AUTUMN MID SEMESTER EXAMINATION-2019
School of Computer Engineering
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
DEEMED TO BE UNIVERSITY, BHUBANESWAR-24

DATA WAREHOUSING AND DATA MINING
[CS-6301]

Time: 1½ Hours **Full Mark: 20**

Answer any four questions including question No.1 which is compulsory.
The figures in the margin indicate full marks. Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

Q.1. [5x1]

(a) How does an ordinal feature differ from a nominal feature?

Ordinal data, unlike nominal data, involves some order; ordinal numbers stand in relation to each other in a ranked fashion. Values have a meaningful order (ranking) but magnitude between successive values is not known.
Nominal means "relating to names." The values of a nominal attribute are symbols or names of things. Nominal data simply names something without assigning it to an order in relation to other numbered objects or pieces of data.

(b) Define anti-monotone property.

If an itemset is frequent, each of its subsets is frequent as well.

(c) Differentiate data query and knowledge query

Ans: View query, pattern query

(d) What condition two item sets A and B will have "No Correlation between them"?

Ans: corr(A,B) =1 means that A and B are independent and there is no correlation between them.

(e) Compute the Euclidean distance between the two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8).

Ans: 6.7

Q.2.

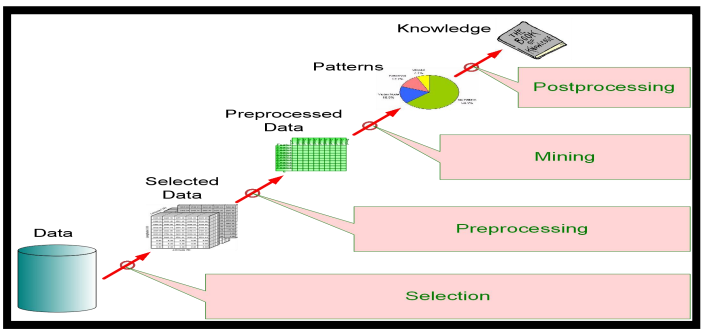
(a) What is Bessel's correction? Calculate and Draw a box plot for the data set given below, [3]

3, 3, 7, 8, 7, 4, 4, 10, 1, 5, 1, 7, 2, 7, 9

Ans: Q1=3, Q2=5, Q3=7, Q3-Q1=4 [2]

(b) With a neat diagram explain the architecture of data mining.

Ans:



Q.3.

You are given the transaction data shown in the Table below from a fast food restaurant. There are 9 distinct transactions (order:1 – order:9) and each transaction involves between 2 and 4 meal items. There are a total of 5 meal items that are involved in the transactions. For all of the parts below the ***minimum support is 2/9*** and the ***minimum confidence is 7/9.*** [5]

Meal Item	List of Item IDs	Meal Item	List of Item IDs
Order:1	M1, M2, M5	Order:6	M2, M3
Order:2	M2, M4	Order:7	M1, M3
Order:3	M2, M3	Order:8	M1, M2, M3, M5
Order:4	M1, M2, M4	Order:9	M1, M2, M3
Order:5	M1, M3		

Find all strong association rules of the form: **X ∧ Y→ Z** and note their confidence values.

Ans:

C1/L1

Itemset	Support Count
M1	6
M2	7
M3	6
M4	2
M5	2

C2/L2

Itemset	Support Count
{M1, M2}	4
{M1, M3}	4
~~{M1, M4}~~	~~1~~
{M1, M5}	2
{M2, M3}	4
{M2, M4}	2
{M2, M5}	2
~~{M3, M4}~~	~~0~~
~~{M3, M5}~~	~~1~~
~~{M4, M5}~~	~~0~~

L2 (after pruning)

Itemset	Support Count
{M1, M2}	4
{M1, M3}	4
{M1, M5}	2
{M2, M3}	4
{M2, M4}	2
{M2, M5}	2

C3 initial

Itemset	Support Count
{M1, M2, M3}	
{M1, M2, M5}	
{M1, M3, M5}	
{M2, M3, M4}	
{M2, M3, M5}	
{M2, M4, M5}	

C3 after checking for Apriori Property

Itemset	Comment
{M1, M2, M3}	
{M1, M2, M5}	
~~{M1, M3, M5}~~	No {M3, M5}
~~{M2, M3, M4}~~	No {M3, M4}
~~{M2, M3, M5}~~	No {M3, M5}
~~{M2, M4, M5}~~	No {M4, M5}

C3 Final/L3

Itemset	Support Count
{M1, M2, M3}	2
{M1, M2, M5}	2

C4/L4

Itemset	Support Count
~~{M1, M2, I3, M5}~~	{No M3, M5}

Solutions Part b)

Rule	Confidence
~~M1 ∧ M2 → M3~~	~~2/4 = .50~~
~~M2 ∧ M3 → M1~~	~~2/4 = .50~~
~~M1 ∧ M3 → M2~~	~~2/4 = .50~~
~~M1 ∧ M2 → M5~~	~~2/4 =.5~~
M1 ∧ M5 → M2	2/2 = 1.0
M2 ∧ M5 → M1	2/2 = 1.0

Q.4.

Use the dataset below to learn and draw a decision tree which predicts if student pass in Machine Learning subject (Yes or No), based on their previous GPA (High, Medium, or Low) and whether or not they studied. [5]

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

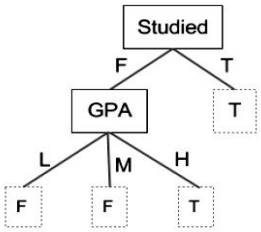
Ans:

$$H(Passed) = -(\frac{2}{6}\log_2\frac{2}{6} + \frac{4}{6}\log_2\frac{4}{6})$$

$$H(Passed) = -(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3})$$

$$H(Passed) = \boxed{\log_2 3 - \frac{2}{3} \approx 0.92}$$

[2.5]



[2.5]

$$H(Passed|GPA) = -\frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) - \frac{1}{3}(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}) - \frac{1}{3}(1\log_2 1)$$

$$H(Passed|GPA) = \frac{1}{3}(1) + \frac{1}{3}(1) + \frac{1}{3}(0)$$

$$H(Passed|GPA) = \boxed{\frac{2}{3} \approx 0.66}$$

$$H(Passed|Studied) = -\frac{1}{2}(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}) - \frac{1}{2}(1\log_2 1)$$

$$H(Passed|Studied) = \frac{1}{2}(\log_2 3 - \frac{2}{3}$$

$$H(Passed|Studied) = \boxed{\frac{1}{2}\log_2 3 - \frac{1}{3} \approx 0.46}$$

Q.5.
Following data shows the number of customers with their corresponding temperature.

Temperature(X)	No. of Customers(Y)
97	14
86	11
89	9
84	9
94	15
74	7

(a) Calculate the covariance between the temperature and customers

Ans:

Mean of X, $\bar{x}$ = (97+86+89+84+94+74)/6 = 524/6= 87.333

Mean of Y, $\bar{Y}$ = (14+11+9+9+15+7)/6 = 65/6= 10.833

Temperature (x- $\bar{X}$)	No of Customers (y- $\bar{Y}$)	Product (x- $\bar{X}$)(y-$\bar{Y}$)
97-87.33 = 9.67	14-10.83 = 3.17	30.65
86-87.33 = -1.33	11-10.83 = 0.17	-0.22
89-87.33 = 1.67	9-10.83 = -1.83	-3.05
84-87.33 = -3.33	9-10.83 = -1.83	6.09
94-87.33 = 6.67	15-10.83 = 4.17	27.81
74-87.33 = -13.33	7-10.83 = -3.83	51.05

$COV(x, y)$ = 112.33/(6–1) = 112.33/5 = 22.46

(b) Show the strength of the correlation between temperature and number of customers.

$$Correlation = \frac{Cov\ (x,y)}{\sigma x * \sigma y}$$

$COV(x, y)$ = covariance of the variables x and y

σx = sample standard deviation of variable x

σy = sample standard deviation of variable y

COV(x, y) = 22.46

σx = 331.28/5=66.25= 8.13

σy = 48.78/5=9.75=3.1

correlation = 22.46/(8.13x 3.1)= 22.46/25.20 =0.8

****