Q3) A data set D consist of {400 – , 100 +} instances. For the D, following are the rule are generated

R1: X --> + (90 - & 100 + instances are covered)
R2: Y --> + (10 - & 30 + instances are covered)
R3: Z --> + (1 - & 4 + instances are covered)

Apply the following measure on rule

l. FOIL information gain
m. m-estimate measure
n. Laplace measure
o. Likelihood ratio statistic
p. Rule accuracy
q. Comment on the rules based on the values achieved from Q1. a-e.

R_1: $A \longrightarrow +$ (covers 4 positive and 1 negative examples),
R_2: $B \longrightarrow +$ (covers 30 positive and 10 negative examples),
R_3: $C \longrightarrow +$ (covers 100 positive and 90 negative examples),

determine which is the best and worst candidate rule according to:

(a) Rule accuracy.

Answer:

The accuracies of the rules are 80% (for R_1), 75% (for R_2), and 52.6% (for R_3), respectively. Therefore R_1 is the best candidate and R_3 is the worst candidate according to rule accuracy.

(b) FOIL's information gain.

Answer:

Assume the initial rule is $\emptyset \longrightarrow +$. This rule covers $p_0 = 100$ positive examples and $n_0 = 400$ negative examples.

The rule R_1 covers $p_1 = 4$ positive examples and $n_1 = 1$ negative example. Therefore, the FOIL's information gain for this rule is

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

The rule R_2 covers $p_1 = 30$ positive examples and $n_1 = 10$ negative example. Therefore, the FOIL's information gain for this rule is

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

The rule R_3 covers $p_1 = 100$ positive examples and $n_1 = 90$ negative example. Therefore, the FOIL's information gain for this rule is

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

Therefore, R_3 is the best candidate and R_1 is the worst candidate according to FOIL's information gain.

(c) The likelihood ratio statistic.

Answer:

For R_1, the expected frequency for the positive class is $5 \times 100/500 = 1$ and the expected frequency for the negative class is $5 \times 400/500 = 4$. Therefore, the likelihood ratio for R_1 is

$$2 \times \left[4 \times \log_2(4/1) + 1 \times \log_2(1/4) \right] = 12.$$

For R_2, the expected frequency for the positive class is $40 \times 100/500 = 8$ and the expected frequency for the negative class is $40 \times 400/500 = 32$. Therefore, the likelihood ratio for R_2 is

$$2 \times \left[30 \times \log_2(30/8) + 10 \times \log_2(10/32) \right] = 80.85$$

For R_3, the expected frequency for the positive class is $190 \times 100/500 = 38$ and the expected frequency for the negative class is $190 \times 400/500 = 152$. Therefore, the likelihood ratio for R_3 is

$$2 \times \left[100 \times \log_2(100/38) + 90 \times \log_2(90/152) \right] = 143.09$$

Therefore, R_3 is the best candidate and R_1 is the worst candidate according to the likelihood ratio statistic.

(d) The Laplace measure.

Answer:

The Laplace measure of the rules are 71.43% (for R_1), 73.81% (for R_2), and 52.6% (for R_3), respectively. Therefore R_2 is the best candidate and R_3 is the worst candidate according to the Laplace measure.

(e) The m-estimate measure (with $k = 2$ and $p_+ = 0.2$).

Answer:

The m-estimate measure of the rules are 62.86% (for R_1), 73.38% (for R_2), and 52.3% (for R_3), respectively. Therefore R_2 is the best candidate and R_3 is the worst candidate according to the m-estimate measure.