BITS PILANI, DUBAI CAMPUS
International Academic City, Dubai
Second Semester 2010 – 2011
Data Mining CS C415 (IV year CS)
Comprehensive Examination

Duration: 3 Hrs
Date: 07.06.2011

Weightage: 40%
MAX Marks: 40 Marks
No. of pages: 04

ANSWER ALL QUESTIONS

1. In the following data set, X is the attribute to be discretized and Y is the class label. Using entropy-based discretization on X, calculate the entropy at cut point 14 and 17. Which one of these will result in the best split? **5 Marks**

X	*Y*
4	P
28	N
0	P
16	N
24	N
26	N
12	P
18	P

2. Design a Naïve Bayesian classification model with six documents D0, D1, D2, D3, D4 and D5 as the training set. The documents are all preprocessed, 6 vocabularies are extracted and are categorized as "terrorism" and "entertainment". The preprocessed results are as shown in the following table, where the numbers indicate the frequency of occurrence of each word in the corresponding document. To avoid "zero frequency" problem use Laplacian estimation with c = 6. **10 Marks**

Training Doc	***kill***	***bomb***	***kidnap***	***music***	***movie***	***TV***	***Category***
D0	2	1	3	0	0	1	Terrorism
D1	1	1	1	0	0	0	Terrorism
D2	1	1	2	0	1	0	Terrorism
D3	0	1	0	2	1	1	Entertainment
D4	0	0	1	1	1	0	Entertainment
D5	0	0	0	2	2	0	Entertainment

Using the model designed, classify the following test document.

Test Doc	*kill*	*bomb*	*kidnap*	*music*	*movie*	*TV*	*Category*
Dt	2	1	2	0	0	1	?

Note : Show the complete detailed workings. Calculate the probabilities with 4 digits to the right of the decimal point.

3. Consider the following data set for a binary class problem.

A	*B*	*Class Label*
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

Calculate the information gain when splitting on A and B. Which attribute would the Decision Tree induction algorithm choose? **3 Marks**

4. Consider the following frequent 3-sequences :

< {1,2,3}>, <{1,2} {3}>, <{1} {2,3}>, <{1,2} {4}>, <{1,3} {4}>, <{1,2,4}>, <{2,3} {3}>, <{2,3} {4}>, <{2} {3} {3}> and <{2} {3} {4}>.
List all candidate 4-sequences produced by the candidate generation step of the **GSP** algorithm. **4 Marks**

5.

a) Convert the following table data to transaction form

A1	*A2*	*A3*
a	b	d
b	c	e

2 Marks

b) Explain in one sentence how association rule mining can be applied to categorical data? Give an example for the same. **2 Marks**

6.

a) Describe briefly the shape of the clusters formed by single-linkage and complete-linkage methods and justify your answer. **3 Marks**

b) What is the disadvantage of the average linkage method? **2 Marks**

c) For the following three clusters, each with 4 members:

Cluster 1: { (1,5) (2,4) (3,3) (2,1)}
Cluster 2: { (5,4) (6,6) (7,5) (8,8)}
Cluster 3 : {(4,1) (3,0) (5,1) (6,2)}
Find the distance between the clusters as in table below, using the three distance measures.
5 Marks

Distance between	***Single-Linkage***	***Complete-Linkage***	***Centroid***
C1 and C2			
C2 and C3			

7. A document in a collection of 10,000 documents has terms A, B and C with frequencies 3, 2 and 1 respectively. If the frequencies of these terms in the collection are 50, 1300 and 250 respectively, find their term frequency(**tf**), inverse-document frequency(**idf**) and **tf-idf.** **4 Marks.**

**********BEST OF LUCK**********

BITS PILANI, DUBAI CAMPUS
International Academic City, Dubai
Second Semester 2010 – 2011
Data Mining CS C415 (IV year CS)
Test – 2 (Open Book)

Duration: 50 minutes **Weightage: 20%**
Date: 24.04.2011 **MAX Marks: 20 Marks**
No. of pages: 2

ANSWER ALL QUESTIONS

1. On a certain transaction database, the Apriori algorithm has identified the following F_3. What are the candidate 4-item sets C_4 ? **3 M**

   $F_3 = \{\{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{1,3,5\}, \{2,3,4\}\}$

2. Let C_1, C_2 and C_3 be the confidence values of the rules $\{p\} \rightarrow \{q\}$, $\{p\} \rightarrow \{q,r\}$ and $\{p,r\} \rightarrow \{q\}$, respectively. If it is assumed that C_1, C_2 and C_3 all have different values, what are the possible relationships that may exist among C_1, C_2 and C_3 ? Which rule has the least confidence? **3 M**

3. The following contingency table summarizes a supermarket transaction data of ***hot dogs*** and ***hamburgers.***

	hot dogs	$\overline{\textbf{\textit{hot dogs}}}$	$\sum_{row}$
hamburgers	2000	500	2500
$\overline{\textbf{\textit{hamburgers}}}$	1000	1500	2500
$\sum_{col}$	3000	2000	5000

   a) Suppose that the association rule $\boldsymbol{hot\ dogs \rightarrow hamburgers}$ is mined. Given a minimum support threshold of 25% and minimum confidence threshold of 50% is this association rule strong? **2 M**

b) Based on the given data, is the purchase of hot dogs independent of the purchase of hamburger? If not, what kind of correlation relationship exists between the two? **2 M**

4. Following is a table of frequent itemsets and their respective supports in a database. Complete the table by identifying each itemset as closed (yes or no), maximal (yes or no) and both closed and maximal (yes or no). **5 M**

Itemset	*Support*	*Closed ?*	*Maximal ?*	*Both ?*
$\{Bread\}$	3			
$\{Cheese\}$	3			
$\{Juice\}$	4			
$\{Milk\}$	3			
$\{Egg\}$	3			
$\{Bread, Cheese\}$	2			
$\{Bread, Juice\}$	3			
$\{Bread, Milk\}$	2			
$\{Cheese, Juice\}$	3			
$\{Juice, Milk\}$	2			
$\{Juice, Egg\}$	2			
$\{Milk, Egg\}$	2			
$\{Bread, Cheese, Juice\}$	2			
$\{Bread, Juice, Milk\}$	2			

5. Using ***minsup*=2,** construct the **FP–Tree** for the following transactional database .Also, identify the frequent itemsets without candidate generation. **5 M**

TID	**List of item_IDs**
100	**I1, I2, I5**
200	**I2, I4**
300	**I2, I3**
400	**I1, I2, I4**
500	**I1, I3**
600	**I2, I3**
700	**I1, I3**
800	**I1, I2, I3, I5**
900	**I1, I2, I3**

**********BEST OF LUCK **********

BITS PILANI, DUBAI CAMPUS
International Academic City, Dubai
Second Semester 2010 – 2011
Data Mining CS C415 (IV year CS)
Test – 1 (Closed Book)

Duration: 50 minutes **Weightage: 25%**
Date: 06.03.2011 **MAX Marks: 25 Marks**
No. of pages: 2

1. With a neat diagram illustrate the process of knowledge discovery in databases (KDD). 4 M

2. Classify the following attributes as binary, discrete or continous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio).

   a. Angles as measured in degrees between 0 and 360.
   b. Bronze, Silver and Gold medals as awarded at the Olympics.
   c. Military rank. 1 x 3 = 3 M

3. What are the various data mining tasks? 3 M

4. For the binary vectors x = (0, 1, 0, 1) and y = (1, 0, 1, 0), compute
   a. cosine similarity measure, cos (x,y)
   b. Jaccard similarity measure, J
   c. Euclidean distance. 3 M

5. What is discretization? Name its different methods. 2 M

6. What is confusion matrix? 2 M

7. For a binary class problem, compute the impurity measure namely, Gini Index for each of the following nodes given their class distribution.

Node N_1	Count
Class = 0	0
Class = 1	10

Node N_2	Count
Class = 0	1
Class = 1	9

Node N_3	Count
Class = 0	5
Class = 1	5

1 x 3 M

8. Using the following as training data set create a decision tree based classifier using **ID3** algorithm to determine the factors affecting sunburn.Use information gain and entropy measures for node impurity.

5 M

[Note : Show the detailed workings and the final decision tree]

Name	Hair	Height	Weight	Lotion	Sunburned
Sarah	Blonde	Average	Light	No	Yes
Dana	Blonde	Tall	Average	Yes	No
Alex	Brown	Short	Average	Yes	No
Annie	Blonde	Short	Average	No	Yes
Emily	Red	Average	Heavy	No	Yes
Pete	Brown	Tall	Heavy	No	No
John	Brown	Average	Heavy	No	No
Katie	Blonde	Short	Light	Yes	No

**********BEST OF LUCK **********

BITS PILANI, DUBAI CAMPUS
International Academic City, Dubai
Second Semester 2010 – 2011
Data Mining CS C415 (IV year CS)
Quiz 2 (Closed Book)

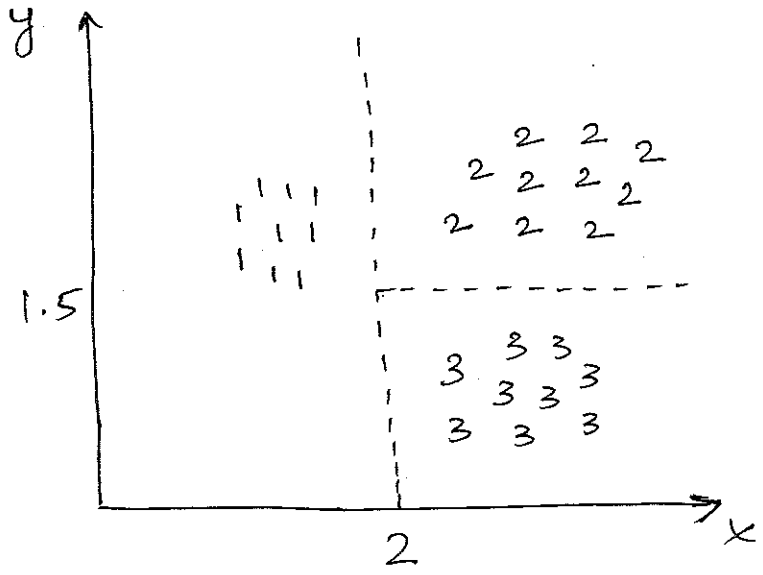
Duration : 20 minutes **Weightage : 7%**
09.05.2011 **MAX : 7 Marks**
No. of pages : 04

ANSWER ALL QUESTIONS

ID No: **Name :**

---

1. Describe the cluster shown below using three rules. **1 Mark**



Ans:

2. Applying k-means clustering using Euclidean distance over the following data set has resulted in 3 clusters C1, C2 and C3. The cluster memberships are C1 = { S1, S9}, C2 = { S2, S5, S6, S10} and C3 = {S3, S4, S7, S8}.

a) Find the centroids of the three clusters. **0.5 x 3 = 1.5 Marks**

b) Find the within cluster distance of C1 and between cluster distance of C1-C2. **1 x 2 = 2 Marks**

Student	***Age***	***Mark1***	***Mark2***	***Mark3***
S1	18	73	75	57
S2	18	79	85	75
S3	23	70	70	52
S4	20	55	55	55
S5	22	85	86	87
S6	19	91	90	89
S7	20	70	65	60
S8	21	53	56	59
S9	19	82	82	60
S10	47	75	76	77

3. Draw the dendrogram after applying hierarchial agglomerative clustering to objects O1, O2, O3 and O4 using complete linkage as distance metric. The following matrix shows the mutual distance of the four objects. Show the detailed workings. **2.5 Marks**

	O1	*O2*	*O3*	*O4*
O1	0			
O2	1	0		
O3	11	2	0	
O4	5	3	4	0

Ans :

BITS PILANI, DUBAI CAMPUS
International Academic City, Dubai
Second Semester 2010 – 2011
Data Mining CS C415 (IV year CS)
Quiz 1 (Closed Book)

Duration : 20 minutes **Weightage : 8%**
28.03.2011 **MAX : 16 Marks**
No. of pages : 03

VERSION A

ANSWER ALL QUESTIONS

ID No: **Name :**

---

1. What are the two important properties of the rule set generated by a rule-based classifier?

   Ans:

   2 M

2. Why is naïve Bayesian Classification called naïve?

   Ans :

   1 M

3. By which of the neural networks model can XOR function be learnt?

   Ans: 1 M

4. Suppose you are running a majority classifier on the following training set which consists of 10 data points. Each data point has a class label of either 0 or 1. A majority classifier is defined to output the class label that is in the majority in the training set, regardless of the input. If there is a tie in the training set, then always output class label 1. What is the training error? (report as a ratio) 2 M

Ans:

5. State whether the following data set results in a linearly separable classification. 3 M

X	Y	Out
0	0	1
0	1	1
1	0	0
1	1	1

Ans:

6. The original data set represents points **(x, y),** where **x** is the attribute and **y** is the class label. A set of two bootstrap samples of the original data set are given below.

Original data	**Bootstrap Sample1**	**Bootstrap Sample 2**
(1.76, 1)	(1.76, 1)	(1.76, 1)
(1.84, 1)	(1.76, 1)	(2.01, 1)
(1.69, 0)	(2.01, 1)	(1.76, 1)
(1.82,1)	(1.82, 1)	(1.76, 1)
(2.01, 1)	(2.01, 1)	(1.69, 0)
(1.73, 0)	(1.76, 1)	(1.82, 1)

Using an ensemble of 1 - nearest neighbor classifiers, classify each of the test data **x** given below [Write only the class. Detailed working is not necessary]

a) X = 1.77 Ans :

b) X = 1.69 Ans :

4M

7. Consider a nearest neighbor classifier that chooses the label for a test point to be the label of its nearest neighboring training example. What is its leave-one-out cross validated error (report as a ratio) for the following data. (+ and - indicate labels of the points. 3 M

Ans:

****