National Textile University

ASSIGNMENT NO# 02

Title:

Dissimilarity Matrix of Mixed Type Attributes with Example

Submitted To:

Respected, Sir Dr. Haseeb Ahmad

Submitted By:

Muhammad Azam	18-NTU-CS-1178
Muhammad Furqan Haider	18-NTU-CS-1183
Usama Sadiq	18-NTU-CS-1191

Mixed Attribute Types

Attribute:

A data field, representing a characteristic or feature of a data object.

e.g., name, address

Types:

- Nominal
- Binary
- Numeric
- Ordinal

Nominal:

Nominal means **"relating to names**." The values of a nominal attribute are. symbols or names of things. Each value represents category, code, or state and so nominal attributes are also referred to as categorical

Hair color = {*auburn, black, blond, brown, grey, red, white*}

marital status, occupation, ID numbers, zip codes

Binary

Binary data has only 2 values/states.
For Example, yes or no, affected or unaffected, true or false.

- **Symmetric:** Both values are equally important (Gender).
- **Asymmetric:** Both values are not equally important (Result).

  e.g., medical test (positive vs. negative)

Numeric:

A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of 2 types, **interval**, and **ratio**.

- An **interval-scaled** attribute has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point, or we can call zero points. Data can be added and subtracted at an interval scale but can not be multiplied or divided. Consider an example of temperature in degrees Centigrade. If a day's temperature of one day is twice of the other day, we cannot say that one day is twice as hot as another day.
  E.g., temperature in C˚or F˚, calendar dates

- A **ratio-scaled** attribute is a numeric attribute with a fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value. The values are ordered, and we can also compute the difference between values, and the mean, median, mode, Quantile-range, and Five number summary can be given.
  e.g., temperature in Kelvin, length, counts, monetary quantities

- **Continuous**: Continuous data have an infinite no of states. Continuous data is of float type.
  e.g., height 5.9, length 34.5

Ordinal:

The Ordinal Attributes contains values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don't indicate how important it is.

Size = {*small, medium, large*}, grades, army rankings

Example

ID	Marks	Sessional	Grade	Gender
1	27	Excellent	A	M
2	23	Very Good	B	F
3	21	Good	B	M
4	20	Average	C	F

Solution:

Dissimilarity Matrix of Numeric Attribute (Marks)

ID	Marks
1	27
2	23
3	21
4	20

$$d_{(i,j)} = \frac{|x_i - x_j|}{max - min}$$

$$d_{(2,1)} = \frac{|27 - 23|}{27 - 20}$$

$$d_{(2,1)} = 0.57$$

$$d_{(3,1)} = \frac{|27 - 21|}{27 - 20}$$

$$d_{(3,1)} = 0.86$$

$$d_{(4,1)} = \frac{|27 - 20|}{27 - 20}$$

$$d_{(4,1)} = 1$$

$$d_{(3,2)} = \frac{|21 - 23|}{27 - 20}$$

$$d_{(3,2)} = 0.28$$

$$d_{(4,2)} = \frac{|20 - 23|}{27 - 20}$$

$$d_{(4,2)} = 0.43$$

$$d_{(4,3)} = \frac{|20 - 21|}{27 - 20}$$

$$d_{(4,3)} = 0.14$$

Dissimilarity Matrix

$$\begin{bmatrix} 0 & & & \\ 0.57 & 0 & & \\ 0.86 & 0.28 & 0 & \\ 1 & 0.43 & 0.14 & 0 \end{bmatrix}$$

Dissimilarity Matrix of Ordinal Attribute (Sessional)

ID	Sessional
1	Excellent
2	Very Good
3	Good
4	Average

Total State =4

Rank

Excellent=4, Very Good=3, Good=2, Average=1

Normalize Ranking

$$z_{if} = \frac{r_{if} - 1}{m_f - 1}$$

For Excellent

$$z = \frac{4 - 1}{4 - 1}$$

$$z = 1$$

For Very Good

$$z = \frac{3 - 1}{4 - 1}$$

$$z = 0.67$$

For Good

$$z = \frac{2-1}{4-1}$$

$$z = 0.33$$

For Average

$$z = \frac{1-1}{4-1}$$

$$z = 0$$

Using Manhattan Distance

$$d_{(i,j)} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

$$d_{(2,1)} = |0.67 - 1|$$

$$d_{(2,1)} = 0.33$$

$$d_{(3,1)} = |0.33 - 1|$$

$$d_{(3,1)} = 0.67$$

$$d_{(4,1)} = |0 - 1|$$

$$d_{(4,1)} = 1$$

$$d_{(3,2)} = |0.33 - 0.67|$$

$$d_{(3,2)} = 0.33$$

$$d_{(4,2)} = |0 - 0.67|$$

$$d_{(4,2)} = 0.67$$

$$d_{(4,3)} = |0 - 0.33|$$

$$d_{(4,3)} = 0.33$$

Dissimilarity Matrix

$$\begin{bmatrix} 0 & & & \\ 0.33 & 0 & & \\ 0.67 & 0.34 & 0 & \\ 1 & 0.67 & 0.33 & 0 \end{bmatrix}$$

Dissimilarity Matrix of Nominal Attribute (Grade)

ID	Grade
1	A
2	B
3	B
4	C

$$d_{(i,j)} = \frac{p - m}{p}$$

In our question total attributes are 4 therefore p=4

$$d_{(2,1)} = \frac{4 - 0}{4}$$

$$d_{(2,1)} = 1$$

$$d_{(3,1)} = \frac{4 - 0}{4}$$

$$d_{(3,1)} = 1$$

$$d_{(4,1)} = \frac{4 - 0}{4}$$

$$d_{(4,1)} = 1$$

$$d_{(3,2)} = \frac{4 - 1}{4}$$

$$d_{(3,2)} = 0.75$$

$$d_{(4,2)} = \frac{4 - 0}{4}$$

$$d_{(4,2)} = 1$$

$$d_{(4,3)} = \frac{4 - 0}{4}$$

$$d_{(4,3)} = 1$$

Dissimilarity Matrix

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 0.75 & 0 & \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

Dissimilarity Matrix of Binary Attribute (Gender)

ID	Gender
1	M
2	F
3	M
4	F

Object i		Object j		
		M	**F**	**Sum**
	M	q	r	q+r
	F	s	t	s+t
	Sum	q+s	r+t	p

Distance Measure for Symmetric Binary Variables

$$d_{(i,j)} = \frac{r+s}{q+r+s+t}$$

$$d_{(2,1)} = \frac{0+1}{0+0+1+0}$$

$$d_{(2,1)} = 1$$

$$d_{(3,1)} = \frac{0+0}{1+0+0+0}$$

$$d_{(3,1)} = 0$$

$$d_{(4,1)} = \frac{0+1}{0+0+1+0}$$

$$d_{(4,1)} = 1$$

$$d_{(3,2)} = \frac{1+0}{0+1+0+0}$$

$$d_{(3,2)} = 1$$

$$d_{(4,2)} = \frac{0+0}{0+0+0+1}$$

$$d_{(4,2)} = 0$$

$$d_{(4,3)} = \frac{0+1}{0+0+1+0}$$

$$d_{(4,3)} = 1$$

Dissimilarity Matrix

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0 & 1 & 0 & \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

Formula For Dissimilarity Matrix of Mixed Type Attributes

$$d_{(i,j)} = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

$$d_{(2,1)} = \frac{(1 * 0.57) + (1 * 0.33) + (1 * 1) + (1 * 1)}{1 + 1 + 1 + 1}$$

$$d_{(2,1)} = 0.72$$

$$d_{(3,1)} = \frac{(1 * 0.86) + (1 * 0.67) + (1 * 1) + (1 * 0)}{1 + 1 + 1 + 1}$$

$$d_{(3,1)} = 0.63$$

$$d_{(4,1)} = \frac{(1 * 1) + (1 * 1) + (1 * 1) + (1 * 1)}{1 + 1 + 1 + 1}$$

$$d_{(4,1)} = 1$$

$$d_{(3,2)} = \frac{(1 * 0.28) + (1 * 0.34) + (1 * 75) + (1 * 1)}{1 + 1 + 1 + 1}$$

$$d_{(3,2)} = 0.59$$

$$d_{(4,2)} = \frac{(1 * 0.43) + (1 * 0.67) + (1 * 1) + (1 * 0)}{1 + 1 + 1 + 1}$$

$$d_{(4,2)} = 0.52$$

$$d_{(4,3)} = \frac{(1 * 0.14) + (1 * 0.33) + (1 * 1) + (1 * 1)}{1 + 1 + 1 + 1}$$

$$d_{(4,3)} = 0.62$$

Dissimilarity Matrix

$$\begin{bmatrix} 0 & & & \\ 0.72 & 0 & & \\ 0.63 & 0.59 & 0 & \\ 1 & 0.52 & 0.62 & 0 \end{bmatrix}$$