BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
First Semester 2011-12
Comprehensive Examination (Closed Book)

No. of Questions: 13

No. of Pages : 2

Course Number & Title : CS C415, Data Mining **Weightage : 40%** **Marks : 40**

Duration:3hrs. Date:02-01-2012 **Time:12.30p.m.–3.30p.m.** **Year: IV year/CS**

ANSWER ALL QUESTIONS SEQUENTIALLY

1. Distinguish between nominal and ordinal attributes with suitable examples. 1 Marks

2. What is the significance of feature subset selection in data mining? Draw a neat flow-chart showing the feature subset selection process. Distinguish between redundant features and irrelevant features with an example for each. 3 Marks

3. For the following vectors, x and y, calculate the indicated similarity or distance measures.
   x = (0, 1, 0, 1), y = (1, 0, 1, 0) cosine, Jaccard 1 Mark

4. Write a simple mathematical expression, to calculate the number of various types of errors in a classification problem, where m is the number of classes. Justify your answer. 2 Marks

5. Using the following weather data set (for playing tennis) as training data design a Naïve Bayesian classifier. Use the model to predict the value of the class attribute **play** in the given test data.
   5 + 1 = 6 Marks

Training Data Set

Item No.	Outlook	Temperature	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Rainy	Mild	High	False	Yes
5	Rainy	Cool	Normal	False	Yes
6	Rainy	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Sunny	Mild	High	False	No
9	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

Test Data

Outlook	Temperature	Humidity	Windy	Play
Sunny	Cool	High	True	?

6. Write the pseudo code of the ensemble bagging algorithm. 3 Marks

7. What are the differences between association rule mining and class association rule mining? 2 Marks

8. Construct the FP – tree for the following transaction data set, where **the supports of the items in descending order is a, b, c, d and e.** 5 Marks

TID	Items
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

9. The distance matrix of 4 objects O_1, O_2, O_3 and O_4 in a data set is given in the following table:-

	O_1	O_2	O_3	O_4
O_1	0			
O_2	1	0		
O_3	11	2	0	
O_4	5	3	4	0

Using hierarchial agglomerative clustering technique with 1)complete linkage and 2) average linkage as inter-cluster distance measure, show the detailed working of clustering these objects by both methods. Draw the final dendrogram.

5 + 1 = 6 Marks

10. Discretize the values { 2, 4, 10, 12, 3, 20, 30, 11, 25} of a numerical attribute into two bins using K-means clustering. Assume the first two values as the initial seeds and absolute value of simple numerical difference as distance measure. Show the detailed working and the final values in the two bins. 5 Marks

11. What is web mining? Name the various web mining tasks? 1 Mark

12. What are outliers? What are the various types of outliers? Describe shortly each type of outlier. 2 Marks

13. What is term-frequency (tf) and inverse-document frequency (idf) used in the vector-space model by text mining algorithms? What is the significance of the measure tf x idf ? 3 Marks

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

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City , Dubai
First Semester 2011-12

Test – 2(Open Book)

No. of Questions: 4	
No. of Pages	: 2

Course Number & Title : CS C415 – Data Mining **Weightage : 20%**
Duration : 50 minutes **Date:13-11-2011** **Year : IV year/CS** **Marks : 20**

Note: Answer All Questions Sequentially

1. A variation of Ada Boost algorithm with decision stump as the base classifier is described below. Here N is the number of samples and the weighted error of the base classifier C_i is given by

$$\varepsilon_i = \sum_{j=1}^{N} W_j\,(i)\delta(C_i(X_j) \neq Y_j) \text{ where}$$

$$\delta(C_i(X_j) \neq Y_j) = \begin{cases} 1 \text{ if } C_i(X_j) \neq Y_j \text{ is true} \\ 0 \text{ otherwise} \end{cases}$$

Importance of the classifier $\alpha_i = \ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$.

Weight update formula is given as $W_j^{i+1} = W_j^i \; x \begin{cases} 1 & if\, C_i(X_j) = Y_j \\ \frac{1-\varepsilon_i}{\varepsilon_i} & if\; C_i(X_j) \neq Y_j \end{cases}$

The final ensemble predicts the class of all training examples as given below
$C^*(X) = \arg\max \sum_{i=1}^{k} \alpha_i\, C_i(X) = Y$ where k is the number of base classifiers.

Initially the weights of all samples are equal, $1/N$. They are updated at the end of each iteration. Weights are rescaled after every time they are updated, so that they sum up to 1. The original data set (X is the feature and Y is the class label) and the samples selected with replacement for each boosting round are given below.

Original data set

x	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y	+	+	+	-	-	-	-	-	+	+

Boosting Round1:

x	0.1	0.1	0.3	0.5	0.5	0.6	0.6	0.8	0.9	1
y	+	+	+	-	-	-	-	-	+	+

Boosting Round2:

x	0.1	0.2	0.3	0.5	0.6	0.7	0.7	0.8	0.9	0.9
y	+	+	+	-	-	-	-	-	+	+

Split points:

Round	Split point
1	0.35
2	0.85

Complete the following table for the above problem, using the samples of each round.

Round	Split point	Left class	Right class	α
1				
2				

Complete the weight table below:

Round	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
1										
2										
end of Round 2										

Show the predicted class of each training sample by the ensemble C^*

X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
C^*										

(10 Marks)

2. In a multi-class problem, where $Y = \{y_1, y_2, y_3, y_4\}$, a test instance is classified as (+, +, -, -), using 1-r approach of extending binary classifiers. What is the predicted class of the test instance? (4 Marks)

3. What do you mean by coverage and accuracy of a rule in a rule-set generated by rule-based classifiers? (3 Marks)

4. The ensemble method namely, Random forests work very well for what nature of data sets? (3 Marks)

****

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City , Dubai
First Semester 2011-12

No. of Questions: 7
No. of Pages : 2

Test – 1(Closed Book)

Course Number & Title : CS C415 – Data Mining **Weightage : 25%**
Duration : 50 minutes **Date:25-9-2011** **Year : IV year/CS** **Marks : 25**

Note: Answer All Questions Sequentially

1. For each of the following problem scenarios, specify which data mining task would provide the correct solution.
   a) When customers visit a web site, what products are they most likely to buy together?
   b) What relationships exist between an individual's height, weight, age and favourite spectator sport?
   c) What characteristics differentiate people who have had back surgery and have returned to work from those who have had back surgery and not returned to their jobs?
   d) A model to decide whether or not to drill for oil. 2 M
2. Derive the formula for min-max normalization of data on [-1, 1] interval. 2 M
3. Given one-dimensional data set X = { -5.0, 23.0, 17.6, 7.23, 1.11}, normalize the data set using
   a) Decimal scaling on interval [-1, 1]. 3 M
   b) Min – Max on interval [0, 1]
4. Distinguish between a) noise and outliers, b) supervised and unsupervised discretization. 2 M
5. In the following data set, X1, X2…X5 represent preprocessed 2-dimensional document vectors. For the query vector (1.4, 1.6), show the order in which these document vectors will be retrieved using Euclidean distance. 5 M

Document	A1	A2
X1	1.5	1.7
X2	2	1.9
X3	1.6	1.8
X4	1.2	1.5
X5	1.5	1.0

6. What is the difference between classification, estimation and prediction? Identify the following tasks as one of the three. 3 M
   a) Which telephone subscribers are likely to change providers during the next three months?
   b) What is the salary of an individual who owns a sports car?
   c) Determine the characteristics that differentiate individuals who have suffered a heart attack from those who have not.
7. You are stranded on a deserted island. Mushrooms of various types grow wildly all over the island, but no other food is anywhere to be found. Some of the mushrooms have been determined as poisonous and others as not(determined by your former companion's trial and error). You are the only one remaining on the island. You have the following data to consider:

Mushroom	Is heavy	Is smelly	Is spotted	Is smooth	Is poisonous
A	No	No	No	No	No
B	No	No	Yes	No	No
C	Yes	Yes	No	Yes	No
D	Yes	No	No	Yes	Yes
E	No	Yes	Yes	No	Yes
F	No	No	Yes	Yes	Yes
G	No	No	No	Yes	Yes
H	Yes	Yes	No	No	Yes

   a) What is the entropy of 'Is Poisonous'? 1 M
   b) If a decision tree based classifier has to be modeled to classify a mushroom as poisonous or not, what would be the root node of the tree? [Show the detailed working] 2.5 M
   c) If the right subtree of the root node is the subset of records where the attribute value of the **root node** = **No**, what is the entropy of this subset? 1 M
   d) If the right subtree of the root node is the subset of records where the attribute value of the **root node** = **No**, which is the best splitting attribute of this subset? 2.5 M
   e) Draw the binary tree modeled upto question number (d) 1 M

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

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
First Semester 2011-12

No. of Questions: 04
No. of Pages : 02

Quiz – 2(Closed Book)

Name : **Id. No. :**

Course Number & Title : CS C415 – Data Mining **Weightage : 7%**
Duration : 20 minutes **Date:19-12-2011** **Year : IV year/CS** **Marks : 7**

Note: Answer All Questions

1. Consider the following set of frequent 3-itemsets from a dataset with only 6 items:
   $\{1,2,3\}, \{1,2,4\}, \{1,2,5\}, \{1,3,4\}, \{2,3,4\}, \{2,3,5\}, \{3,4,6\}$
   a) List all C_4 obtained by the candidate generation procedure in Apriori algorithm 1 M
   Ans:

   b) List all C_4 that survive the candidate pruning step of the Apriori algorithm
   Ans: 1 M

2. Write in one line the difference between FP growth and Apriori method of candidate generation. 1 M
   Ans:

3. Following is the set of frequent item sets for a data base with minsup = 2. Identify each frequent item set as closed? maximal? Both? - Write Yes/No 3 M

Itemset	Support	Closed?	Maximal?	Both?
$\{A\}$	3			
$\{B\}$	3			
$\{C\}$	4			
$\{D\}$	3			
$\{E\}$	3			
$\{A,B\}$	2			
$\{A,C\}$	3			
$\{A,D\}$	2			
$\{B,C\}$	3			
$\{C,D\}$	2			
$\{C,E\}$	2			
$\{D,E\}$	2			
$\{A,B,C\}$	2			
$\{A,C,D\}$	2			

4. What is the confidence of the rule A $\rightarrow \emptyset$ in a given data set? 1 M

Ans:

****

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
First Semester 2011-12

No. of Questions: 08
No. of Pages : 03
VERSION : A

Quiz – 1(Closed Book)

Course Number & Title : CS C415 – Data Mining **Weightage : 8%**
Duration : 20 minutes **Date:10-10-2011** **Year : IV year/CS** **Marks : 16**

Note: Answer All Questions

1. Construct the confusion matrix for a binary classification model, where the target values are either **buyer** or **non-buyer**. The model correctly predicted a buyer 516 times and incorrectly predicted a buyer 10 times. It correctly predicted a non-buyer 725 times and incorrectly predicted a non-buyer 25 times. 1.5 Marks
   Ans:

2. The data mining task is to predict, whether a customer will respond to a promotional mailing. The target has 2 categories: YES (the customer responds) and NO (the customer does not respond). Suppose a positive response generates $500 and that it costs $5 to do the mailing. Answer the following which are evaluations of the relative cost of different misclassifications on the test data. 2.5 Marks

   a) If the model predicts YES and the actual value is YES, the cost of misclassification is _____.

   b) If the model predicts YES and the actual value is NO, the cost of misclassification is _____.

   c) If the model predicts NO and the actual value is YES, the cost of misclassification is _____.

   d) If the model predicts NO and the actual value is NO, the cost of misclassification is _____.

e) What is the cost matrix of this model?

Ans :

From the benefits perspective of the model, answer the following

f) What is the benefit of correctly predicting a YES (a responder)? 1 Mark

g) What is the benefit of correctly predicting a NO (a non-responder)? How is this benefit achieved? 2 Marks

3. Information retrieval systems (like search engines) are meant for retrieving relevant documents from a document collection in response to a user query. From the context of information retrieval systems define the following terminologies in one line. 2 Marks

True Positive :

True Negative :

False Positive:

False Negative :

4. Mention any two performance issues of the kNN classifiers. 2 Marks

Ans:

5. What is the disadvantage of information gain measure used in decision tree induction?
Ans: 1 Mark

6. In PEBLS, an instance based classifier the distance between nominal features is computed using _____ metric.
1 Mark

7. What is the relationship between k-fold cross validation and leave-one-out cross validation techniques? 1 Mark
Ans:

8. Define the modified Bayes theorem for classification. 2 Marks

Ans:

****