Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2019-2020
M.Tech (Data Science and Engineering)
Mid-Semester Exam (EC-2 Regular)

Course No. : DSECLZC415
Course Title : Data Mining
Nature of Exam : Open Book
Weightage : 30%
Duration : 90 minutes
Date of Exam : 21/06/2020 (AN), 2:00 pm to 3:30 pm

No. of Pages = 3
No. of Questions = 4

Note to Students:
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. You have been given a task to perform the data preprocessing of the data retrieved from multiple sources, before you start applying the data mining task. Identify, (atleast 5) data quality issues with the sample data set retrieved from the master data set. Suggest, how do you resolve these quality issues (python code is not required)? **[5]**

TXN-ID	NAME	AGE	HEIGHT	WEIGHT	BLOOD GROUP	COVID-19 RESULT
T001	RAMA	45	145	62kg	O+ve	Positive
T002	SEETHA	43	168	45kg	B+ve	Negative
T003	Akbar	38	172	60kg	Iam+ve	Positive
T004	BIRBAL	45	168	52kg	AB+ve	Negative
T005	THenali	22	157	78kg	B-ve	1
T006	Venkat	36	157	54kg	O-ve	Negative
T007	Rajuu	350	132	48kg	O+ve	Positive
T008	HARI	32	180	120lbs	AB-ve	Negative
T009	Inba	25		85kg	O+ve	0
T010	SysUsr789	20	165	68kg	O-ve	Negative

The attribute value SysUsr789 for the Name in the given data (T010 record) is not consistent with other names and it has alpha numeric when compared with other data types. (0.5)
This data quality issue can be resolved by replacing that field with right name/data type for consistency. (0.5)
The Age 350 is the outlier in T007 record and Height for Inba is missing (T009) (0.5)
These data issues can be resolved by filling the mean value of age and height. (0.5)
There is a mismatch in the data type units in T008, the Weight Unit for Hari is 120lbs whereas all other attributes are having Kg values. (0.5)
This is the data type issue and it can be done through data transformation by either manual or automatic edits of erroneous data (0.5)
The blood group has different representation in T004 record, inconsistent format of Iam+ve is being used in the blood group. (0.5)
This can be replaced with either NULL or by applying binning techniques (0.5)

Transaction id T005 has Covid Result-Representation Mismatch as 1 and in T009 it has 0, instead of indicating positive and negative values (1).
This data quality issue can be solved by applying data transformation such as data smoothing to make the simple changes as there are only two values which requires replacement. (0.5)

Q.2. Answer the following:

a) Find Minkowski distance of order = 3, between two objects represented by the coordinates (12, 36, 42, 20) and (17,35, 43, 26). **[2]**

((|12-17|)**3 + (|36-35|)**3 + (|42-43|)**3 + (|20-26|)**3)**(1/3) = 7

1 mark for the formula

1 mark for the calculations and correct answer

b) Apply equi-width and equi-depth binning method on following dataset to create sets of 3 bins. [23, 8, 2, 20, 11, 1, 29, 30, 21] **[3]**

For equi-width, bin width will be (30-1)/3=9.66, so ranges will be 1-10.66, 10.66-20.33, 20.33-30 [0.5 mark]
Equi-width bins : [1,2,8], [11,20], [21, 23, 29, 30], [1mark]
For equi-depth, each bin will have 9/3 = 3 elements [0.5]
Equi-depth bins : [1,2,8], [11,20,21], [23,29,30] [1 mark]

c) Suppose the stock closing price of two companies A and B are as follows **[4]**

Company	Mon	Tues	Wed	Thurs	Fri
A	100	110	105	100	95
B	150	148	155	155	100

What inference can you draw from this dataset on the dependency of stock prices in companies A and B? Explain how you arrive at this.

The Pearson correlation coefficient is used to measure the strength of a linear association between two variables, where the value $r = 1$ means a perfect positive correlation and the value $r = -1$ means a perfect negataive correlation.

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{\sqrt{\Sigma(X-\bar{X})^2}\sqrt{(Y-\bar{Y})^2}}$$

Where, $\bar{X}$ - mean of X variable
$\bar{Y}$ - mean of Y variable

a

A	B	Ai - Mean(A)	Bi - Mean(B)	(Ai - Mean(A))X (Bi-mean(B)	(((Ai-mean(A))^2)	(((Bi-mean(B))^2)
100	150	-2	8.4	-16.8	4	70.56
110	148	8	6.4	51.2	64	40.96
105	155	3	13.4	40.2	9	179.56
100	155	-2	13.4	-26.8	4	179.56
95	100	-7	-41.6	291.2	49	1730.56

r = 339/534.9355101 = 0.63

A and B are positively correlated

Second Method

Covariance is used to measure of the relationship between two random variables, to check whether the variables are positively related or inversely related.

$$\text{Cov}(X, Y) = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n}$$

Where,

- Xi – the values of the X-variable
- Yj – the values of the Y-variable
- $\bar{X}$ – the mean (average) of the X-variable
- $\bar{Y}$ – the mean (average) of the Y-variable
- n – the number of the data points

Positive covariance: Indicates that two variables tend to move in the same direction.
Negative covariance: Reveals that two variables tend to move in inverse directions.

Cov(A,B) = 339/5= 67.8

Hence, A and B are positively corelated.

1 mark for identifying the measure to be used
1 mark for the formula
2 marks for calculation and correct interpretation

Q.3. Answer the following: **[5+3+3]**

a) Consider the following training data set (with three attributes, such as Past trend, Open Interest, Trading Volume and class/target variable is "return") for a binary class problem. The attributes are nominal with two possible values. We intend to create a decision tree model using Information Gain. Which attribute would the decision tree induction algorithm choose for the root node?

Past Trend	Open Interest	Trading Volume	Return
Positive	Low	High	Up
Negative	High	Low	Down
Positive	Low	High	Up
Positive	High	High	Up
Negative	Low	High	Down
Positive	Low	Low	Down
Negative	High	High	Down
Negative	Low	High	Down
Positive	Low	Low	Down
Positive	High	High	Up

Entropy for the entire dataset [1mark]
Up=4
Down=6
Total =10
Entropy =$-6/10\log_2 6/10 - 4/10\log_2 4/10 = 0.97$
Past trend [1 mark]
P(Past Trend=Positive): 6/10
P(Past Trend=Negative): 4/10
If (Past Trend = Positive & Return = Up), probability = 4/6
If (Past Trend = Positive & Return = Down), probability = 2/6
If (Past Trend = Negative & Return = Up), probability = 0
If (Past Trend = Negative & Return = Down), probability = 4/4
Information gain for past trend
Information gain for past trend –positive
4/6log4/6-2/6log2/6=0.92
Information gain for past trend –negative
0
Gain =0.97-0.6*0.92-0=0.418

Open Interest[1 mark]
P(Open Interest=High): 4/10
P(Open Interest=Low): 6/10
If (Open Interest = High & Return = Up), probability = 2/4
If (Open Interest = High & Return = Down), probability = 2/4
If (Open Interest = Low & Return = Up), probability = 2/6
If (Open Interest = Low & Return = Down), probability = 4/6
Information gain for open interest –high
-2/4log2/4-2/4log2/4=1
Information gain for open interest –low
-2/6log2/6-4/6log4/6
Gain for Open Interest = 0.97-(0.4+0.6*0.918)=0.192

Trading Volume[1 mark]
P(Trading Volume=High): 7/10
P(Trading Volume=Low): 3/10
If (Trading Volume = High & Return = Up), probability = 4/7
If (Trading Volume = High & Return = Down), probability = 3/7
If (Trading Volume = Low & Return = Up), probability = 0
If (Trading Volume = Low & Return = Down), probability = 3/3
Information gain for trading volume –high
-4/7log4/7-3/7log3/7=0.98522
Information gain for trading volume-low
0-3/3log3/3=0

Gain for trading volume =0.97-.7*.98=0.280

Since the Gain for "past trends" is the highest, hence, it will be selected as the root node [1 mark]

b) Suppose a training set consists of 100 positive examples and 100 negative examples for each of the rules R1 and R2. Considering R1 covers 30 positive examples and 90 negative examples; R2 covers 60 positive examples and 60 negative examples, Identify which rule is better, using FOIL gain.

Foil_Gain(R2, R0)= p2 [log p2/(p2+n2) -log (p0/(p0+n0)]

= 60 [log 60/(60+60) -log (100/(100+200)]
=60 [log 60/(120) -log (100/(200)]
=0

1M

Foil_Gain(R1, R0)= p_1 [log $p_1/(p_1+n_1)$ -log (p0/(p0+n0)]

= 30 [log (30/(30+90) - log 100/(200)] = -30

1M

Clearly R2 is better compared to R1 1M

c) Consider the confusion matrices for two models M1 and M2 are given below. Evaluate the performance of models using F1-score, and Identify which one is better.

M1:

Predicted Class → Actual Class ↓	Positive	Negative	
Positive	1800	200	2000
Negative	400	1600	2000
	2200	1800	4000

M2:

Predicted Class → Actual Class ↓	Positive	Negative	
Positive	1600	400	2000
Negative	800	1200	2000
	2400	1600	4000

F1 =2 *precision *recall/(precision + recall)
precision=TP/(TP+FP); recall=TP/(TP+FN)
Model1: precision=1800/2200=9/11; recall =1800/(2000)=9/10;
F1(M1)= (2*9/11*9/10)/[9/11+9/10]= 6/7= 0.85 1 M

Precision (M2) =1600/2400 =2/3; Recall(M2)=1600/2000 =4/5
F1(M2)= (2*2/3*4/5)/[2/3 +4/5]= 0.72 1 M
Since F1 score of M1 is better than that of M2, M1 is a better model 1 M

Q.4. Answer the following: **[1+4]**

a) Consider the following data describing three customers (A, B, C) and their preferences for four

products P1, P2, P3, P4 where "1" indicates the customer prefers that product. Which similarity measure is appropriate in this case to measure the similarity between customers?

b) Identify the customer pairs that are more similar with respect to the rest of them by computing the similarity measure.

	P1	**P2**	**P3**	**P4**
A	0	1	1	0
B	1	1	0	0
C	1	1	0	1

a) Since the bit vectors are asymmetric binary in the given case Jaccard would be better measure. One can also use cosine similarity measure: 1M

b) J(A,B) =f11/(f10+f01+f11) =1/[3]=1/3=0.33 1M

J(B,C) =2/(2+1)=2/3 = 0.66; 1M
J(A,C) =1/(2+1+1) =1/4=0.25 1M
B, C are more similar 1M

OR
If one uses cosine similarity
s(A,B)=1/2=0.5; s(B,C) =2/(√ 6)=2/2.45=0.816; s(A,C)=1/√6=0.41
B, C are more similar