C_1)$

$\cancel{2 * (0.68}$ $\frac{2 * 3}{(2 * 3 + \cancel{2*0})}$

$= \frac{6}{8} = 0.75$

$F_1$ measure ($C_2$) $= \dfrac{2 \times 4}{2 \times 4 + 1 + 1} = 0.8$

Accuracy is similar in both cases but in $F_1$ measure $C_2 > C_1$

$\boxed{C_2 \text{ is better over } C_1}$ → in terms of $F_1$ measure

(3) Rule based Binary Classification

(a) No rules are not mutually exhaustive exclusive since values are different from each other, we can have air conditioner and high mileage working with good engine

(b) Rule Set is Exhaustive — Yes

It covers all combination and every instance ~~subsets~~ gives trigger to one rule atleast

Air Conditioner = working Engine = good
AC = working Engine = bad
AC = NOT working Engine = bad

(c) Yes we need ordering because rule set is not mutually ~~exhaustive~~ exclusive

(NO)
(d) We don't need exhaustive default class for rule because instance triggers at least one rule

(5) $C_{4.5}$ Rules of strength

-> Easy to implement and produce decision tree
-> Indirect learning algorithm - produce rules to help decision making with ability to build multiple decision trees
-> Use both Categorical and Endless values

Weakness

-> Builds Empty branches. It has many nodes Zero Values

Ripper

→ Designed for Constructing models from dataset

→ Works with noisy data as it uses validation set to fix model overfitting

Weakness

Complex to create , Hard to implement.

→ C4.5 rules algorithm generates classification rules from global perspective as Rules are from decision tree which are induced with objective of partitioning the feature space into homogenous regions without focusing on class

→ Ripper generates one ~~rule~~ class at a time
Biased towards generated classes ~~at first~~ which are produced at first instance

(b) $C_{4.5}$ has class ordering scheme which is easier to interpret than scheme used by Ripper as there is large difference in class size

(6A) Accuracy (R)

$$\frac{n_{correct}}{n_{covers}}$$

$R_1 \rightarrow$ positive = 5 negative = 1

$R_2 \rightarrow$ positive = 20 negative = 5

$R_3 \Rightarrow$ positive = 50 negative = 40

Total $\rightarrow$ $\underline{R_1}$ = 5+1 = 6

$$\frac{\text{Positive}}{\text{Total}} = \frac{5}{6} \Rightarrow 0.83 \Rightarrow 83\%.$$

$\underline{R_2}$

20 + 5

Total = 25 $= \frac{20}{25} = 80\%$

$\underline{R_3}$

50 + 40 = 90 $= \frac{50}{90} = 55\%$

→ Best candidate — $R_1$ — highest Accuracy

— $R_3$ — Lowest Accuracy

(b) Foil information gain

$$Gain(R_0, R_1) = P_1 \times \left[\log_2\left(\frac{P_1}{P_1 + n_1}\right) - \log\left(\frac{P_0}{P_0 + n_0}\right)\right]$$

$R_0$ - initial Rule

150 positive negative = 200

$$R_1 = 5\left[\log_2\left(5/6\right) - \log_2\left(\frac{50}{250}\right)\right] =$$

$$5\left(-0.263 - (-2.32)\right)$$

$$= 10.32$$

$$R_2 = 20\left[\log_2\left(20/25\right) - \log_2\left(\frac{50}{250}\right)\right]$$

$$20\left(-0.32 - (-2.32)\right)$$

$$= 40.02$$

$$R_3 = 50\left[\log_2\left(\frac{50}{90}\right) - \log_2\left(\frac{50}{250}\right)\right]$$

$$= 50(-0.84 - (-2.32))$$

$$= \cancel{60}.\quad 74\%$$

$R_3$ is highest and Best $\quad$ $R_i$ is worst candidate