survival8

BITS WILP Data Mining Mid-Sem Exam 2017-H2

Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
First Semester 2017-18
Mid Semester Test (EC2 Regular)
Course No: IS ZC415
Course Title: Data Mining
Nature of Exam: Closed Book
Weightage: 30%
Duration: 2 Hours
Date of Exam: 23/Sep/2017 (AN)
No of pages: 2
No of questions: 4

Page: 1

Pages

- Postings Index
- Index of BITS WILP Exam Papers and Content
- Index of Lessons in Technology
- Index of Guest Interviews
- Downloads
- Book Requests

Blog Archive

▼ 2020 (31)
▼ May (1)
Covid-19 and response of IT companies (by Divjot S...
► April (6)
► March (12)
► February (6)
► January (6)
► 2019 (48)
► 2018 (31)
► 2017 (15)
► 2016 (6)

Popular Posts

You Are a Badass. How to stop doubting your greatness and start living an awesome life (Jen Sincero, 2013)
INTRODUCTION The language used in the book extremely funny and Jen Sincero still makes sure that she m...

Covid-19 and response of IT companies (by Divjot Singh)
As the Covid-19 pandemic ravages the world, many domains like airlines, tourism and services...

Innovation to beat the Coronavirus (Covid19)
Coronovirus' Exponential growth and decline In the first phase of the pandemic, we saw a...

Download fiction books (March 2018)
Download fiction books for free: Link for Google Dr...

Birla Institute of Technology & Science, Pilani
Work-Integrated Learning Programmes Division
First Semester 2017-2018

Mid-Semester Test
(EC-2 Regular)

Course No.	: IS ZC415
Course Title	: DATA MINING
Nature of Exam	: Closed Book
Weightage	: 30%
Duration	: 2 Hours
Date of Exam	: 23/09/2017 (AN)

No. of Pages = 2
No. of Questions = 4

Note:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q.1 (a) You are given the following equal frequency binned data:

bin 1	5,10,11,13
bin 2	15,35,50,55
bin 3	72,92,204,215

Show the output of smoothing by bin mean, bin median, and bin boundary. [3]

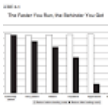
Q.1 (b) Give an example each where the following similarity measures would be useful:
(a) Euclidean
(b) Manhattan
(c) Cosine [3]

Q.1 (c) Consider the problem of spam detection. Suppose we have a dataset of spam and non-spam emails and we train a naive Bayes classifier on this dataset. For ten instances, the figure below shows the predictions of the trained classifier of the probability of an email being spam. The classifier classifies an instance as spam if and only if the predicted probability is greater than 0.990. Draw the confusion matrix. [2]

Predicted Probability	Actual Label
0.001	not spam
0.100	not spam
0.500	not spam
0.600	not spam
0.980	not spam
0.400	spam
0.800	spam
0.900	spam
0.995	spam
0.999	spam

Q.1 (d) Given the following marks scored by a student in two subjects, compute z-scores to find out in which subject the student has done better comparatively. [2]

	Marks	Std Dev. of marks	Mean mark of class
Subject 1	70	15	60
Subject 2	65	6	60



Intelligent investor (Ben Graham & Jason Zweig, 4e)
Reading from "A Note About Benjamin Graham by Jason Zweig" Here are Graham...

Life Lessons By Steve Jobs
Steve Jobs' last words will change your views on life. The billionaire passed away at the ...

Effects of news and world events on Nifty50 and stock market
Day: 10th Aug 2017 Sensex tanks 267 points. Nifty hits one-month low. 1. Market outlook: ...

Why Bill Gates would raise chickens
I'm excited about the poverty-fighting power of poultry. If you were living on $2 a day, wh...

The Essays Of Warren Buffett (Lessons For Corporate America)
INTRODUCTION Buffett has applied the traditional principles as chief executive officer of Berkshi...

How To Talk TO Anyone (92 Little Tricks For Big Success In Relationships, by Leil Lowndes) - Book Summary
There are two kinds of people in this life: Those who walk into a room and say, "Well, here I...

About Me

Ashish Jain

View my complete profile

Q.2. The following table shows the time in minute and number of messages received at that time. Use regression to find out the number of messages at time 3. [4]

Sample	Time	Number of messages
1	1	2
2	2	5

Q.3. Given below is a database of a car insurance company:

Name	AgeGroup	CarType	CrashRisk
Ben	30-40	Family	Low
Paul	20-30	Sports	High
Bill	40-50	Sports	High
James	30-40	Family	Low
John	20-30	Family	High
Steven	30-40	Sports	High

The last column is the class attribute showing the risk of a crash.

a. Draw the complete decision tree produced on the above dataset by using entropy. What will be the class label for nodes with no training samples? Show the split test you used at each node. For each leaf node, show the class and the records associated with it. [6]

b. Using the produced classifier, determine the class label of the following records
   i) {Pete, 20-30, Sports}
   ii) {Bob, 40-50, Family} [2]

Q.4. Given that min support is 2, and min confidence is 70% in the set of transactions below:

Transaction ID	Items
1	a, b, c
2	b, c, d, e
3	c, d
4	a, b, d
5	a, b, c

(a) Find all frequent itemsets using Apriori. [4]
(b) Which of the frequent 2-itemsets in the above set of transactions are closed and which of them (frequent 2-itemsets) are maximal? [2]
(c) Find all strong rules possible from frequent 3-itemsets in the above set. [2]

****

Solutions:

Q.1 (a) You are given the following equal frequency binned data:

bin 1	5,10,11,13
bin 2	15,35,50,55
bin 3	72,92,204,215

Show the output of smoothing by bin mean, bin median, and bin boundary. [3]

Answer 1(A):
Binning Methods for Data Smoothing
Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Q.1 (b) Give an example each where the following similarity measures would be useful:
(a) Euclidean
(b) Manhattan
(c) Cosine [3]

Anwer 1(B)
Euclidean distance is widely used in the Geometry where shortest distance between two points is often required to calculated as in distances between two celestial objects in space.

Manhattan distance is used in the navigation systems to calculate the distance between two points through the obstacle that are there in the path. This is also known as 'taxi cab' distance.

Cosine distance is used in 'web search, information retrieval' where two documents are represented as vectors with terms as dimensions and similarity between two documents is calculated in the form of cosine distance between them.

Q.1 (c) Consider the problem of spam detection. Suppose we have a dataset of spam and non-spam emails and we train a naive Bayes classifier on this dataset. For ten instances, the figure below shows the predictions of the trained classifier of the probability of an email being spam. The classifier classifies an instance as spam if and only if the predicted probability is greater than 0.990. Draw the confusion matrix. [2]

Predicted Probability	Actual Label
0.001	not spam
0.100	not spam
0.500	not spam
0.600	not spam
0.980	not spam
0.400	spam
0.800	spam
0.900	spam
0.995	spam
0.999	spam

Answer 1(C):

	Predicted spam	Predicted n.s.
Actual spam	TP	FN
Actual n.s.	FP	TN

Predicted probability	Prediction	Actual
0.001	n.s.	n.s.
0.100	n.s.	n.s
0.500	n.s.	n.s.
0.600	n.s.	n.s.
0.980	n.s.	n.s.
0.400	n.s.	spam
0.800	n.s.	spam
0.900	n.s.	spam
0.995	spam	spam
0.999	spam	spam

TP = 2
FN = 3
FP = 0
TN = 5

...

A **confusion matrix** is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are

known.

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

...

Q.1 (d) Given the following marks scored by a student in two subjects, compute z-scores to find out in which subject the student has done better comparatively. [2]

	Marks	Std Dev. of marks	Mean mark of class
Subject 1	70	15	60
Subject 2	65	6	60

...

In z-score normalization (or *zero-mean normalization*), the values for an attribute, *A*, are normalized based on the mean and standard deviation of *A*. A value, *v*, of *A* is normalized to *v'* by computing:

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

	Marks	Z-score
Subject 1	70	(70 – 60)/15 = 0.666
Subject 2	65	(65 - 60)/6 = 0.833

Student did better subject 2.

Q.2. The following table shows the time in minute and number of messages received at that time. Use regression to find out the number of messages at time 3. [4]

Sample	Time	Number of messages
1	1	2
2	2	5

Answer 2:
Use this:

Linear Regression

- Linear regression: involves a response variable y and a single predictor variable x

  $y = w_0 + w_1 x$

  where w_0 (y-intercept) and w_1 (slope) are regression coefficients
- Method of least squares: estimates the best-fitting straight line

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2} \qquad w_0 = \bar{y} - w_1 \bar{x}$$

- Multiple linear regression: involves more than one predictor variable
  - Training data is of the form $(\mathbf{X_1}, y_1), (\mathbf{X_2}, y_2), \ldots, (\mathbf{X_{|D|}}, y_{|D|})$
  - Ex. For 2-D data, we may have: $y = w_0 + w_1 x_1 + w_2 x_2$
  - Solvable by extension of least square method or using SAS, S-Plus
  - Many nonlinear functions can be transformed into the above

...

x-mean = 1.5
y-mean = 3.5
W1 = ((1-1.5)*(2 – 3.5) + (2-1.5)*(5-3.5)) / ((1-1.5)^2 + (2-1.5)^2) = 3
W0 = 3.5 – 3*(1.5) = -1

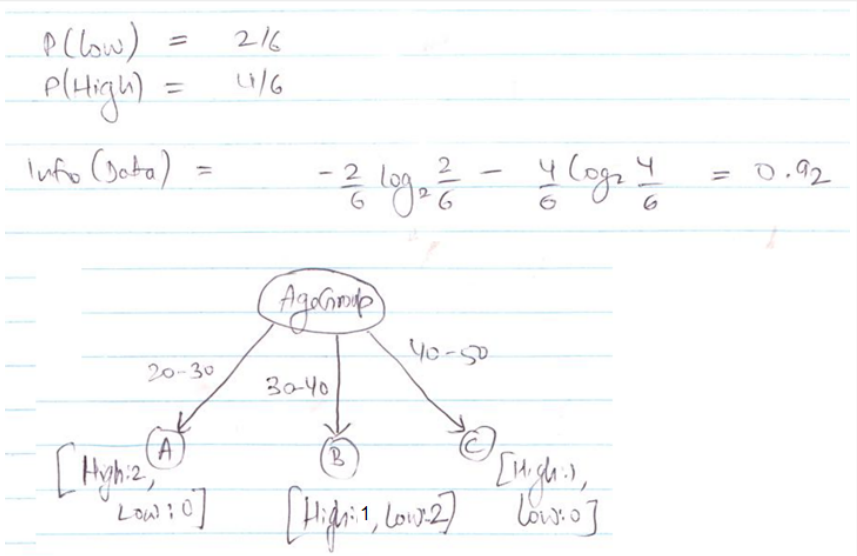
Q.3. Given below is a database of a car insurance company:

Name	AgeGroup	CarType	CrashRisk
Ben	30-40	Family	Low
Paul	20-30	Sports	High
Bill	40-50	Sports	High
James	30-40	Family	Low
John	20-30	Family	High
Steven	30-40	Sports	High

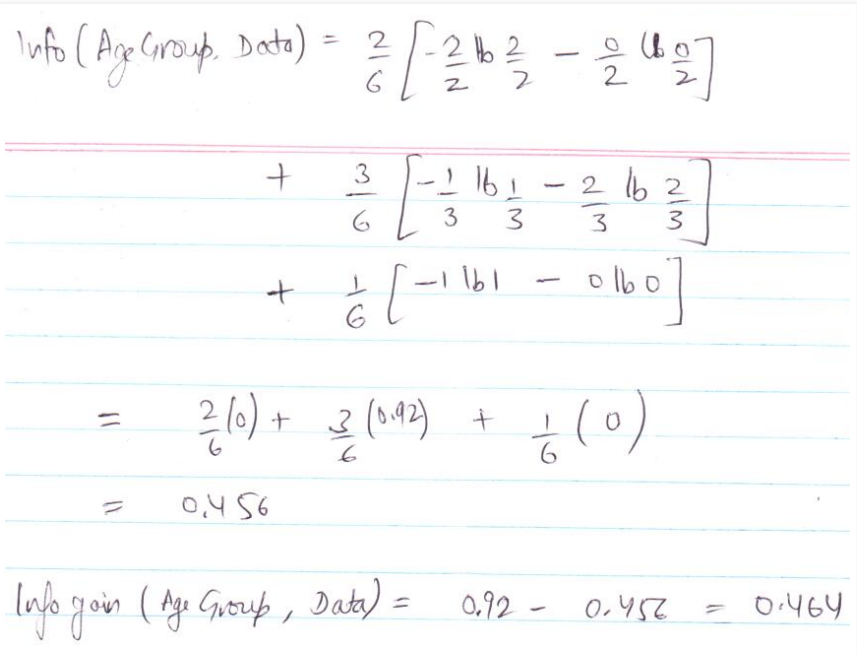
The last column is the class attribute showing the risk of a crash.

a. Draw the complete decision tree produced on the above dataset by using entropy. What will be the class label for nodes with no training samples? Show the split test you used at each node. For each leaf node, show the class and the records associated with it. [6]

"What will be the class label for nodes with no training samples?"
Answer 3(A)
Info(D) = -(2/6)(log (2/6) / log (2)) - (4/6)(log (4/6) / log (2)) = 0.92



...



...

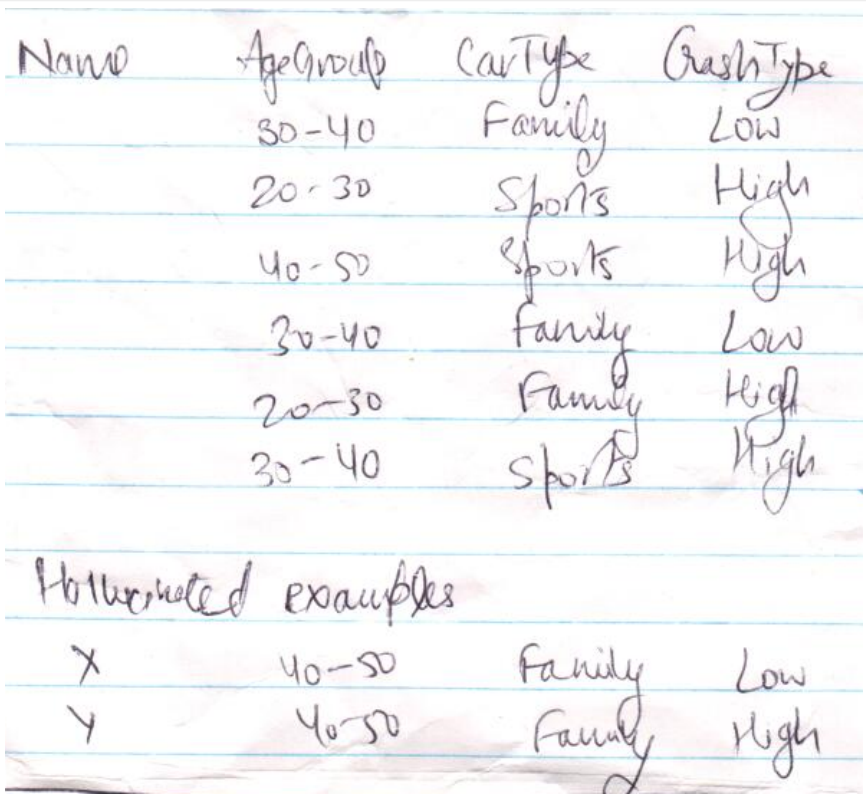
Intermediate calculation: [-(1/3)*(log(1/3) / log (2)) -(2/3)*(log(2/3) / log (2))] = 0.92
For "What will be the class label for nodes with no training samples?", from ML course:

One danger of this maximum likelihood estimate is that it can sometimes result in θ estimates of zero, if the data does not happen to contain any training examples satisfying the condition in the numerator. To avoid this, it is common to use a "smoothed" estimate which effectively adds in a number of additional "hallucinated" examples, and which assumes these hallucinated examples are spread evenly over the possible values of X_i. This smoothed estimate is given by

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \wedge Y = y_k\} + l}{\#D\{Y = y_k\} + lJ} \quad (7)$$

where J is the number of distinct values X_i can take on, and l determines the strength of this smoothing (i.e., the number of hallucinated examples is lJ). This expression corresponds to a MAP estimate for θ_{ijk} if we assume a Dirichlet prior distribution over the θ_{ijk} parameters, with equal-valued parameters. If l is set to 1, this approach is called Laplace smoothing.

...



Question 3(B)

b. Using the produced classifier, determine the class label of the following records
i) {Pete, 20-30, Sports}
ii) {Bob, 40-50, Family} [2]

...

Q.4. Given that min support is 2, and min confidence is 70% in the set of transactions below:

Transaction ID	Items
1	a, b, c
2	b, c, d, e
3	c, d
4	a, b, d
5	a, b, c

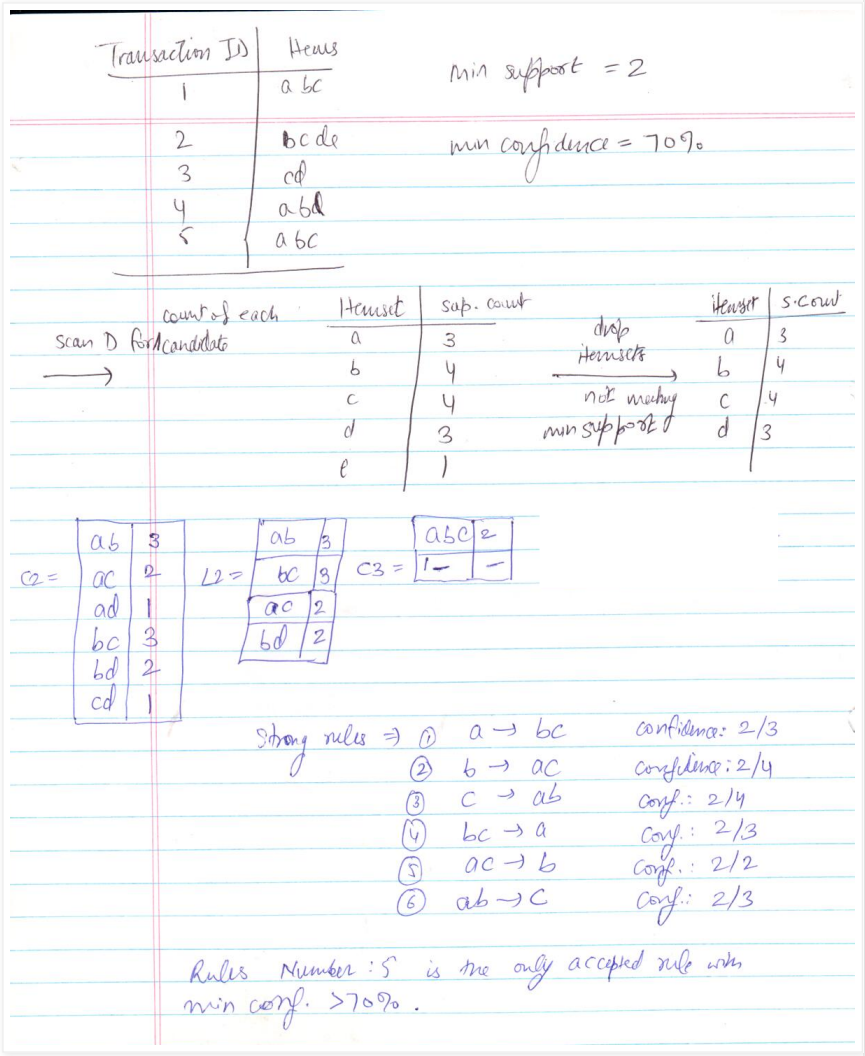
(a) Find all frequent itemsets using Apriori. [4]
(b) Which of the frequent 2-itemsets in the above set of transactions are closed and which of them (frequent 2-itemsets) are maximal? [2]
(c) Find all strong rules possible from frequent 3-itemsets in the above set. [2]

Answer 4:
Example:

- Support
  - Usefulness of discovered rules
- Confidence
  - Certainty of discovered rules

computer => antivirus software [support = 2%, confidence = 60%]

- A support of 2% means that 2% of all the transactions under analysis show that computer and a.v. are purchased together.
- A confidence of 60% means that 60% of the customers who purchased a computer also bought the software.



…

Answer 4(B)

From Stackoverflow.com

Ques: I want to find out the **maximal frequent item sets** and the **closed frequent item sets**.

Frequent item set X∈F is **maximal** if it does not have any frequent supersets.

Frequent item set X ∈ F is **closed** if it has no superset with the same frequency

---

So I counted the occurrence of each item set.

```
{A} = 4 ;  {B} = 2  ; {C} = 5  ; {D} = 4  ; {E} = 6

{A,B} = 1; {A,C} = 3; {A,D} = 3; {A,E} = 4; {B,C} = 2;
{B,D} = 0; {B,E} = 2; {C,D} = 3; {C,E} = 5; {D,E} = 3

{A,B,C} = 1; {A,B,D} = 0; {A,B,E} = 1; {A,C,D} = 2; {A,C,E} = 3;
{A,D,E} = 3; {B,C,D} = 0; {B,C,E} = 2; {C,D,E} = 3

{A,B,C,D} = 0; {A,B,C,E} = 1; {B,C,D,E} = 0
```

Min_Support set to 50%

Does **maximal** = {A,B,C,E}?

Does **closed** = {A,B,C,D} and {B,C,D,E}?

…

Ans:

Note:

- Did not check the support counts
- Let's say min_support=0.5. This is fulfilled if min_support_count >= 3

```
{A} = 4  ; not closed due to {A,E}
{B} = 2  ; not frequent => ignore
{C} = 5  ; not closed due to {C,E}
{D} = 4  ; closed, but not maximal due to e.g. {A,D}
{E} = 6  ; closed, but not maximal due to e.g. {D,E}

{A,B} = 1; not frequent => ignore
{A,C} = 3; not closed due to {A,C,E}
{A,D} = 3; not closed due to {A,D,E}
{A,E} = 4; closed, but not maximal due to {A,D,E}
{B,C} = 2; not frequent => ignore
{B,D} = 0; not frequent => ignore
{B,E} = 2; not frequent => ignore
{C,D} = 3; not closed due to {C,D,E}
{C,E} = 5; closed, but not maximal due to {C,D,E}
{D,E} = 4; closed, but not maximal due to {A,D,E}

{A,B,C} = 1; not frequent => ignore
{A,B,D} = 0; not frequent => ignore
{A,B,E} = 1; not frequent => ignore
{A,C,D} = 2; not frequent => ignore
{A,C,E} = 3; maximal frequent
{A,D,E} = 3; maximal frequent
{B,C,D} = 0; not frequent => ignore
{B,C,E} = 2; not frequent => ignore
{C,D,E} = 3; maximal frequent

{A,B,C,D} = 0; not frequent => ignore
{A,B,C,E} = 1; not frequent => ignore
{B,C,D,E} = 0; not frequent => ignore
```

Answer to problem:
Frequent item set X∈F is **maximal** if it does not have any frequent supersets.
Frequent item set X ∈ F is **closed** if it has no superset with the same frequency

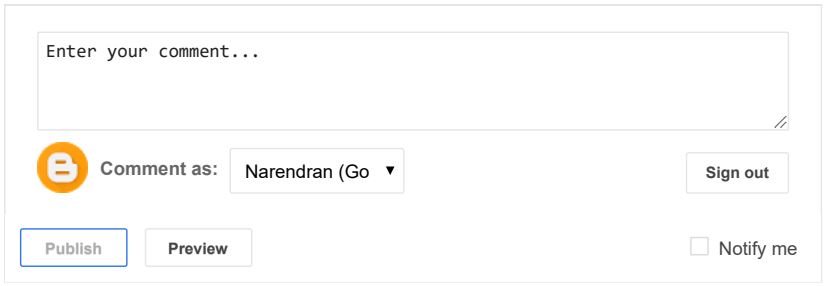
Closed 2-itemsets: "ab, bc, bd"
Maximal 2-itemsets: "bd"

*****

Tag: BITS WILP Data Mining Mid-Sem Exam 2017-H2

No comments:

Post a Comment



Home

Subscribe to: Posts (Atom)

Followers

Followers (0)

Follow