CS795: Topics in Data Mining and Security
Summer 2010
Final Exam
June 23, 2010
3:00pm-6:00pm
Points: 100

Name:

UID:

Question	Maximum points possible	Points obtained
1a.	10	
1b.	10	
2a.	10	
2b.	10	
3a.	10	
3b.	10	
4a.	10	
4b.	10	
5a.	10	
5b.	10	
BONUS	5	
Total	100	

CALCULATORS ARE PERMITTED

OPEN BOOK. OPEN NOTES. OPEN MIND.

ANSWER ALL QUESTIONS

Question 1 [Points 20] Referring to the personal data in Table 1, answer the following:

a. Identify two association rules with a minimum coverage of 3 and a minimum accuracy of 50%.

b. Considering the age and salary as the chosen attributes, build a kD-tree. Determine the nearest neighbor of a test instance whose age=35 and salary = 45K.

Instance#	Education	Age	Salary	Sex	State of health
1	**High School**	**25**	**25K**	**M**	**1**
2	**BS**	**35**	**55K**	**F**	**2**
3	**MS**	**55**	**85K**	**M**	**3**
4	**High School**	**30**	**40K**	**F**	**1**
5	**BS**	**40**	**60K**	**M**	**2**
6	**BS**	**20**	**50K**	**F**	**1**
7	**PhD**	**62**	**150K**	**M**	**4**
8	**MS**	**52**	**90K**	**F**	**3**
9	**High School**	**22**	**30K**	**M**	**1**
10	**PhD**	**65**	**120K**	**M**	**4**

Table 1. Personal data

Question 2. [Points 20]

a. When three data mining methods M1, M2, and M3 were provided with five test instances (with data similar to Table 1), they predicted the state of the heath of the 5 individuals as follows. The actual health state is also provided in the table. Evaluate the performance of these methods using root mean squared error, mean absolute error and correlation error as metrics and rank the methods accordingly.

Test instance #	Actual	M1	M2	M3
1	4	3	4	2
2	4	4	4	3
3	1	4	2	1
4	2	3	2	2
5	3	3	4	2
6	1	1	1	1
7	2	4	2	2

b. Suppose method M1 had predicted the state of health of a person as 1, what could be the actual value (or range) of the person using (i) Mean absolute error (ii) Mean squared error computed in 2a above.

c. Suppose a data mining method is known to estimate the salary of a person with 10% accuracy 70% of the time, and with 30% accuracy 30% of the time. If it predicts an individual salary as 50K, what could be the actual salary (range)?

Question 3. [Points 20] Using data in Table 1, answer the following.

a. Assuming the state of health as the outcome, determine the attributes that are redundant and can be removed. Explain the technique that you have used for such elimination.
b. Discretize the salary attribute with state of health as the outcome.

Question 4. [Points 20]

(a) Three hospitals are attempting to collaborate to find the causes of a specific disease without revealing their own data to others. The idea is for one site to determine an association rule and indicate its coverage and accuracy to other sites. They would test this rule and broadcast the coverage and accuracy of the rule on their data. Each would then compute the overall coverage and accuracy of the rule among all databases without sharing their data.
(a) Determine a single rule that site A can derive. Show what it would broadcasts to others.
(b) Show what sites B and C would compute in turn and broadcast.
(c) Show what each of the three sites would derive from the global information.

Data at A:

Patient#	Age	Sex	Exercise	Smoker	Diet	Disease severity
1	80	M	4	N	4	1
2	20	F	1	Y	1	3
3	50	M	3	Y	1	3
4	50	F	1	N	1	2
5	70	M	2	N	2	3

Data at B:

Patient#	Age	Sex	Exercise	Smoker	Diet	Disease severity
1	60	M	3	N	3	1
2	85	F	2	N	2	2
3	40	M	4	Y	4	2
4	35	F	1	Y	1	3

Data at C:

Patient#	Age	Sex	Exercise	Smoker	Diet	Disease severity
1	25	M	2	Y	4	3
2	40	M	3	Y	4	2

3	65	F	4	N	2	1
4	25	M	4	N	3	1
5	95	F	3	N	3	3
6	75	M	1	Y	4	3
7	40	F	2	Y	2	2

Question 5 [Points 20]. This question relates to data obfuscation. A governmental committee had asked the three sites to come together and share their data. However, they were given the option of obfuscating the data so as not to compromise the privacy of the individuals. The following changes were carried out on the data at the individual sites:

Site A: (i) Randomly increased or decreased the age by 10. (ii) Reversed the sex attribute

Site (B): (i) Each of the smoker entry was reversed with a 50% probability (i.e., half the entries were changed in the smoker attribute). (ii) The age was either increased or decreased by 5 randomly.

Site (C): (i) The sex element was reversed (ii) The diet element was perturbed by 1 (positively or negatively), always keeping the value between the legal values 1-4.

On the global data, test the validity of the rule you have developed in Question 4. Compute its coverage and accuracy. Compare these with those calculated in Question 4. Has obfuscation made any impact on the rule? Is the rule still valid?

BONUS: [Points 5] Find a linear equation to fit the three points (1,3), (2,5), and (3,6)..

Answer sheet1:

Answer sheet 2:

Answer sheet3

Answer sheet4

Answer sheet5

Answer sheet6

Answer sheet7

Answer sheet8

Answer sheet9

Answer sheet10