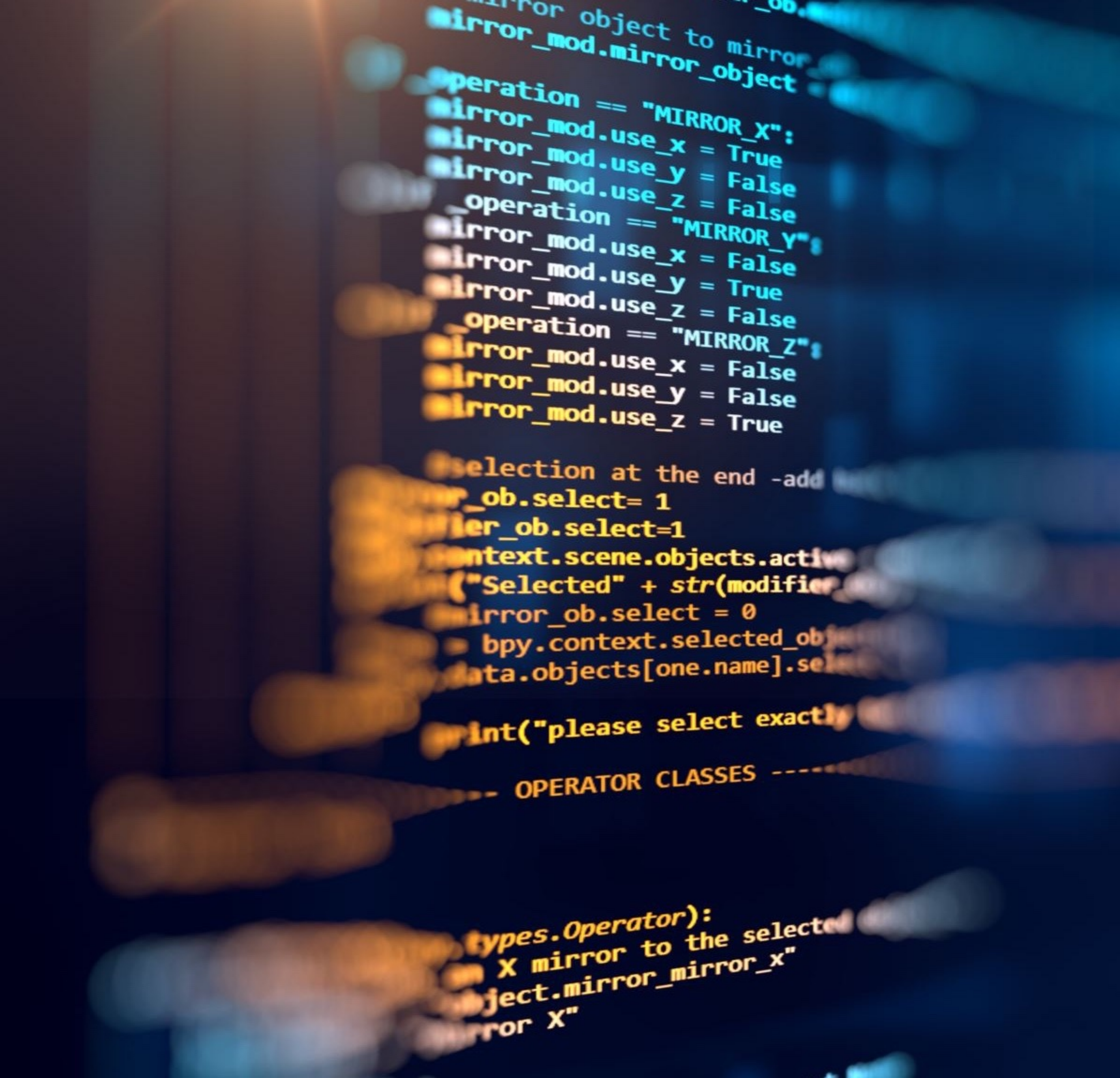

DATA SCIENCE

SIKKIM UNIVERSITY
MCA 3RD SEMESTER

UNIT –II (PROXIMITY MEASURE)

CONTENT

- PROXIMITY MEASURE
- IDENTITY
- NON-NEGATIVITY
- SYMMETRICITY
- TYPES OF PROXIMITY MEASURES

PROXIMITY MEASURE

- **Proximity measures are mainly mathematical techniques that calculate the similarity/dissimilarity of data points**.
- Usually, proximity is measured in terms of similarity or dissimilarity i.e., how alike objects are to one another.
- While implementing clustering algorithms ,outlier analysis and nearest neighbour, it is important to be able to quantify the proximity of objects to one another.
- Distance or similarity measures are essential in solving many pattern recognition problems such as classification and clustering.
- Various distance/similarity measures are available in the literature to compare two data distributions.
- As the names suggest, a similarity measures how close two distributions are.
- For multivariate data, complex summary methods are developed to answer this question.
- **Proximity measures are different for different types of attributes.**

SIMILARITY AND DISSIMILARITY

- **Similarity Measure** - Numerical measure of how alike two data objects often fall between 0 (no similarity) and 1 (complete similarity).

- Measure is higher when objects are more alike.
- Often falls in the range [0,1].

- **Dissimilarity Measure** -Numerical measure of how different two data objects are range from 0 (objects are alike) to ∞ (objects are different). Numerical measure of how different two data objects are.

– Lower when objects are more alike.

– Minimum dissimilarity is often 0.

– Upper limit varies.

- **Proximity** - refers to a similarity or dissimilarity

DISSIMILARITY MATRIX

- Dissimilarity matrix is a matrix of pairwise dissimilarity among the data points.
- It is often desirable to keep only lower triangle or upper triangle of a dissimilarity matrix to reduce the space and time complexity.

- ***It's square and symmetric(**A^T**= A for a square matrix A, where** A^T **represents its transpose).***
- ***The diagonals members are zero, meaning that zero is the measure of dissimilarity between an element and itself.***

MEASURES OF SIMILARITY AND DISSIMILARITY

- **Distance**, such as the Euclidean distance and Minkowski distance, is a dissimilarity measure and has some well-known properties:
- Common Properties of Dissimilarity Measures :-

1. $d(p, q) \geq 0$ for all p and q, and $d(p, q) = 0$ if and only if $p = q$,
2. $d(p, q) = d(q,p)$ for all p and q,
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all p, q, and r, where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q.

- A distance that satisfies these properties is called a **metric**.

COMMON PROPERTIES OF SIMILARITY MEASURES

Similarities have some well-known properties:

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$,

2. $s(p, q) = s(q, p)$ for all p and q, where $s(p, q)$ is the similarity between data objects, p and q.

EXAMPLE

- Suppose we have four objects A,B,C,D and need to find proximity measure, then we first create a **dissimilarity or similarity** matrix.

DISMILIARITY MATRIX

	A	B	C	D
A	0	d(A,B)	d(A,C)	d(A,C)
B	d(B,A)	0	d(B,A)	d(B,D)
C	d(C,A)	d(C,B)	0	d(C,D)
D	d(D,A)	d(D,B)	d(D,C)	0

	A	B	C	D
A	0			
B	d(B,A)	0		
C	d(C,A)	d(C,B)	0	
D	d(D,A)	d(D,B)	d(D,C)	0

Note: ‘d’ refers to distance metric.

DISTANCE METRIC

- Distance metric, metric, or distance function, “is a function that defines a distance between each pair of elements of a set.”
- A distance metric d(·)d(·) requires the following four axioms to be true for all elements x, y, and z in a given set.

1. Non-negativity:

$d(x,y) \geq 0$ - The distance must always be greater than zero.

2. Identity of indiscernibles:

$d(x,y)=0 \Leftrightarrow x=y$ – The distance must be zero for two elements that are the same (i.e., indiscernible from each other).

3. Symmetry:

- $d(x,y)=d(y,x)$ – The distances must be the same, no matter which order the parameters are given.

4. Triangle inequality:

- $d(x,z) \leq d(x,y)+d(y,z)$ – For three elements in the set, the sum of the distances for any two pairs must be greater than the distance for the remaining pair.

SIMILARITY AND DISSIMILARITY BETWEEN SIMPLE ATTRIBUTES

Attribute Type	Similarity	Dissimilarity
Nominal	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$
Ordinal	$s = 1 - \frac{\lVert p - q \rVert}{n - 1}$	$d = \frac{\lVert p - q \rVert}{n - 1}$
	(values mapped to integer 0 to n-1, where n is the number of values)	
Interval or Ratio	$s = 1 - \lVert p - q \rVert, s = \frac{1}{1 + \lVert p - q \rVert}$	$d = \lVert p - q \rVert$

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

- Nominal attributes can have two or more different states e.g. an attribute ‘color’ can have values like ‘Red’, ‘Green’, ‘Yellow’, ‘Blue’, etc.
- **Dissimilarity for nominal attributes is calculated as the ratio of total number of mismatches between two data points to the total number of attributes.**
- Nominal means **“relating to names.”** The values of a nominal attribute are symbols or names of things.
- Each value represents some kind of category, code,
or state and so nominal attributes are also referred to as **categorical**.

Examples: ID numbers, eye color, zip codes.

- Let **M** be the total number of states of a nominal attribute.
- Then the states can be numbered from 1 to **M.**
- However, the numbering does not denote any kind of ordering and can not be used for any mathematical operations.

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

- NOMINAL DATA EXAMPLE

COLOUR CODE
RED
BLUE
GREEN
YELLOW

GRADE
A
B
C
D

CODE
CODE A
CODE B
CODE C
CODE D

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

DISSIMILARITY MEASUREMENT

$$d(i,j) = (p-m)/p$$

p = number of attribute
m = number of match between i and j.
d(i,j) = dissimilarity between i and j.

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

EXAMPLE 1

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

Number of attribute =1

ID	TEST RESULT
1	CODE A
2	CODE B
3	CODE C
4	CODE A

Number of attribute =2

ID	TEST 1 RESULT	TEST 2 RESULT
1	CODE A	CODE C
2	CODE B	CODE B
3	CODE C	CODE A
4	CODE A	CODE A

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

DATA

ID	TEST RESULT
1	CODE A
2	CODE B
3	CODE C
4	CODE A

DISSIMILARITY MATRIX

	1	2	3	4
1	0			
2	d(2,1)	0		
3			0	
4				0

For d(2,1)
p =1
m = 0

d(2,1) = (p-m)/p
=(1-0)/1
=1

DISSIMILARITY MATRIX

	1	2	3	4
1	0			
2	1	0		
3			0	
4				0

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

ID	TEST RESULT
1	CODE A
2	CODE B
3	CODE C
4	CODE A

	1	2	3	4
1	0			
2		0		
3			0	
4				0

DISSIMILARITY MATRIX

For d(3,1)
p =1
m = 0

d(3,1) = (p-m)/p
=(1-0)/1
=1

For d(4,1)
p =1
m = 1

d(4,1) = (p-m)/p
=(1-1)/1
=0

For d(3,2)
p =1
m = 0

d(3,2) = (p-m)/p
=(1-0)/0
=1

For d(4,2)
p =1
m = 0

d(4,2) = (p-m)/p
=(1-0)/1
=1

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

For d(4,3)
p =1
m = 0

d(4,3) = (p-m)/p
=(1-0)/0
=1

DISSIMILARITY MATRIX

	1	2	3	4
1	0			
2	1	0		
3	1	1	0	
4	0	1	1	0

Here all data are dissimilar except (4,1)

DATA

ID	TEST RESULT
1	CODE A
2	CODE B
3	CODE C
4	CODE A

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

EXAMPLE 2

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

DATA

ID	ATTRIBUTE 1	ATTRIBUTE 2
1	20	AA
2	40	BB
3	20	AA
4	30	CC

DISSIMILARITY MATRIX

	1	2	3	4
1	0			
2		0		
3			0	
4				0

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

ID	ATTRIBUTE 1	ATTRIBUTE 2
1	20	AA
2	40	BB
3	20	AA
4	30	CC

For d(2,1)
p =2
m = 0

d(2,1) = (p-m)/p
=(2-0)/2
=1

For d(3,1)
p =2
m = 2

d(3,1) = (p-m)/p
=(2-2)/2
=0

For d(3,2)
p =2
m = 0

d(3,2) = (p-m)/p
=(2-0)/2
=1

	1	2	3	4
1	0			
2		0		
3			0	
4				0

For d(4,1)
p =2
m = 0

d(4,1) = (p-m)/p
=(2-0)/2
=1

For d(4,2)
p =2
m = 0

d(4,2) = (p-m)/p
=(2-0)/2
=1

For d(4,3)
p =2
m = 0

d(4,3) = (p-m)/p
=(2-0)/2
=1

DISSIMILARITY MATRIX

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

ID	ATTRIBUTE 1	ATTRIBUTE 2
1	20	AA
2	40	BB
3	20	AA
4	30	CC

For d(2,1)
p =2
m = 0

d(2,1) = (p-m)/p
=(2-0)/2
=1

For d(3,1)
p =2
m = 2

d(3,1) = (p-m)/p
=(2-2)/2
=0

For d(3,2)
p =2
m = 0

d(3,2) = (p-m)/p
=(2-0)/2
=1

	1	2	3	4
1	0			
2	1	0		
3	0	1	0	
4	1	1	1	0

DISSIMILARITY MATRIX

For d(4,1)
p =2
m = 0

d(4,1) = (p-m)/p
=(2-0)/2
=1

For d(4,2)
p =2
m = 0

d(4,2) = (p-m)/p
=(2-0)/2
=1

For d(4,3)
p =2
m = 0

d(4,3) = (p-m)/p
=(2-0)/2
=1

PROXIMITY MEASURE FOR NOMINAL ATTRIBUTES

	1	2	3	4
1	0			
2	1	0		
3	0	1	0	
4	1	1	1	0

DISSIMILARITY MATRIX

Here all data are not matching except d(3,1)

DATA

ID	ATTRIBUTE 1	ATTRIBUTE 2
1	20	AA
2	40	BB
3	20	AA
4	30	CC

PROBLEM

Find proximity measures for following nominal Attributes

Roll No	Marks	Grades
1	96	A
2	87	B
3	83	B
4	96	A

Solution:

Applying the formula for finding the proximity of nominal attributes we get:

- d(1,1)= (p-m)/p = (2-2)/2 = 0
- d(2,1)= (p-m)/p = (2-0)/2 = 1
- d(3,1)= (p-m)/p = (2-2)/2 = 1
- d(4,1)= (p-m)/p = (2-2)/2 = 0
- d(4,3)= (p-m)/p = (2-0)/2 = 1
- d(2,2)= (p-m)/p = (2-2)/2 = 0
- d(3,2)= (p-m)/p = (2-1)/2 = 0.5
- d(4,2)= (p-m)/p = (2-0)/2 = 1
- d(3,3)= (p-m)/p = (2-2)/2 = 0
- d(4,4)= (p-m)/p = (2-2)/2 = 0

	1	2	3	4
1	0			
2	1	0		
3	1	0.5	0	
4	0	1	1	0

DISSIMILARITY MATRIX

PROXIMITY MEASURES FOR ORDINAL ATTRIBUTES

PROXIMITY MEASURES FOR ORDINAL ATTRIBUTES

- An ordinal attribute is an attribute whose possible values have a meaningful order or ranking among them, but the magnitude between successive values is not known. However, to do so, it is important to convert the states to numbers where each state of an ordinal attribute is assigned a number corresponding to the order of attribute values.

Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}.

- Since a number of states can be different for different ordinal attributes, it is therefore **required to scale the values to a common range,** e.g [0,1]. This can be done using the given formula,

$$z_{if}=(r_{if}-1)/(M_f-1)$$

- where M is a maximum number assigned to states and r is the rank(numeric value) of a particular object.

- The similarity can be calculated as:

$$s(i, j)=1-d(i, j)$$

EXAMPLE

Object ID	Attribute
1	High
2	Low
3	Medium
4	High

- In this example, we have four objects having ID from 1 to 4.
- Here for encoding our attribute column, we consider ***High=1, Medium=2, and Low=3***. And, the value of M_f=3(since there are three states available)
- Now, we normalize the ranking in the range of 0 to 1 using the above formula.
- So, ***High=(1-1)/(3-1)=0, Medium=(2-1)/(3-1)=0.5, Low=(3-1)/(3-1)=1.***

- Finally, we are able to calculate the dissimilarity based on difference in normalized values corresponding to that attribute.
- – d(1,1)= 0-0 = 0
- – d(2,1)= 1-0= 1
- – d(3,1)= 0.5-0 = 0.5
- – d(4,1)= 0-0 =0
- – d(4,3)= 0.5-0=0

– d(2,2)= 3-3 = 0

– d(3,2)= 0.5-0 = 0.5

– d(4,2)= 1-0 = 1

– d(3,3)= 0.5-0.5 = 0

– d(4,4)= 0-0 = 0

	1	2	3	4
1	0			
2	1	0		
3	0.5	0.5	0	
4	0	1	0	0

DISSIMILARITY MATRIX

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

- BINARY ATTRIBUTES
  - DISSIMILARITY ATTRIBUTES
    - SYMMETRIC BINARY
      - $(r+s) / (q+r+s+t)$
    - ASYMMETRIC BINARY
      - $(r+s) / (q+r+s)$
  - SIMILARITY ATTRIBUTES
    - SYMMETRIC BINARY
      - SIMPLE MATCHING COEFFICIENT (SMC)
      - $(M_{11} + M_{00}) / (M_{11} + M_{10} + M_{01} + M_{00})$
    - ASYMMETRIC BINARY
      - JACCARD COEFFICIENT
      - $q / (q + r + s)$
      - OR
      - $(M_{11} + M_{00}) / (M_{11} + M_{10} + M_{01} + M_{00})$

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

	1	0	
1	q	r	M_{11}
0	s	t	M_{10}
	M_{01}	M_{00}	

q	**M_{11}**
r	M_{10}
s	M_{01}
t	M_{00}

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

DISSIMILARITY (ASYMETRIC BINARY)

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

DATA

	TEST 1	TEST 2	TEST 3	TEST 4	TEST 5	TEST 6
JACK	1	0	1	0	0	0
JIM	1	0	1	0	1	0
MARY	1	1	0	0	0	0

DISSIMILARITY (ASYMETRIC BINARY)

$$d(i,j) = (M_{10} + M_{01}) / (M_{11} + M_{10} + M_{01})$$

	JACK	JIM	MARY
JACK			
JIM			
MARY			

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

DATA

	TEST 1	TEST 2	TEST 3	TEST 4	TEST 5	TEST 6
JACK	1	0	1	0	0	0
JIM	1	0	1	0	1	0
MARY	1	1	0	0	0	0

DISSIMILARITY (ASYMETRIC BINARY)

$$d(i,j) = (M_{10} + M_{01}) / (M_{11} + M_{10} + M_{01})$$

	JACK	JIM	MARY
JACK	0		
JIM		0	
MARY			0

DISSIMILARITY (ASYMETRIC BINARY)

$$d(i,j) = (M_{10} + M_{01}) / (M_{11} + M_{10} + M_{01})$$

DATA

	TEST 1	TEST 2	TEST 3	TEST 4	TEST 5	TEST 6
JACK	1	0	1	0	0	0
JIM	1	0	1	0	1	0
MARY	1	1	0	0	0	0

	JACK	JIM	MARY
JACK	0		
JIM		0	
MARY			0

DISSIMILARITY MATRIX

$$\begin{aligned} d(JIM,JACK) &= (M_{10} + M_{01}) / (M_{11} + M_{10} + M_{01}) \\ &= (1+0) / (2+1+0) \\ &= 1/3 \\ &= 0.33 \end{aligned}$$

	JACK	JIM	MARY
JACK	0		
JIM	0.33	0	
MARY			0

DISSIMILARITY (ASYMETRIC BINARY)

DATA

	TEST 1	TEST 2	TEST 3	TEST 4	TEST 5	TEST 6
JACK	1	0	1	0	0	0
JIM	1	0	1	0	1	0
MARY	1	1	0	0	0	0

DISSIMILARITY MATRIX

	JACK	JIM	MARY
JACK	0		
JIM	0.33	0	
MARY			0

$$d(i,j) = (M_{10} + M_{01}) / (M_{11} + M_{10} + M_{01})$$

$$\begin{aligned} d(MARY,JACK) &= (M_{10} + M_{01}) / (M_{11} + M_{10} + M_{01}) \\ &= (1+1) / (1+1+1) \\ &= 2/3 \\ &= 0.67 \end{aligned}$$

DISSIMILARITY MATRIX

	JACK	JIM	MARY
JACK	0		
JIM	0.33	0	
MARY	0.67		0

DISSIMILARITY (ASYMETRIC BINARY)

DATA

	TEST 1	TEST 2	TEST 3	TEST 4	TEST 5	TEST 6
JACK	1	0	1	0	0	0
JIM	1	0	1	0	1	0
MARY	1	1	0	0	0	0

$$d(i,j) = (M_{10} + M_{01}) \,/\, (M_{11} + M_{10} + M_{01})$$

$$\begin{aligned} d(MARY,JIM) &= (M_{10} + M_{01}) \,/\, (M_{11} + M_{10} + M_{01}) \\ &= (1+2) \,/ (1+1+2) \\ &= 3 \,/ 4 \\ &= 0.75 \end{aligned}$$

DISSIMILARITY MATRIX

	JACK	JIM	MARY
JACK	0		
JIM	0.33	0	
MARY	0.67		0

DISSIMILARITY MATRIX

	JACK	JIM	MARY
JACK	0		
JIM	0.33	0	
MARY	0.67	0.75	0

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

DISSIMILARITY (SYMETRIC BINARY)

DISSIMILARITY (SYMETRIC BINARY)

DATA

	TEST 1	TEST 2	TEST 3	TEST 4	TEST 5	TEST 6
JACK	1	0	1	0	0	0
JIM	1	0	1	0	1	0
MARY	1	1	0	0	0	0

$d(i,j) = (M_{10} + M_{01}) / (M_{11} + M_{10} + M_{01} + M_{00})$

$d(JIM, JACK) = (1+0)/(2+1+0+3)$
$= 1/6 = 0.166$

$d(i,j) = (M_{10} + M_{01}) / (M_{11} + M_{10} + M_{01} + M_{00})$

$d(MARY, JACK) = (1+1)/(1+1+1+3)$
$= 2/6 = 1/3 = 0.33$

	JACK	JIM	MARY
JACK	0		
JIM		0	
MARY			0

	JACK	JIM	MARY
JACK	0		
JIM	0.166	0	
MARY			0

	JACK	JIM	MARY
JACK	0		
JIM	0.166	0	
MARY	0.33		0

DISSIMILARITY (SYMETRIC BINARY)

DATA

	TEST 1	TEST 2	TEST 3	TEST 4	TEST 5	TEST 6
JACK	1	0	1	0	0	0
JIM	1	0	1	0	1	0
MARY	1	1	0	0	0	0

$d(i,j) = (M_{10} + M_{01}) / (M_{11} + M_{10} + M_{01} + M_{00})$

$d(MARY,JIM) = (1+2)/(1+1+2+2)$

$= 3/6 = ½ = 0.5$

	JACK	JIM	MARY
JACK	0		
JIM	0.166	0	
MARY	0.33		0

	JACK	JIM	MARY
JACK	0		
JIM	0.166	0	
MARY	0.33	0.5	0

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

SIMILARITY (SYMMETRIC BINARY) SIMPLE MATCHING COEFFICIENT (SMC)

EXAMPLE

(x) = (1,0,0,0,0,0,0,0,0,0)

(y)= (0,0,0,0,0,0,1,0,0,1)

Formula :

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

$$S(x, y) = (0+7)/(0+1+2+7)$$

$$= 7/10$$

$$= 0.7$$

PROXIMITY MEASURE FOR BINARY ATTRIBUTES

SIMILARITY (ASYMMETRIC BINARY)

JACAARD COEFFICIENT

EXAMPLE

(x) = (1,0,0,0,0,0,0,0,0,0)

(y)= (0,0,0,0,0,0,1,0,0,1)

Formula :

S=q / (q + r + s)

or

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

S (x , y)= (0)/(0+1+1)

= 0/2

=0

PROBLEM

Consider the list of items bought by two customers as follows among 1000 available items:-

C1 = {sugar, coffee, tea, rice, egg}

C2 = {sugar, coffee, bread, biscuit}

Find the similarity between the items bought by two customers using SCM method and Jaccard coefficient.

Solution:

M_{11} = Items present in C1 & C2 = {sugar, coffee} =2

M_{10}= Items present in C1 but NOT in C2 = {tea, rice, egg} =3

M_{01}= Items present in C2 but NOT in C1 = {bread, biscuit} =2

M_{00}= Items NOT present both in C1 and C2. = Total item – (M_{11} + M_{10} + M_{01})

= 1000- (2+3+2)=993

Jaccard coefficient

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

S(C1,C2) = (2)/(2+3+2) = 2/7 = 0.285

SMC

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

S(C1,C2) = (2+993) /(2+3+2+993) =995/1000 =0.995

PROXIMITY MEASURE FOR NUMERIC ATTRIBUTES

EXAMPLE

Distance for numeric attribute can be measured using : Euclidean Distance or Manhattan Distance.

Data

	attribute1	attribut2
P1	0	2
P2	2	0
P3	3	1
P4	5	1

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Distance Matrix

	p1	p2	p3	p4
p1				
p2				
p3				
p4				

EXAMPLE

Distance for numeric attribute can be measured using : Euclidean Distance or Manhattan Distance.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Data

	attribute1	attribut2
P1	0	2
P2	2	0
P3	3	1
P4	5	1

	p1	p2	p3	p4
p1	0			
p2	2.8	0		
p3	3.2	1.4	0	
p4	5.1	3.2	2.0	0

Distance Matrix
(Euclidean Function)

EXAMPLE

Formula

Manhattan $\sum_{i=1}^{k} |x_i - y_i|$

Data

	attribute1	attribut2
P1	0	2
P2	2	0
P3	3	1
P4	5	1

	p1	p2	p3	p4
p1	0			
p2	4	0		
p3	4	2	0	
p4	6	4	2	0

Distance Matrix
(Manhattan Function)

PROXIMITY MEASURE FOR MIXED ATTRIBUTES

EXAMPLE

DATA

ID	TEST1	TEST2	TEST3
1	CODE A	EXCELLENT	45
2	CODE B	FAIR	22
3	CODE C	GOOD	64
4	CODE A	EXCELLENT	28

TEST 1 = NOMINAL ATTRIBUTE
TEST 2 = ORDINAL ATTRIBUTE
TEST 3 = NUMERICAL ATTRIBUTE

FORMULA:

δ_{ij} =0 , if x_i or x_j is missing
or x_i =0 or x_j =0
δ_{ij} =1, otherwise

$$d(x_i, x_j) = \frac{\sum_{n=1}^{p} \delta_{ij}^{(n)} d_{ij}^{(n)}}{\sum_{n=1}^{p} \delta_{ij}^{(n)}}$$

SOLUTION

DATA

ID	TEST1	TEST2	TEST3
1	CODE A	EXCELLENT	45
2	CODE B	FAIR	22
3	CODE C	GOOD	64
4	CODE A	EXCELLENT	28

DISSIMILARITY MATRIX OF NOMINAL ATTRIBUTE

	1	2	3	4
1	0			
2	1	0		
3	1	1	0	
4	0	1	1	0

DISSIMILARITY MATRIX OF ORDINAL ATTRIBUTE

	1	2	3	4
1	0			
2	1	0		
3	0.5	0.5	0	
4	0	1.0	0.5	0

SOLUTION

DATA

ID	TEST1	TEST2	TEST3
1	CODE A	EXCELLENT	45
2	CODE B	FAIR	22
3	CODE C	GOOD	64
4	CODE A	EXCELLENT	28

ID	TEST 3
1	45
2	22
3	64
4	28

Need to normalize these numerical values so that it can be mapped in the range [0-1]

DISSIMILARITY MATRIX OF NUMERICAL ATTRIBUTE

	1	2	3	4
1	0			
2		0		
3			0	
4				0

Formula :

$d_{ij} = | x_1 – x_{2| /}$ /(max - min)

d(2,1) = |41 -22| /(64 -22) =23/42 =0.55

d(3,1) = |45 -64| /(64 -22 =0.45

d(3,2) = |22 -64| /(64 -22) =1.0

Solving further we get the final distance matrix given in next slide.

DISSIMILARITY MATRIX OF NUMERICAL ATTRIBUTE

	1	2	3	4
1	0			
2	0.55	0		
3	0.45	1	0	
4	0.40	0.14	0.86	0

DATA

ID	TEST1	TEST2	TEST3
1	CODE A	EXCELLENT	45
2	CODE B	FAIR	22
3	CODE C	GOOD	64
4	CODE A	EXCELLENT	28

FINAL DISSIMILARITY MATRIX OF MIXED ATTRIBUTE

	1	2	3	4
1	0			
2		0		
3			0	
4				0

d(2,1) =[(1x1)+(1x1)+(1x0.55)] / (1+1+1) = 0.85
d(3,1) =[(1x1)+(1x0.5)+(1x0.45) /(1+1+1) =0.65
Solving other values in similar manner we get the final dissimilarity matrix for mixed attributes as given in next slide.

FINAL DISSIMILARITY MATRIX OF MIXED ATTRIBUTE

	1	2	3	4
1	0			
2	0.85	0		
3	0.65	0.83	0	
4	0.13	0.71	0.79	0

NEXT CLASS

- CLASSIFICATION
- MACHINE LEARNING MODEL
- CLASSIFICATION & PREDICTION
- DATA SEPARABILITY
- DECISION BOUNDARY