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

Solution:

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=inf in the Minkowski distance

Linf	P1	P2	P3
P1	0.000	4.000	3.000
P2	4.000	0.000	2.000
P3	3.000	2.000	0.000

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

Solution:

AVG AGE:	26.6
STD AGE	11.9498954
AVG INCOME	22600
STD INCOME	16697.30517
AVG EDU	12.2
STD EDU	5.069516742
AVG EDU	158
STD EDU	19.23538406

AGE-AVG	INCOME-AVG	EDU-AVG	HEIGTH-AVG
-16.6	-22600.00	-8.2	-28
-6.6	-7600.00	0.8	22
1.4	-2600.00	0.8	2
8.4	17400.00	5.8	-8
13.4	15400.00	0.8	12

Corr(Age,Icome)= ((-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

Exercise 3

Given the following two vectors compute the cosine similarity

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

Solution

D1 • D2 = 4*2 + 0*0+ 2*0 + 0*2 + 1*2 = 10
||D1|| $=(4^2 + 2^2 + 1^2)^{0.5} = (16+4+1)^{0.5} = 21^{0.5} = 4.58$
||D2|| $=(2^2 + 2^2 + 2^2)^{0.5} = (4+4+4)^{0.5} = 12^{0.5} = 3.46$

COS (D1,D2) = (D1 • D2)/ (||D1|| * ||D2||) = 10/(4.58*3.46) = 0.63

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

Solution

$\mathbf{M_{01}}$ = 2 (the number of attributes where p was 0 and q was 1)
$\mathbf{M_{10}}$ = 0 (the number of attributes where p was 1 and q was 0)
$\mathbf{M_{00}}$ = 1 (the number of attributes where p was 0 and q was 0)
$\mathbf{M_{11}}$ = 3 (the number of attributes where p was 1 and q was 1)

SMC = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$ = (3+1) / (2+0+3+1) = 4/6 = 0.67

J = (M_{11}) / $(M_{01} + M_{10} + M_{11})$ = 03/ (2 + 3) = 3/5 = 0.6

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.

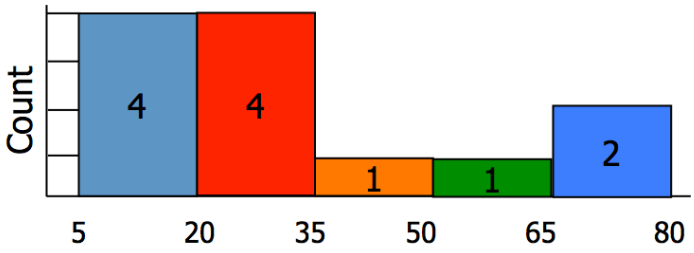
AGE: 10,10,15,28,30,20,80,60,30,35,70,5

SOLUTION

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)
C: [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}