CS795: Topics in Data Mining and Security
Summer 2009
Exam I
June 6, 2009
8:30am-12:00pm
Points: 100
Instructor: Ravi Mukkamala

Name:

UID:

Question	Maximum points possible	Points obtained
1a.	15	
1b.	15	
2a.	10	
2b.	15	
2c.	15	
3a.	15	
3b	15	
Total	100	

CALCULATORS ARE PERMITTED

OPEN BOOK. OPEN NOTES. OPEN MIND.

ANSWER ALL QUESTIONS

1. **[Points 30] Table 1** has 12 instances from a passenger car evaluation survey database. It has price ($), capacity (#of passengers), and safety as the attributes. The outcome is acceptability: Low, Medium, and High.
    a. Using each attribute (1R), form rudimentary rules to predict the decision.
    b. Construct a decision tree (with all relevant attributes) to model the decision process. Due to short time, do the work only for the 1st level of the tree. Clearly show work to justify your choice for the root attribute.
2. **[Points 40]**

    (a) A company received 75000 electronic applications for the position of a manager. It used two methods for screening:

    I. A data mining tool DM1 to do the screening. DM1 selected 5000 applicants.
    II. As a second opinion, it asked another company to manually go through all the 75000 applications and screen them out. They selected a total of 1500 applicants.
    III. DM1 correctly identified only 1200 of the 1500 potential candidates.

Assuming that the manual operation is the correct standard, evaluate the following for DM1: (i) # of True positives (ii) # of True Negatives (iii) # of False positives (iv) # of False negatives (v) True positive rate (vi) False positive rate (vii) Recall (viii) Precision and (ix) F-measure.

(b) A marketing firm wishes to conduct a mailing promotion campaign in a city. It costs the company $5 to print and mail each packet. If a correct customer is reached, then a $400 income can be expected. The city has 400,000 residents out of which only 5000 are potential customers. It has access to two data mining classifiers: (i) Classifier C1 selects 10,000 residents for mailing campaign out of which 200 are potential customers. (ii) Classifier C2 selects 5000 customers out of which 150 are potential customers. Calculate

(i) Lift factor with each of the classifiers

(ii) Based on the net benefit (income-cost), determine which method the firm should adopt.

(c) Table 2 provides data for a sample of 12 customers using classifiers C3 and C4. Draw an ROC curve showing the result of a random classifier, C3, and C4. Which one would you recommend based on this graph?

3. [Points 30]

(a) The decision tree in **Figure 1** currently has a depth of 3 with the maximum path having three internal nodes. Prune the tree (using subtree replacement) to the extent that is possible. Draw the confusion matrix prior to pruning and after pruning. Show your work.

(b) Using the data in Table 1, derive the rules with exceptions. Draw graph similar to Figure 6.7 (in the textbook).

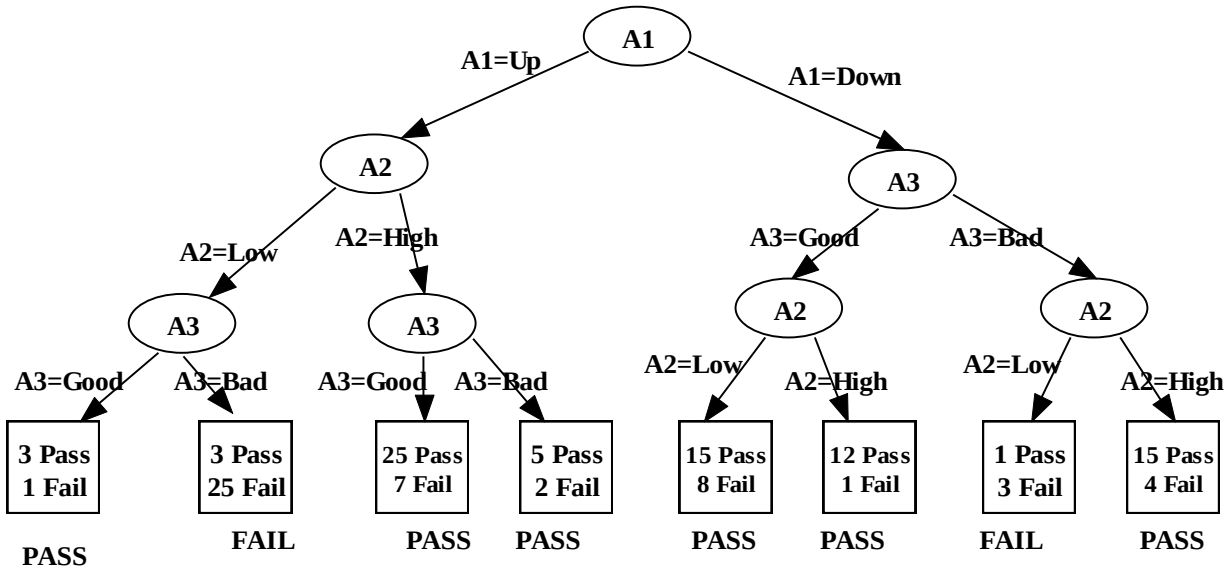

Figure 1. Decision tree to be pruned (A1-A3 are the attributes; PASS?FAIL represent the final decision for test instances that reach that node.)

Instance#	Price	Capacity	Safety	Acceptability
1	9000	4	High	High
2	15000	7	High	Low
3	11000	5	Medium	Medium
4	10000	7	High	Medium
5	12000	5	High	Medium
6	7000	4	High	High
7	8500	7	High	High
8	25000	2	High	Low
9	18000	4	Low	Low
10	10000	7	High	High
11	17000	8	Medium	Low
12	5000	4	Medium	Medium

Table 1. A subset of passenger car evaluation survey data

Customer id	Prob. Of Yes with C3	Prob. of Yes with C4	Actual response
1	0.69	0.95	YES
2	0.90	0.92	YES
3	0.40	0.97	NO
4	0.30	0.40	YES
5	0.90	0.95	NO
6	0.70	0.70	YES
7	0.40	0.35	NO
8	0.90	0.40	YES
9	0.35	0.95	NO
10	0.45	0.35	YES
11	0.95	0.95	YES
12	0.20	0.35	NO

Table 2. Data for classifiers C3 and C4

Answer sheet1:

Answer sheet2:

Answer sheet3

Answer sheet4

Answer sheet5

Answer sheet6

Answer sheet7

Answer sheet8

Answer sheet9

Answer sheet10