BITS PILANI, DUBAI CAMPUS
II Semester12-13
Comprehensive Examination - Closed Book

Course No. & Title: CS C415, DATA MINING **Date: 04.06.13** **Duration: 3 Hrs**

Weightage : 40% **Max Marks: 40**

ANSWER ALL QUESTIONS

1. Explain in detail the feature subset selection process with suitable illustrations. 3 M

2. Express how dissimilarity and similarity is calculated for each of the simple attribute types mentioned in column 1 of the following table. 3 M

Attribute Type	Dissimilarity (d)	Similarity (s)
Nominal		
Ordinal		
Interval/Ratio		

3. Explain in detail the methods used for evaluating the performance of a classifier. 3 M

4. The following table summarizes a data set with three attributes A, B, C and two class labels + , - . Construct the decision tree using **classification error** to decide the best splitting attribute in each iteration. Show the detailed working. 5 M

A	B	C	No. of Instances	
			+	-
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

5. Consider the one-dimensional data set shown in the following table where x is the attribute and y is the class label.

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-

a) Classify the data point x = 5.0 according to its 1, 3, 5 and 9 nearest neighbors, using simple majority voting.

b) Classify the data point x = 5.0 according to its 1, 3, 5 and 9 nearest neighbors, using distance-weighted voting approach. 2 + 2 = 4 M

6. Consider the following set of frequent 3-itemsets L_3. Assume that there are only five items in the data set. 3 M

{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4}, {1, 3, 5}, {2, 3, 4}, {2, 3, 5}, {3, 4, 5}

a) List all C_4 obtained by a candidate generation procedure using the F_{k-1} x F_1merging strategy.

b) List all C_4 obtained by the candidate generation procedure in Apriori.

c) List all C_4 that survive the candidate pruning step of the Apriori algorithm.

7. Using the Apriori algorithm with minsup = 30% on the following transaction data set, identify each of the itemsets listed in the second table as one of N, F or I. The itemsets are labeled as :

N : If the candidate itemset is not considered to be a candidate itemset by the Apriori algorithm. Either it is not generated at all or generated but removed during the candidate pruning step.

F : If the candidate itemset is found to be frequent.

I : if the candidate itemset is found to be infrequent after support counting. 5 M

Transaction Database

Transaction ID	Items Bought
1	a, b, d, e
2	b, c, d
3	a, b, d, e
4	a, c, d, e
5	b, c, d, e
6	b, d, e
7	c, d
8	a, b, c
9	a, d, e
10	b, d

Candidate Itemset

Itemset	A	B	C	D	E	AB	AC	AD	AE
Label									

Itemset	BC	BD	BE	CD	CE	DE	ABC	ABD	ABE
Label									

Itemset	ACD	ACE	ADE	BCD	BCE	BDE	CDE	ABCD	ABCE
Label									

Itemset	ABDE	ACDE	BCDE	ABCDE
Label				

8. Use the **similarity matrix** in the following table to perform complete link hierarchical clustering. Show the detailed working and draw the dendrogram. 5 M

	P1	**P2**	**P3**	**P4**	**P5**
P1	1.00	0.10	0.41	0.55	0.35
P2	0.10	1.00	0.64	0.47	0.98
P3	0.41	0.64	1.00	0.44	0.85
P4	0.55	0.47	0.44	1.00	0.76
P5	0.35	0.98	0.85	0.76	1.00

9. Compute the entropy and purity of each cluster and for each clustering, given the following confusion matrix 5 M

Cluster	**Entertainment**	**Finance**	**Foreign**	**Metro**	**National**	**Sports**	**Total**
1	1	1	0	11	4	676	693
2	27	89	333	827	253	33	1562
3	326	465	8	105	16	29	949
Total	354	555	341	943	273	738	3204

10. Define the TFIDF weighting used in text mining for document representation. Given a document, drawn from a collection of 1000 documents, in which the following terms given in the table below occur. Calculate the TFIDF weights for each term (without normalizing). 4 M

Term	**Frequency in current document**	**No. of documents containing the term**
student	2	800
course	10	700

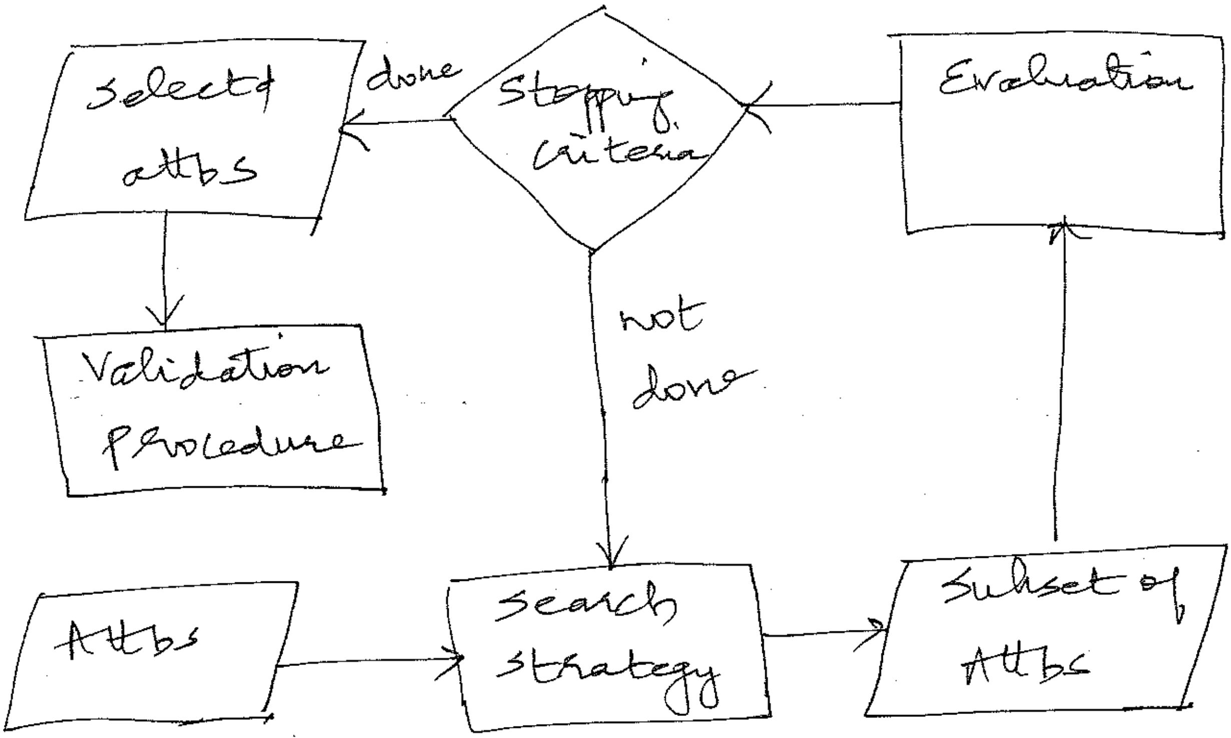
****

4/6/13 CS C415, Data Mining - Compre Exam 40% 40 Mark
IV CS

Marking scheme

1. What is feature subset selection?

Approaches; significance, procedure using a flow-chart. Evaluation . . . . .



3M

2.

Attb type	Dis (d)	sim (s)
Nominal	$d = \begin{cases} 0 & \text{if } x=y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x=y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x-y \rvert / (n-1)$	$s = 1-d$
Interval / Ratio	$d = \lvert x-y \rvert$	$s = -d, s = \frac{1}{1+d}$ $s = e^{-d}$ $s = 1 - \frac{d - mind}{maxd - mind}$

3M

3) Evaluating the performance of a Classifier

Holdout, Random subsampling, CV, Bootstrap

in detail. 3M

4) classfn error $(t) = 1 - \max_{i} \left[P(i/t) \right]$

$E_{orig} = 1 - \max\left(\frac{50}{100}, \frac{50}{100}\right) = \frac{50}{100}$ 5M

Split using A:

$A = T \quad [+ : 25, \; - : 0] \quad E_{A=T} = 0$

$\quad = F \quad [+ : 25, \; - : 50] \quad E_{A=F} = \frac{25}{75}$

$\therefore$ Gain split$_A$ $= E_{orig} - \left[\frac{25}{100} \times 0 + \frac{75}{100} \times \frac{25}{75}\right]$

$= \frac{1}{4}$ or $\frac{25}{100}$

Split using B:

	B = T	B = F
+	30	20
–	20	30

$E_{B=T} = \frac{20}{50}$

$E_{B=F} = 0.4$ or $\frac{20}{50}$

$\therefore$ Gain$_B$ $= \frac{10}{100}$ or 0.1

Split on C:

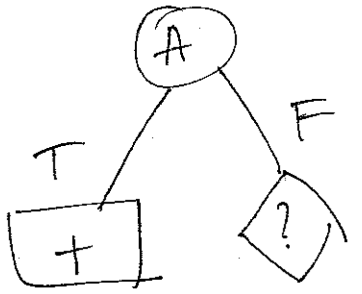
	C=T	C=F
+	25	25
−	25	25

$E_{C=T} = \frac{25}{50}$

$E_{C=F} : \frac{25}{50}$

∴ Gain $= 0$

∴ Attb A is the root.



B	C	+	−
T	T	0	20
F	T	0	5
T	F	25	0
F	F	0	25

only for A = F:

$E_{orig} = \frac{25}{75}$ or 0.33

Split on B:

	T	F
+	25	0
−	20	30

$E_{B=T} = \frac{20}{45}$

$E_{B=F} = 0$

$\text{Gain}_B = 5/75$

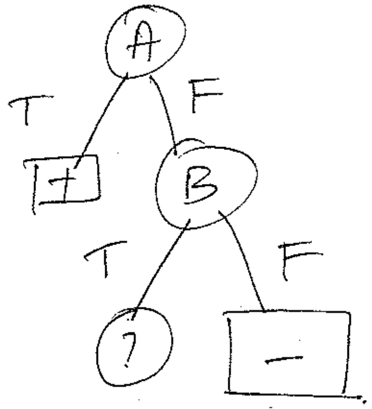
Split on C:

	T	F
+	0	25
−	25	25

$E_{C=T} = 0$

$E_{C=F} = \frac{25}{50} = 0.5$

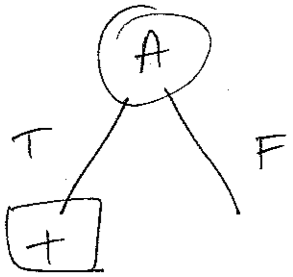
$\therefore$ $\text{Gain}_C = 0$



Split using C:

C	+	−
T	0	20
F	25	0

$\therefore$ Final Tree:



5)

a) Simple majority voting:- Test: $x = 5.0$

1-NN : +

3-NN : −

5-NN : +

9-NN : −

2 + 2 = 4M

b) Distance weighted voting:-

$$w_i = 1/d^2$$

1-NN :	+	5-NN :	+
3-NN :	+	9-NN :	+

6)

a) C_4 using the $F_{k-1} \times F_1$ strategy:-

{1,2,3,4} {1,2,3,5} {1,2,3,6}

{1,2,4,5} {1,2,4,6} {1,2,5,6} 3M

{1,3,4,5} {1,3,4,6} {2,3,4,5}

{2,3,4,6} {2,3,5,6}

b) C_4 using cand. gen in Apriori:-

{1,2,3,4} {1,2,3,5} {1,2,4,5}

{2,3,4,5} {2,3,4,6}

c) C_4 after cand. pruning step of Apriori:-

{1,2,3,4}.

7)

A : F

B : F

C : F

D : F

E : F

AB : F

AC : I

AD : F

AE : F

BC : F

BD : F

BE : F

CD : F

CE : I

DE : F

5M

ABC : N

ABD : I

ABE : I

ACD : N

ACE : N

ADE : F

BCD : I

BCE : N

BDE : F

CDE : N

ABCD : N

ABCE : N

ABDE : N

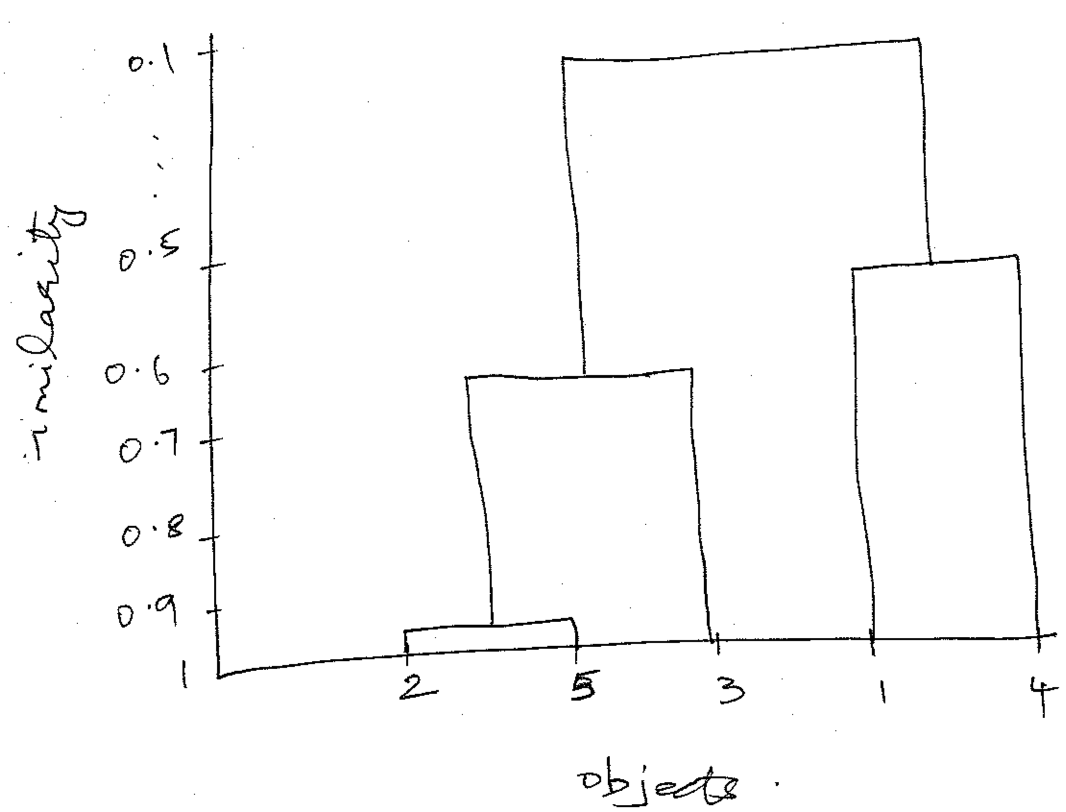
ACDE : N

BCDE : N

ABCDE : N

8) Complete linkage dendrogram..

5M



9)

cluster	Entropy	Purity
1	0.20	0.98
2	1.84	0.53
3	1.70	0.49
Total	1.44	0.61

5M

10) TFIDF : Term freq. x Inverse doc. freq 4 M
Defn :

tf : term freq.

idf : $\log_2(N/df)$

Student : $2 \times \log_2\left(\frac{1000}{800}\right) = 0.64$

Course : $10 \times \log_2\left(\frac{1000}{700}\right) = 5.15$

$\underline{\underline{2+2\,M}} = \underline{\underline{4M}}$

BITS PILANI, DUBAI CAMPUS
Dubai International Academic City, Dubai
Second Semester 2012-13

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

Test – 2(Open Book)

Course Number & Title : CS C415 – Data Mining **Weightage :** 20%
Duration : 50 minutes **Date:** 28.04.13 **Year :** IV year/CS **Marks :** 20

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

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

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

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

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

The final ensemble predicts the class of all training examples as given below

$$C^*(X) = \arg\max \sum_{i=1}^{k} \alpha_i\, C_i(X) = Y \text{ where } k \text{ is the number of base classifiers.}$$

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	+	+	+	-	-	-	-	-	+	+

x ⍰ 0.35 then y = + else 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	+	+	+	-	-	-	-	-	+	+

x ⍰ 0.85 then y = - else y = +

≤

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

Round	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. What are the disadvantages of the Euclidean distance measure used in kNN. Suggest ways for modifying it in order to improve the performance of the traditional kNN classifier.

(4 Marks)

3. Let $C_1, C_2,$ *and* C_3 be the confidence of the rules $\{p\} \rightarrow \{q\}, \{p\} \rightarrow \{q, r\}$ *and* $\{p, r\} \rightarrow \{q\}$ respectively. If $C_1, C_2,$ *and* C_3 are different values, which rule has the least confidence?

(3 Marks)

4. A training set that contains 100 positive examples and 400 negative examples. For the rule given R1 : A ⟶ + which covers 4 positive examples and 1 negative example, what is the Foil's information gain? Assume initial rule is $\emptyset \rightarrow +$, covers p0 = 100 +ve and n0 = 100 –ve examples. (3 Marks)

****

II Sem 12-13

CS C415, Data Mining   T2 - Open Book   20% wt

20 Marks, 28/4/13.

Marking Scheme

1.

Round	α
1	1.386
2	1.466

Round	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.
2	0.0625 →							→	0.25	0.2
3	0.1667 →		→	0.0285 →				→	0.1539	→

X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
c^*	–	–	–	–	–	–	–	–	+	+

10M

2. Method with justification 4M

3) $c_1 = \dfrac{s(p \cup q)}{s(p)}$ $\qquad c_2 = \dfrac{s(p \cup q \cup r)}{s(p)}$

$c_3 = \dfrac{s(p \cup r \cup q)}{s(p \cup r)}$.

$\because$ $s(p) \geq s(p \cup q) \geq s(p \cup q \cup r)$

$\therefore$ $c_1 \geq c_2$ and $c_3 \geq c_2$.

$\therefore$ c_2 has the least confidence.

$\underline{\underline{3M}}$

4) R1: $p_1 = 4$, $N_1 = 1$ $\qquad p_0 = 100$, $n_0 = 100$

$\therefore$ Gain $= 4 \times \left(\log_2 \dfrac{4}{5} - \log_2 \dfrac{100}{500}\right) = \underline{\underline{8}}$.

3M

x——x

BITS PILANI, DUBAI CAMPUS
II Semester12-13
Test 1 – Closed Book

Course No. & Title: CS C415, DATA MINING   Date: 10.03.13   Duration: 50 mins
Weightage : 25%   Max Marks: 25

ANSWER ALL QUESTIONS

1. A set of m objects are divided into k groups, where the size of the ith group is m_i. What is the difference between the following two sampling schemes, which are used to create a sample of size n < m (sampling with replacement)
   a) Randomly selects n * m_i / m elements from each group
   b) Randomly selects n elements from the data set. 3 M

2. Define Jaccard measure and cosine measure. How are they similar? 4 M

3. Draw a neat flow-chart showing the feature subset selection process. Distinguish between redundant features and irrelevant features with an example for each. 3 M

4. Using the following as training data set, develop an ID3 model to classify birds. Show the detailed working and the complete decision tree. If this tree is used for feature selection, what are the selected features? 10 M

Name	Eggs	Pouch	Flies	Feathers	Class
Cockatoo	Yes	No	Yes	Yes	Bird
Dugong	No	No	No	No	Mammal
Echidna	Yes	Yes	No	No	Marsupial
Emu	Yes	No	No	Yes	Bird
Kangaroo	No	Yes	No	No	Marsupial
Koala	No	Yes	No	No	Marsupial
Kookaburra	Yes	No	Yes	Yes	Bird
Owl	Yes	No	Yes	Yes	Bird
Penguin	Yes	No	No	Yes	Bird
Platypus	Yes	No	No	No	Mammal
Possum	No	Yes	No	No	Marsupial
Wombat	No	Yes	No	No	Marsupial

5. In the following table O_i and X_i are the objects and the attributes respectively. Construct their distance matrix using the measures a) City Block , Manhattan and b) Euclidean 5 M

Object	X1	X2	X3	X4
O1	5	6	4	9
O2	8	9	3	2
O3	3	4	5	3

****

II Sem 12 - 13

CS CA15, Data Mining Test - 1 - Closed Book

IV C.S. 10.03.13 25%

Marking Scheme

1.

a) Selects same no/. of objects from each group.

b) ~~this~~ varies

3M

2). $$J = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

$$\cos(x, y) = \frac{(x \cdot y)}{|x| \; |y|}$$

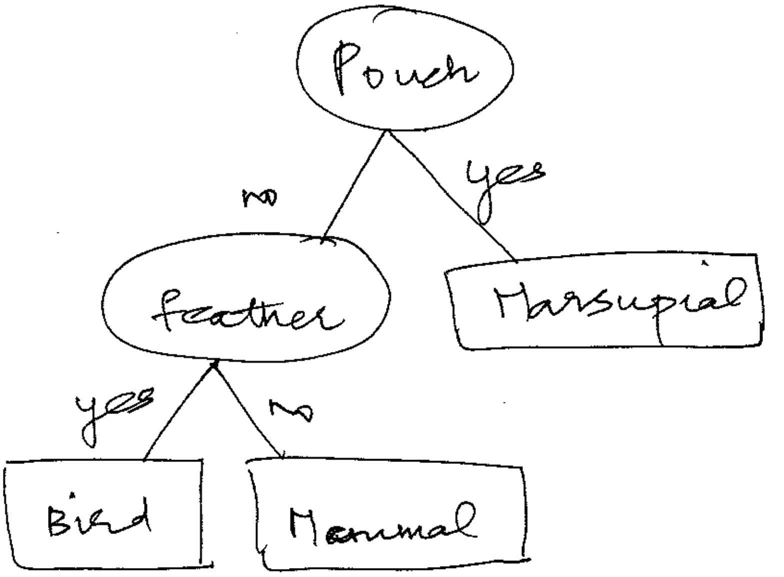
Both ignore 00 matches. $1 + 1 + 2M$ $= 4M$

3) Flow-chart:

Diff betⁿ redundat & irrelevat with ex:

$1+1+1 = 3M$

4)



10M

selected features }: Pouch, feather

5) Euclidean

$$\begin{pmatrix} 0 & & \\ 8.25 & 0 & \\ 6.7 & 7.42 & 0 \end{pmatrix}$$

City-Block

$$\begin{pmatrix} 0 & & \\ 3.5 & 0 & \\ 2.75 & 3.25 & 0 \end{pmatrix}$$

5M

BITS PILANI, DUBAI CAMPUS

II Semester12-13

Quiz 2 – Closed Book

Course No. & Title: CS C415, DATA MINING	Date: 19.05.13	Duration: 20 mins
Weightage : 7%	Max Marks: 7	
Name:	ID No.	

ANSWER ALL QUESTIONS

1. Define the following performance measures of a classifier:- 1 Mark

   a) Specificity :

   Ans :

   b) F1 Score :

   Ans:

2. a) Draw the confusion matrix of a classifier which always predicts the positive class. 0.5 + 1Mark

   Ans:

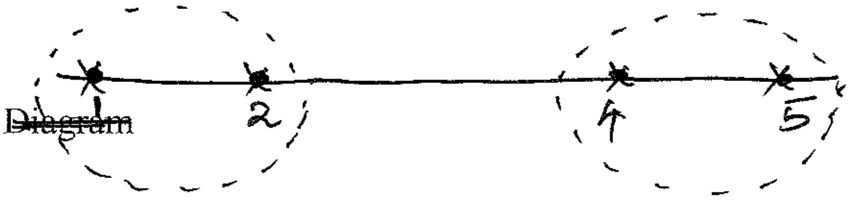
b) What is the TP rate, FP rate, Precision, F1 Score and Accuracy of this classifier.

Ans:

3. ____________ and ____________ are the FP rate and TP rate of a best classifier on the ROC space. 1 Mark

4. ____________ and ____________ are the FP rate and TP rate of a worst possible classifier on the ROC space. 1 Mark

5. The following one dimensional objects are grouped into k=2 clusters. Calculate the cluster validity measures WSS and BSS. 1.5 Marks



~~Diagram~~

Ans :

6. Define the entropy of a cluster and a clustering. 1 Mark

Ans:

****

IV CS

19/5/13 Data Mining - Quiz - 2

7% Marking Scheme

1. a) Specificity : True Negative rate = TN/N

Propr. of negative instances that are correctly classifie as negative.

~~1~~. b) F1 Score: (2x Precn x Recall)/(Precn + Recall)

0.5 +0.5 = 1M

2. a)

0.5 + 1 = 1.5M

$$\begin{array}{ccc} \text{Actual} & \multicolumn{2}{c}{\text{Predicted}} \\ & + & - \\ + & P & 0 \\ - & N & 0 \end{array}$$

b) TP rate = P/P = 1 FP Rate = N/N = 1.

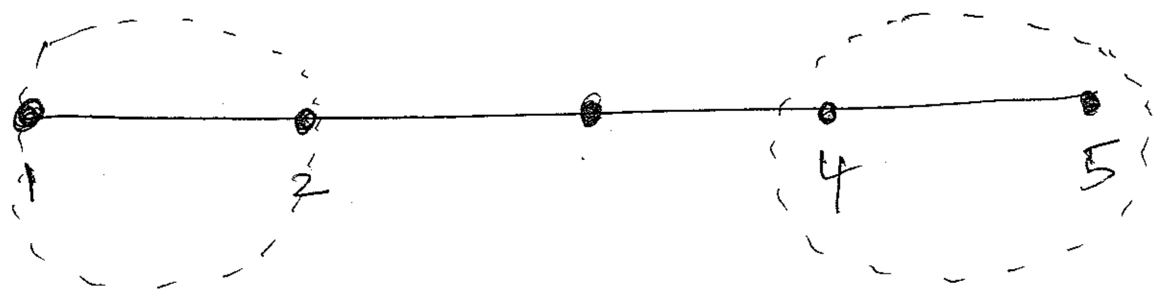
Precn = P/(P+N) F1 score = 2 x P/(2 x P+N)

Accuracy = P/(P+N)

3) a) 0 and 1 for best - - - 1M

4) 1 and 0 for worst - - - 1M

5.



$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2(4.5-3)^2$$

$$= 9.$$

1.5M

6. Entropy of a cluster j =

1M

$$e_j = \sum_{i=1}^{L} P_{ij} \log_2 P_{ij}$$

L - - - no. of classes

P_{ij} = prob of the i^{th} class in the j^{th} cluster.

clustering entropy :- $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$

K - - no. of clusters

m . . total no. of objects

m_i - - no. of objs in the i^{th} cluster

BITS PILANI, DUBAI CAMPUS

II Semester12-13

Quiz 1 – Closed Book

Course No. & Title: CS C415, DATA MINING   Date: 24.02.13   Duration: 20 mins

Weightage : 8%   Max Marks: 8

Name: **ID No.**

ANSWER ALL QUESTIONS

1. Nominal and ordinal variables together are called (quantitative/qualitative) ____________________variables. 0.5 mark

2. What are symmetric binary variables? Give an example. 1 mark
   Ans:

3. What is a ternary data? Give an example. 1 mark
   Ans:

4. If a nominal data is numerically ordered, can we use the median as its summary? Explain in one line. 0.5 mark
   Ans:

5. A cyclic ordinal data is circularly ordered. Give 2 examples. 0.5 mark
   Ans:

6. Differentiate between supervised and unsupervised learning. Give an example for both.
   Ans: 1 Mark

7. Name two data mining algorithms where discretization is useful. 0.5 mark

8. Discretize {1,1,2,2,2,3} into 2 bins using equal frequency binning. What is the problem you observe on the discretized results. 1 Mark
   Ans:

9. Identify the type of variable to represent each of the following: 1.5 Mark
   a) The newspapers you read. Ans:
   b) How many pages did you read in your data mining book? Ans:
   c) Your typing speed in words per min. Ans:
   d) Customer rating of a service in maximum score 10 Ans:
   e) Amount of sugar in a cup of orange juice. Ans:
   f) Wind classifications(breeze, gale, storm, hurricane} Ans:

10. Write two disadvantages of min-max normalization? 0.5 mark
    Ans:

****

II Sem 12-13

24/2/13 CSCH5, Data Mining Quiz1 – Marking Scheme

8% weightage 8 Marks.

1. quantitative 0.5M

2. The two choices of a binary variable have equal importance. 1M

3 If a nominal var assumes values in exactly 3 mutually exclusive categories that do not have any logical order – ternary variable. Its data – ternary dat

Ex: primary color = {R, G, B} . . 1M

4 No. Nominal data has no order . . 0.5M

5. seasons = {spring, summer, autumn, winter}

hour time = {AM, PM} . . . 0.5M

6. Difference with ex 1M

7. Decision Trees, ARM 0.5M

8. B1 = {1, 1, 2} B2 = {2, 2, 3}

Duplicates in B1 and B2 . . . 1M

9. a) Nominal

b) Ratio c) Ratio d) Ordinal ~~e) Nominal~~

e) Ratio ~~g) ordinal~~ f) ordinal 1.5M

10. Disadvs ? . . . - - . 0.5M

1

- - - - - - -