Vedant Sahai
Batch - C
Roll No 8364

FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERING
Department of Computer Engineering
ASSIGNMENT 2
(2019-2020)

Class/Sem./Branch –TE/VI/COMP **Course code: CSC603**
Subject: Data warehousing and mining (DWM) **Date: 01/02/2020**
DATE OF SUBMISSION: 11-02-2020

Course outcomes: On successful completion of course learner will be able to:

CSC603.3	**Identify appropriate techniques / algorithms to solve real world problems of data exploration in data mining.**
Improvements	**Q 12 and 13 is added for metadata of dimension and normalization**

Exercise 1

Given the following points compute the distance matrix by using

a) Manhattan distance (provide the formula)
b) Euclidean distance (provide the formula)
c) Supremum distance (provide the formula)

Points	X	Y
P1	6	3
P2	2	2
P3	3	4

Exercise 2

Given the following table compute the correlation matrix.

AGE	INCOME	EDUCATION	HEIGHT
10	0	4	130
20	15000	13	180
28	20000	13	160
35	40000	18	150
40	38000	13	170

Exercise 3

Given the following two vectors compute the cosine similarity

D1= 4 0 2 0 1
D2= 2 0 0 2 2

Exercise 4

Given the following two binary vectors compute the Jaccard and Simple Matching Coefficient:

p = 0 0 1 1 0 1
q = 1 1 1 1 0 1

Exercise 5

Apply discretization on the attribute AGE and provide the corresponding histogram by using: a) Natural Binning with number of classes K=5 and b) Equal-frequency binning with number of classes K=3.

AGE: 10,10,15,28,30,20,80,60,30,35,70,5

Exercise 6

Suppose that the minimum and maximum values for the attribute *income* are $12,000 and $98,000, respectively. We would like to map *income* to the range [0.0, 1.0]. Find min-max normalization, for the value of $73,600 *income*

Exercise 7

Explain attribute types with examples and operations in table format for college system. (Consider students and employee)

Exercise 8

Draw the box plot, Histogram for the data in exercise 5

Exercise 9

Give the cosine similarity for following lines using binary data formula and nominal data formula and comment on the answer.

L1= I like the data mining than DBMS

L2= Raj loves data mining than DBMS

Exercise 10

Find the dissimilarity among the objects for following data

Name	Gender	Fever	cough	Test1	Test2	Test3	Test4
Jack	M	Y	N	P	N	N	N
Marry	F	Y	N	P	N	P	N
JIM	M	Y	Y	N	N	N	N

Exercise 12: Suppose that a data warehouse for *Big University* consists of the four dimensions *student, course, semester*, and *instructor*, and two measures *count* and *avg grade*. At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg grade* measure stores the actual course grade of the student. At higher conceptual levels, *avg grade* stores the average grade for the given combination.

Give the metadata for each dimension.

Exercise 13: Use the two methods below to *normalize* the following group of data:

200*;* 300*;* 400*;* 600*;* 1000

(a) min-max normalization by setting *min* = 0 and *max* = 1

(b) z-score normalization

ASSIGNMENT-2

— SOLUTIONS —

Q1.3 a.) The Manhattan distance is obtained setting r=1 in the Minkowski distance

$$dist = \left(\sum_{k=1}^{n} |p_k - q_k|^r \right)^{\frac{1}{r}}$$

L1	P1	P2	P3
P1	0	5	4
P2	5	0	3
P3	4	3	0

b.) the Euclidean distance is obtained setting r=2 in the Minkowski distance

L2	P1	P2	P3
P1	0.000	4.123	3.162
P2	4.123	0.000	2.236
P3	3.162	2.236	0.000

c.) The Euclidean distance is obtained setting r=∞ is the Minkowski distance.

L1	P1	P2	P3
P1	0.000	4.000	3.000
P2	4.000	0.000	2.000
P3	3.000	2.000	0.000

Q2) Avg age = 26.6
Std age = 11.9498954
Avg Edu = 12.2
Std Edu = 5.069516742
Avg income = 22600
Std income = 16697.30517
Avg Height = 158
Std Height = 19.23538406

Age - Avg	Income - Avg	Edu - Avg	Height - Avg
-16.6	-22600	-8.2	-28
-6.6	-7600	0.8	22
1.4	-2600	0.8	2
8.4	17400	5.8	-8
13.4	15400	0.8	12

$$Corr(Age, Income) = \frac{((-16.6 * -22600) + (-6.6 * -7600) + (1.4 * -2600) + (8.4 * 17400) + (13.4 * 15400))}{4 * 11.9498954 * 16697.30517} = 0.97$$

...

Correlation	Age	Income	Education	Height
Age	1.00	0.97	0.79	0.45
Income	0.97	1.00	0.86	0.39
Education	0.79	0.86	1.00	0.54
Height	0.45	0.39	0.54	1.00

Q3.) D1 = 4 0 2 0 1 ; D2 = 2 0 0 2 2.

$D1 \cdot D2 = 4*2 + 0*0 + 2*0 + 0*2 + 1*2$

$= 10$

$||D11|| = [4^2 + 2^2 + 1^2]^{0.5} = 4.58$

$||D21|| = [2^2 + 2^2 + 2^2]^{0.5} = 3.46$

$\cos(D1, D2) = (D1 \cdot D2) / (||D11|| * ||D21||)$

$= 10 / (4.58 * 3.46) = 0.63$

Q4.) p = 0 0 1 1 0 1 ; q = 1 1 1 1 0 1

M_{01} = 2 (the no. of attributes where p was 0 and q was 1)

M_{10} = 0 (" " " " " " 1 " " 0)

M_{00} = 0 (" " " " " " 0 " " 0)

M_{11} = 3 (" " " " " " 1 " " 1)

$SMC = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

$= (3+1) / (2+0+3+1) = 0.67$

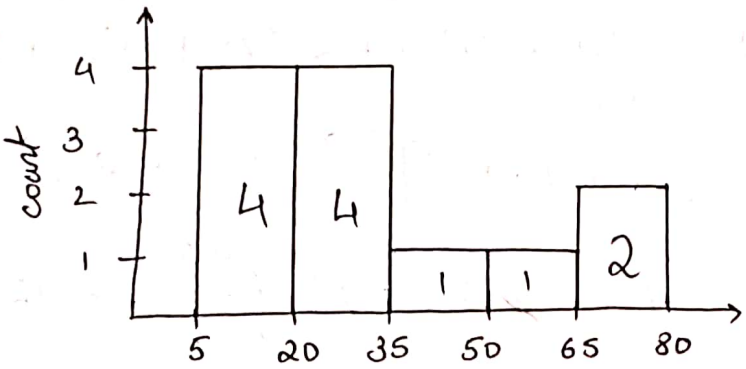
$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0.6$

Q5.) Age: 10, 10, 15, 28, 30, 20, 80, 60, 30, 35, 70.5

a.) Natural Binning with number of classes k = 5

delta = (max - min) / k = (80 - 5) / 5 = 15

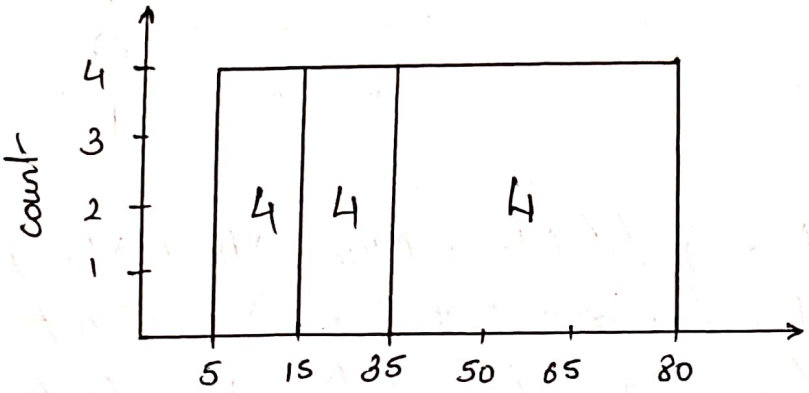
C1: [5, 20] ; C2: [20, 35] ; C3: [35, 50]; C4: [50, 65]
C5 : [65, 80]



b.) Equal frequency binning with number of classes k = 3

F = N/K = 12/3 = 4

C1: {5, 10, 10, 15}
C2: {20, 28, 30, 30}
C3: {35, 60, 70, 80}



Q6.3 Min-max normalization = $\frac{v - min_F}{max_F - min_F}(new\_max_F - new\_min_F) + new\_min_F$

where v is the current value of the feature F according to the question

$v = 73600$; $min_F = 12000$; $max_F = 98000$

$new\_max_F = 1.0$; $new\_min_F = 0$

$\therefore$ Min-max normalization = $\frac{73600 - 12000}{98000 - 12000}(1.0 - 0.0) + 0$

$= 0.716$

Q7.3

ATTRIBUTE TYPE	DESCRIPTION	EXAMPLES	OPERATION
Nominal	the values of a nominal attribute are just different names i.e; nominal attributes provide only enough informatn to distinguish one object from another	zip codes, student, teacher ID numbers, sex	mode entropy, contingency, correlation, χ^2 test.
Ordinal	the values of an ordinal attribute provide enough informatn to order objects	grades, age, Street-numbers	median, percentiles, rank correlation, run tests, sign tests

Interval	For interval attributes, the differences between values are meaningful, i.e. a unit of measurement exists	calendar dates, classes, divisions	mean, standard deviation, Pearson's correlation, t & F tests
Ratio	For ratio variables, both differences and ratios are meaningful	age, height, weight, monetary quantities	geometric mean, harmonic mean, percent variation

Q9.] L1: I like the data mining than DBMS; L2: Raj loves data mining than DBMS

$L1 \cup L2 = \{0: I;\ 1: like,\ 2: love,\ 3: Data,\ 4: mining,\ 5: DBMS,\ 6: Raj\}$

Binary data format:— $L1 = \{1101110\}$; $L2 = [0011111]$

$D_1 * D_2 = 3$

$|D_1| = \sqrt{5} = |D_2|$

$\therefore \cos\theta = \frac{D_1 * D_2}{|D_1||D_2|} = \frac{3}{5} = \underline{0.6}$

Normal Data format:—

$D_1 = [0.45, 0.45, 0, 0.45, 0.45, 0]$

$D_2 = [0, 0, 0.45, 0.45, 0.45, 0.45, 0.45]$

$D_1 D_2 = 0.1075$

$|D_1| = |D_2| = \sqrt{1.0125}$

$\therefore \cos\theta = \frac{D_1 \cdot D_2}{|D_1||D_2|} = \frac{0.1075}{1.0125}$

$\therefore \cos\theta = 0.6$

Q10.) Gender is a symmetric attribute
The remaining attributes are asymmetric binary
Let the values Y & P be 1 and the value N is 0

$$\therefore d(\text{jack, mary}) = \frac{0+1}{2+0+1} = \underline{0.33}$$

$$\therefore d(\text{jack, jim}) = \frac{1+1}{1+1+1} = \underline{0.67}$$

$$\therefore d(\text{mary, jim}) = \frac{1+2}{1+1+2} = \underline{0.75}$$

Q12.) Student < student_id, student_name, address_id, major, status, university >
Course < course_id, course_name, department >
Instructor < instructor_id, inst_name, department >
Semester < semester_id, semester name, year >
Address < address_id, street, city, state, zip_code, country >

Q13.) a.) min-max normalization by setting min = 0 & max = 10

$min_f = 200$ & $max_f = 1000$

$$\therefore \text{min-max} = \frac{v - min_f}{max_f - min_f}(max - min) + min$$

for $v = 200$; min-max = 0 ; for $v = 600$; min-max = 0.5
for $v = 300$; min-max = 0.125; for $v = 100$; min-max = 1
for $v = 400$; min-max = 0.25

b.) z-score normalization using the mean absolute deviation instead of the standard deviation

we have mean $M = \frac{1}{n} \sum_{i=1}^{n} x_i = 500$

$$MAP = \frac{1}{n} \sum_{i=1}^{n} (x_i - M) = 240$$

$$\therefore z = \frac{v - \bar{A}}{\sigma_A} \quad ; v \text{ is old entry, } \bar{A} \text{ is mean & } \sigma_A \text{ is std deviation}$$

for $v = 200$; $\therefore z = -1.25$
for $v = 300$; $\therefore z = -0.833$
for $v = 400$; $\therefore z = -0.417$
for $v = 600$; $\therefore z = 0.417$
for $v = 1000$; $\therefore z = 2.083$