

## Assignment#4 Solutions (Chapter 5)

4. Consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules,

$R_1: A \rightarrow +$  (covers 4 positive and 1 negative examples),

$R_2: B \rightarrow +$  (covers 30 positive and 10 negative examples),

$R_3: C \rightarrow +$  (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  $\emptyset \rightarrow +$ . 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 information gain for this rule is

$$4 [\log(4/5) - \log(100/500)] = 8.$$

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

$$30 [\log(30/40) - \log(100/500)] = 57.2$$

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

$$100 [\log(100/190) - \log(100/500)] = 139.6$$

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

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

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

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

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

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

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.

5. Figure 5.4 illustrates the coverage of the classification rules  $R_1$ ,  $R_2$ , and  $R_3$ . Determine which is the best and worst rule according to:

a) The likelihood ratio statistic.

**Answer:** There are 29 positive examples and 21 negative examples in the data set.  $R_1$  covers 12 positive examples and 3 negative examples. The expected frequency for the positive class is  $15 \times 29/50 = 8.7$  and the expected frequency for the negative class is  $15 \times 21/50 = 6.3$ . Therefore, the likelihood ratio for  $R_1$  is

$$2 \times [ 12 \times \log_2(12/8.7) + 3 \times \log_2(3/6.3) ] = 4.71.$$

$R_2$  covers 7 positive examples and 3 negative examples. The expected frequency for the positive class is  $10 \times 29/50 = 5.8$  and the expected

frequency for the negative class is  $10 \times 21/50 = 4.2$ . Therefore, the likelihood ratio for  $R2$  is

$$2 \times [ 7 \times \log_2(7/5.8) + 3 \times \log_2(3/4.2) ] = 0.89.$$

$R3$  covers 8 positive examples and 4 negative examples. The expected frequency for the positive class is  $12 \times 29/50 = 6.96$  and the expected frequency for the negative class is  $12 \times 21/50 = 5.04$ . Therefore, the likelihood ratio for  $R3$  is

$$2 \times [ 8 \times \log_2(8/6.96) + 4 \times \log_2(4/5.04) ] = 0.5472.$$

$R1$  is the best rule and  $R3$  is the worst rule according to the likelihood ratio statistic.

b) The Laplace measure.

**Answer:** The Laplace measure for the rules are 76.47% (for  $R1$ ), 66.67% (for  $R2$ ), and 64.29% (for  $R3$ ), respectively. Therefore  $R1$  is the best rule and  $R3$  is the worst rule according to the Laplace measure.

c) The m-estimate measure (with  $k = 2$  and  $p_+ = 0.58$ ).

**Answer:** The m-estimate measure for the rules are 77.41% (for  $R1$ ), 68.0% (for  $R2$ ), and 65.43% (for  $R3$ ), respectively. Therefore  $R1$  is the best rule and  $R3$  is the worst rule according to the m-estimate measure.

d) The rule accuracy after  $R1$  has been discovered, where none of the examples covered by  $R1$  are discarded).

**Answer:** If the examples for  $R1$  are not discarded, then  $R2$  will be chosen because it has a higher accuracy (70%) than  $R3$  (66.7%).

e) The rule accuracy after  $R1$  has been discovered, where only the positive examples covered by  $R1$  are discarded).

**Answer:** If the positive examples covered by  $R1$  are discarded, the new accuracies for  $R2$  and  $R3$  are 70% and 60%, respectively. Therefore  $R2$  is preferred over  $R3$ .

f) The rule accuracy after  $R1$  has been discovered, where both positive and negative examples covered by  $R1$  are discarded.

**Answer:** If the positive and negative examples covered by  $R1$  are discarded, the new accuracies for  $R2$  and  $R3$  are 70% and 75%, respectively. In this case,  $R3$  is preferred over  $R2$ .