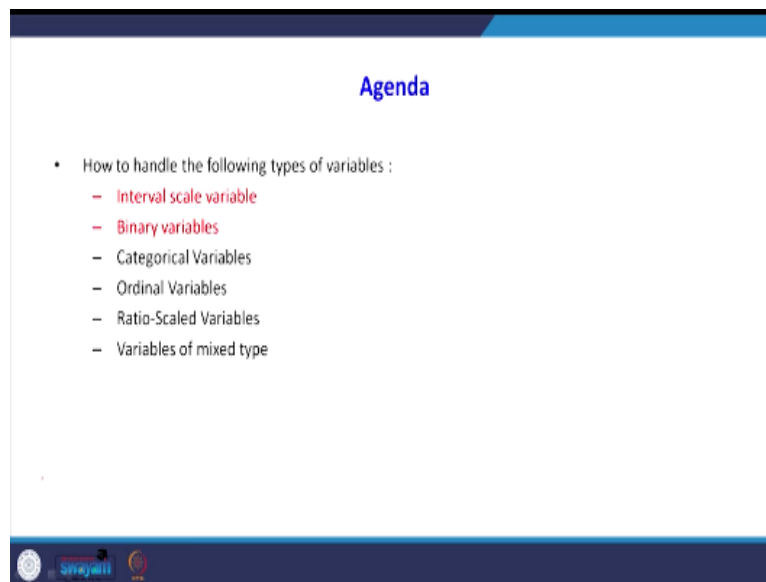


Data Analytics with Python
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 52
Cluster Analysis - IV

In my previous lecture, I have explained how to handle missing data while doing clustering analysis. Second thing I have explained how to find dissimilarity and similarity matrix. The third one I have explained there is a binary variable, how to find out the dissimilarity and similarity matrix.

(Refer Slide Time: 00:46)



We will now handle other variables. For example, the agenda for these classes, if there is an interval scale variable, how to use that dataset for the cluster analysis and binary variable and how to use that dataset for the cluster analysis; we have done in our previous lecture. In this class, we will see; if there is a categorical variable, how to use that dataset for clustering analysis. Next we will see there is an ordinal variable, how to use for our cluster analysis.

And if there are ratio-scaled variables and how to do that data for our cluster analysis, and finally the data may be mixed type, the combination of these above data, in that case how to use that dataset for our clustering analysis. The agenda for this lecture is how to handle the following

data types. One is interval-scaled variable and binary variable that I have covered in my previous lectures.


In this lecture if the variable nature is categorical, ordinal, ratio and combination of above dataset that is mixed type, let us see how to see this kind of variables and how to use these kind of variables for our clustering analysis.

(Refer Slide Time: 02:04)

Categorical Variables

Categorical Variables

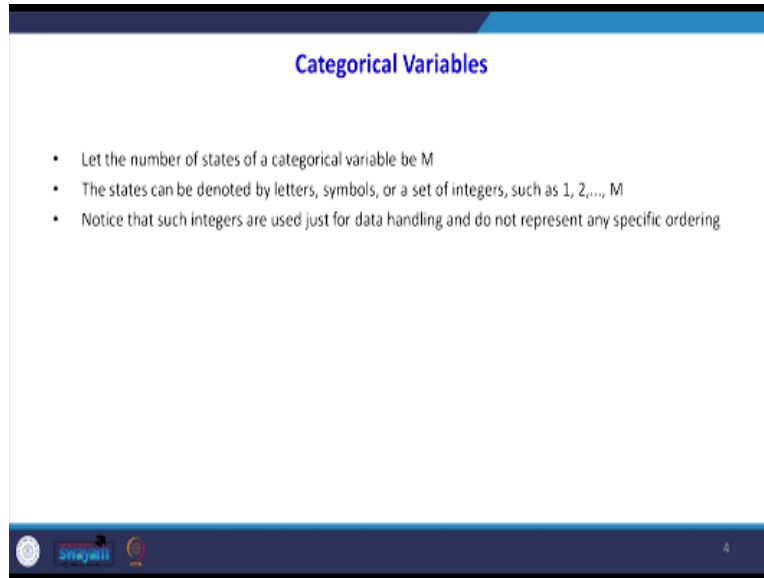
- A categorical variable is a generalization of the binary variable in that it can take on more than two states
- For example, map color is a categorical variable that may have, say, five states: red, yellow, green, purple, and blue



3

First we will take an example, categorical variable. A categorical variable is a generalization of binary variable in that it can take on more than two states. It is similar to binary variable but in binary variable only 2 option will be there. But in categorical variable there maybe more than 2 states. For example, map color is a categorical variable that may have say 5 states red, yellow, green, purple and blue. This is an example of categorical dataset.

(Refer Slide Time: 02:36)



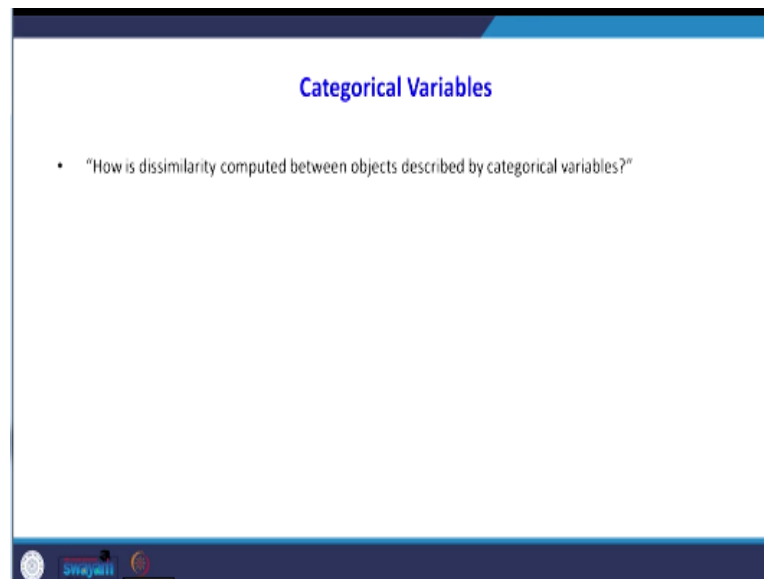
A presentation slide with a blue header and footer. The title 'Categorical Variables' is in blue text. The slide contains three bullet points. The footer includes a logo on the left and a small number '4' on the right.

Categorical Variables

- Let the number of states of a categorical variable be M
- The states can be denoted by letters, symbols, or a set of integers, such as $1, 2, \dots, M$
- Notice that such integers are used just for data handling and do not represent any specific ordering

Let the number of states of a categorical variable be capital M . The states can be denoted by letters, symbols, or a set of integers such as $1, 2$ up to M . Notice that such integers are just used for data handling and do not represent any specific ordering.

(Refer Slide Time: 02:57)



A presentation slide with a blue header and footer. The title 'Categorical Variables' is in blue text. The slide contains one bullet point. The footer includes a logo on the left and a small number '5' on the right.

Categorical Variables

- "How is dissimilarity computed between objects described by categorical variables?"

The question which we are going to answer in this lecture is how dissimilarity computed between objects described by categorical variables.

(Refer Slide Time: 03:05)

Categorical Variables

- The dissimilarity between two objects i and j can be computed based on the ratio of mismatches:

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

where 'm' is the number of matches (i.e., the number of variables for which 'i' and 'j' are in the same state), and 'p' is the total number of variables

Weights can be assigned to increase the effect of 'm' or to assign greater weight to the matches in variables having a larger number of states

The dissimilarity between two objects i and j can be computed based on the ratio of mismatches. So the formula to find out the dissimilarity for a categorical variable is d of $(i, j) = (p - m)$ divided by p , notice that it is a small m , m is the number of matches that is the number of variables for which i and j are in the same state and the p is the total number of variables which can be assigned to increase the effect of m or to assign greater weight to the matches in the variables having a larger number of states. So we can give weightage also for different states.

(Refer Slide Time: 03:48)

Dissimilarity between categorical variables

- Suppose that we have the sample data as shown in the table
- Let only the object-identifier and the variable (or attribute) test-1 are available, which is a categorical data

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210

Finding Groups in Data: An Introduction to Cluster Analysis
Authors: Edward R. Kaufman, Peter J. Rousseeuw
March 1990, John Wiley & Sons, Inc.

Now we will take an example with the help of example I will explain how to find out the dissimilarity between categorical variables. So the source for this example is, this book finding groups in data and introduction to cluster analysis by Kaufman and Rousseeuw, the publisher is

John Wiley. Suppose that we have the sample data as shown in the table, the table having 1, 2, 3 there are 4 columns.

One is object identifier, second column is test-1 categorical data, third column is test-2 ordinal, the fourth one is test-3 ratio-scaled. Let only the object identifier and variable test-1 are available which is a categorical data, so for this example we are going to consider only 2 column, one is this object identifier, second one is the test-1 which is categorical data.

(Refer Slide Time: 04:47)

Dissimilarity matrix				
	1	2	3	4
1	0			
2	$d(2,1)$	0		
3	$d(3,1)$	$d(3,2)$	0	
4	$d(4,1)$	$d(4,2)$	$d(4,3)$	0

The dissimilarity matrix, see that we are showing only the lower triangle that is 1, 2, 3, 4 is a object identifier, the diagonal will be 0 because the dissimilarity is 0 for this same value of when $i = j$. So this location is d 2, 1 second row first column, this location is third row first column third row second column, this location is fourth row first column fourth row second column and fourth row third column.

(Refer Slide Time: 5:18)

Dissimilarity between categorical variables

- Since here we have one categorical variable, test-1, we set $p = 1$ in Equation

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

So that $d(i, j)$ evaluates to '0' if objects i and j match, and '1' if the objects differ

- Thus, we get

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

$$d(2, 1) = (1 - 0) / 1 = 1$$

$$d(4, 1) = (1 - 1) / 1 = 0$$

object Identifier	test-1 (categorical)
1	code-A
2	code-B
3	code-C
4	code-A

Since here we have only one categorical variable, test-1, we set $p = 1$ in equation because p is the number of variables. So that $d(i, j)$ evaluate 0 if the object i and j match and 1 if the object differ, thus we get 0, 1 0, 1 1 0, 0 1 1 0. I will tell you how we got this matrix. Suppose let us find out how this location has come this value is 1, the distance between you see that this matrix $d(2, 1)$ so $d(2, 1)$ is we will use this $p - m$ divided by small p so p is number of variables because here only one variable minus m .



When you compare 1 and 2 because the 2, 1 or 2, 1 see code-A and code-B it is not matching so the value of m is 0 so $1 - 0$ divided by ($p = 1$), that is why we got this one value. Let us find out this value $d(4, 1)$. So there is a one variable is this 4 another variable this, so we will use the same formula $d(i, j)$ equal to; so $p = 1$, see code A and; for object identifier the code A is same for 1 and 4. So here the m is 1, so p also 1, m also 1, so this value is 0. This way the all other values were found.

(Refer Slide Time: 06:52)

Ordinal Variables

- A discrete ordinal variable resembles a categorical variable, except that the 'M' states of the ordinal value are ordered in a meaningful sequence
- Ordinal variables are very useful for registering subjective assessments of qualities that cannot be measured objectively *V. good, Good, bad*
- For example, professional ranks are often enumerated in a sequential order, such as Assistant, Associate, and full for Professors
- A continuous ordinal variable looks like a set of continuous data of an unknown scale; that is, the relative ordering of the values is essential but their actual magnitude is not

99 - 1
50 - 2



10

Now let us go to the next type of variable, ordinal variable. So ordinal variable is similar to categorical variable but the order is more important. So in my; when I am the explaining the previous data that is a categorical variable one example is the pin code, for example in India the pin code is an example for our categorical data, so that number is not representing any meaning. So now we will start with ordinal variables.

A discrete ordinal variable resembles a categorical variable, except that M states of the ordinal values are ordered in a meaningful sequence, this is term is very important because there is a ranking, there is a order in each value. Ordinal variables are very useful for registering subjective assessments of qualities that cannot be measured objectively. For example, say very good, good, bad so this way we can give the rank 1, 2, 3.

So here 1, 2, 3 says that is a rank for that. For example, professional ranks are often enumerated in a sequential order, such as Assistant, Associate, and full professors. So the order is more important. A continuous ordinal variable look like a set of continuous data of an unknown scale; that is the relative ordering of the values is essential but their actual magnitude is not. The problem with the ordinal scale in your class.

Suppose you are giving rank, those who got 99 is rank number 1, so those who are got number 2 50 marks rank number 2. See that this 1 and 2 signifies the rank but it is not the actual value

because this fellow got 99 marks, this fellow got 50 marks. So we are losing some important information when you go for ordinal dataset. Because for us this 1 and 2 is more important, how much mark they got is not important for us.

(Refer Slide Time: 08:58)

The slide is titled "Ordinal Variables" in blue text. It contains three bullet points:

- For example, the relative ranking in a particular sport (e.g., gold, silver, bronze) is often more essential than the actual values of a particular measure
- Ordinal variables may also be obtained from the discretization of interval-scaled quantities by splitting the value range into a finite number of classes
- The values of an ordinal variable can be mapped to ranks

Below the text is a hand-drawn diagram in red ink. It shows a step-like structure with three levels. The top level is labeled '1', the middle level is labeled '2', and the bottom level is labeled '3'. The diagram illustrates the concept of discrete ranks or classes.

At the bottom of the slide, there is a footer with logos on the left and the number '11' on the right.

For example, the relative ranking is a particular sport for example, gold, silver, bronze is often more essential than the actual value of a particular measures, because see there may be a different 3 scales, so rank 1, rank 2, rank 3 so this person gold, this fellow silver, this fellow bronze. Here what is more important is the; the rank is more important not the actual measures. So ordinal variable may also be obtained from discretization of intervals quantities by splitting the value range into finite number of classes.

Sometimes what happen, if there is a interval-scaled dataset that can be converted into ordinal variables. The values of ordinal variables can be mapped into ranks, so after converting ordinal then we can bring different ranks.

(Refer Slide Time: 09:52)

Dissimilarity computation

- The treatment of ordinal variables is quite similar to that of interval-scaled variables when computing the dissimilarity between objects
- Suppose that 'f' is a variable from a set of ordinal variables describing 'n' objects
- The dissimilarity computation with respect to 'f' involves the following steps:
- The value of 'f' for the i^{th} object is x_{if} , and 'f' has M_f ordered states, representing the ranking $1, \dots, M_f$.
- Replace each x_{if} by its corresponding rank, $r_{if} \in \{1, \dots, M_f\}$.

Let us see how to find out dissimilarity matrix for our ordinal dataset. The treatment of ordinal variable is quiet similar to that of interval-scaled variables when computing the dissimilarity between objects. Suppose that f is variable from a set of ordinal variables describing n objects. The dissimilarity computation with respect to f involves the following steps. The first one is the value of f for the i^{th} object is $x(i, f)$ and f has M of ordered states, representing the ranking 1 to M.

So the M is the maximum number of rank. The $x(i, f)$ is the particular variable. So what we have to do we have to replace each $x(i, f)$ by its corresponding rank r_{if} , so r_{if} is the current rank the M_f is the maximum rank.

(Refer Slide Time: 10:51)

Dissimilarity computation

	A	B	C	D	E	F	G
B	69.8						
C	2.0	70.8					
D	71.6	5.7	72.5				
E	108.6	42.2	109.9	43.9			
F	95.7	26.3	96.8	26.4	19.1		
G	5.8	75.7	5.3	77.4	114.3	101.6	
H	17.7	87.2	17.3	89.2	125.0	112.9	12.2

Look at this data. This is; this also we have seen our previous lecture, this is Euclidean distance. So this is an example of dissimilarity computation. We have seen this previous table that is the Euclidean distances. This is an example of interval-scaled data. From this interval-scaled data we can convert this table into in ordinal dataset. So what we have to do, so suppose the lowest distance is highest rank.


So the lowest one is this one value, so this can be ranked as 1, the second one is 5., so this is rank 2. So next one is 5.7 rank 3, so; and so on, for each variable that is the interval dataset, you can convert into ordinal dataset by giving rank like 1, 2, 3 and so on. So the highest value will have the highest rank.

(Refer Slide Time: 11:52)

Standardization of ordinal variable

- Since each ordinal variable can have a different number of states, it is often necessary to map the range of each variable onto $[0.0, 1.0]$ so that each variable has equal weight.
- This can be achieved by replacing the rank r_{if} of the i^{th} object in the f^{th} variable by:

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



14

Standardization of ordinal variable. Since each ordinal variable can have a different number of states, it is often necessary to map the range of each variable onto 0 to 1 scale, so that each variable has equal weight. This can be achieved by replacing the rank r_{if} of the i^{th} object in the f^{th} variable by z_{if} equal to $(r_{if} - 1)$ divided by $(M_f - 1)$. So the r_{if} represents the current rank M_f represents the maximum rank.

(Refer Slide Time: 12:31)

Dissimilarity computation

- Dissimilarity can then be computed using any of the distance measures described earlier (like that for interval data)


15

Now let us see how to find out the dissimilarity computation. The dissimilarity can be computed using any of the distance measures described by earlier like that interval data.

(Refer Slide Time: 12:44)

Example

- Suppose that we have the sample data of the following Table ,
- Except that this time only the object-identifier and the continuous ordinal variable, test-2, are available
- There are three states for test-2, namely fair, good, and excellent, that is $M_f = 3$

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210

Now, let us take ordinal data the I will explain how to find out the dissimilarity matrix. Suppose that we have the sample data of following table, the same table which I have seen. So there are columns object identifier, test-1 categorical, test-2. Now we are going to consider these two column, one is object identifier next one is test-2 that is ordinal dataset. Suppose that we have the sample dataset for the following table except that is this time only the object identifier and the continuous ordinal variable, test-2 are available for us. There are 3 states of the test-2 namely good, excellent and fair, so the aim of three, because that is a maximum number of states.

(Refer Slide Time: 13:41)

Example

- For step 1, if we replace each value for test-2 by its rank, the four objects are assigned the ranks 3, 1, 2, and 3, respectively

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

- Step 2 normalizes the ranking by mapping rank 1 to 0.0, rank 2 to 0.5, and rank 3 to 1.0
- For step 3, we can use, say, the Euclidean distance, which results in the following dissimilarity matrix:

The step 1, if you replace each value of the test-2 by its rank, the four objects are assigned and the rank 3, 1, 2, 3 respectively. How is 3, 1, 2, 3? So this we are going to call test-3, this is also 3,

this is 1, this is 2. So how you are ranking this variable, 1 is fair, 2 is good, 3 is excellent. That is why we got this one, 3, 1; 3, 1, 2 and 3. Step 2, normalize the ranking by mapping 1 to 0, rank 2 to 0.5, rank 3 to 1. How we got this well, rank 1 = 0.

Because the r if value is 1 so $1 - 1$ divided by ; there are; the value of Mf is 3 so $3 - 1 = 2$; it is 0. The second one rank 2, $2 - 1$ divided by $3 - 1 = 2$ so what will happen, it is 1 divided by 2 it is a 0.5. The third one, so value of r if will be $3 - 1$ divided by this also $3 - 1$ so it is $2/2$ that is equal to 1. That is a way to standardize. Step 3, we can use say the Euclidean distance, which results the following dissimilarity matrix.

(Refer Slide Time: 15:16)

Dissimilarity computation

Object identifier

1 → 3 → 1

2 → 1 → 0

3 → 2 → 0.5

4 → 3 → 1

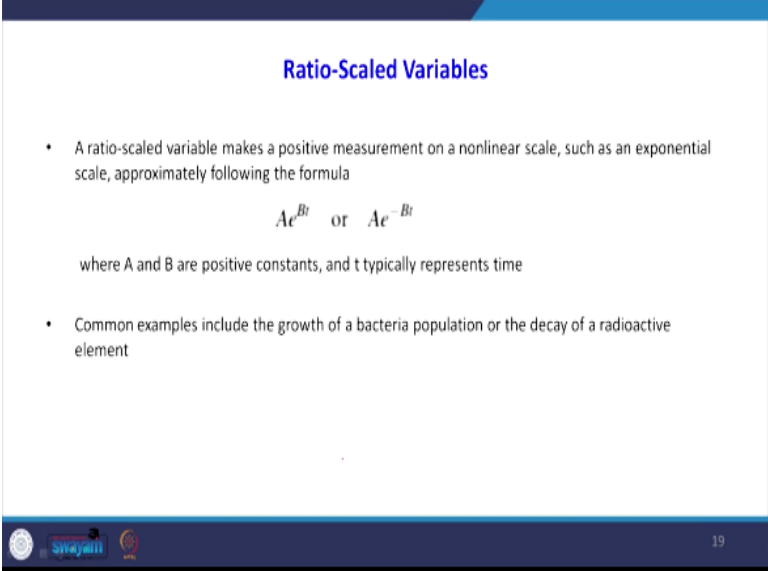
	1	2	3	4
1	0			
2	1	0		
3	0.5	0.5	0	
4	0	1.0	0.5	0

So this is our Euclidean distance. So the first column for example it says if the object identifier. This was our ranking ordinal dataset. This is our standardized value, so 3 is mapped into 2, 1; how we got this one see. $3 = 1$ so $1 = 0$, $2 = 0.5$, $3 = 1$. This is our standardized data. This is in the Z scale. This is our object identifier. This was our ordinal data. Now let us see how this matrix has come. Suppose if you want to know the distance between object identifier 2 and 1 so between 2 and 1 the distance is 1, so root of 1 square is 1.

So let us see how to find out distance between 3 and 1. So 3 is 1, so the formula is $1 - 0.5$ that is a 0.5 whole square. When you take square root this is 0.5. Let us see 3 to 2. So 3 to 2 $0 - 0.5$ whole square, square root that value is a 0.5. Let us see 4 to 1, what is the distance. So 4 to 1 is 1,

this distance is 1, $1 - 1$ that whole square, you take square root that as a 0. Let us see how we got this 1. So that is a 4 to 2. So 4 to 2 is 1 square, then square root that value is 1. So 4 to 3, so this distance is 0.5 - 1 whole square, square root that value is 0.5. So this is an Euclidean distance. This is our standardized z value.

(Refer Slide Time: 17:18)



Ratio-Scaled Variables

- A ratio-scaled variable makes a positive measurement on a nonlinear scale, such as an exponential scale, approximately following the formula

$$Ae^{Bt} \text{ or } Ae^{-Bt}$$

where A and B are positive constants, and t typically represents time

- Common examples include the growth of a bacteria population or the decay of a radioactive element

19

Now let us go to the next type of variable that is a Ratio-Scaled variable. A ratio-scaled variable makes a positive measurement on a nonlinear scale, such as an exponential scale, approximately the following the formula. Ae^{Bt} or Ae^{-Bt} where, A and B are positive constants, and t is typically representing the time. So common example include the growth of a bacteria population or the decay of radioactive element, so that is an example of ratio-scale. Here the concept of ratio-scale is, there is a meaning for our absolute 0. When there is a absolute 0 you can do all kind of arithmetic operation using ratio-scaled data.

(Refer Slide Time: 18:08)

Computing the dissimilarity between objects

- There are three methods to handle ratio-scaled variables for computing the dissimilarity between objects:
 1. Treat ratio-scaled variables like interval-scaled variables
 - This, however, is not usually a good choice since it is likely that the scale may be distorted ✓
 2. Apply logarithmic transformation to a ratio-scaled variable f having value x_{if} for object i by using the formula $y_{if} = \log(x_{if})$
 - The y_{if} values can be treated as interval valued, Notice that for some ratio-scaled variables, log-log or other transformations may be applied, depending on the variable's definition and the application ✓

So computing the dissimilarity between the objects. There are three methods to handle the ratio-scaled variable for computing the dissimilarity between the objects. The first method is, treat ratio-scaled variable like interval-scaled variables. This, however, is not usually a good choice since it is likely that the scale may be distorted. But most of the marketing examples we will not differentiate the ratio-scale and interval-scale even though we collect the interval scale.

So we will use as a ratio-scale for finding all kind of statistical test. The second method is apply logarithmic transformation to a ratio-scaled variable f having the value x_{if} for a object i and using the formula $Y_{if} = \log(x_{if})$. It is nothing but if there is a ratio-scale just you take log of that, so that can be used for further analysis for finding the dissimilarity matrix. The Y_{if} values can be treated as interval valued, notice that for some ratio-scaled variables, so log of log or other transformation may be applied, depending upon the variables definition and the applications, because for we can use any kind of different transformations.

(Refer Slide Time: 19:31)

Computing the dissimilarity between objects

3. Treat x_{ij} as continuous ordinal data and treat their ranks as interval-valued
- The latter two methods are the most effective, although the choice of method used may depend on the given application

The third method is treat X if as a continuous ordinal data and treat their rank as interval-values. So this is the third method. The latter two methods are the most effective, although the choice of method may depend upon the given application.

(Refer Slide Time: 19:49)

Example

- This time, we have the sample data of the following Table,
- Except that only the object-identifier and the ratio-scaled variable, test-3, are available

object identifier	test-1 (categorical)	test-2 (ordinal)	test-3 (ratio-scaled)
1	code-A	excellent	445
2	code-B	fair	22
3	code-C	good	164
4	code-A	excellent	1,210

Now let us taken an example of ratio-scaled then I will explain how to find out the dissimilarity matrix. This time we have the sample data of the following table except the only object identifier, so we are going to consider this one and the ratio-scaled variable this column. So we are going to consider only these two column for finding the dissimilarity matrix, so in the third column ratio-scaled 445, 22, 164, 1210.

(Refer Slide Time: 20:23)

Example

- Let's try a logarithmic transformation
- Taking the log of test-3 results in the values 2.65, 1.34, 2.21, and 3.08 for the objects 1 to 4, respectively
- Using the Euclidean distance on the transformed values, we obtain the following dissimilarity matrix:

$$\begin{matrix}
 & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\
 \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{bmatrix} 0 & & & \\ 1.31 & 0 & & \\ 0.44 & 0.87 & 0 & \\ 0.43 & 1.74 & 0.87 & 0 \end{bmatrix}
 \end{matrix}$$

object identifier	test-3 (ratio-scaled)
1	445
2	22
3	164
4	1,210

Let us try a logarithmic transformation. So just we are going to take the log of the column 3 values. Taking the log of test-3 results in the values become 2.65, 1.34, 2.21 and 3.08 for subject 1 to 4 respectively. When you look at this after taking log transformation the value it is compressed, see that it is scaled down; that is a purpose of scaling. So instead of using 445 you can use 2.65; it is easy to handle.

So instead of using 22 for cluster analysis application you can use 1.34. So the benefit of taking log of this one is it is compressed in a smaller scale. So using the Euclidean distance on the transformed values, we obtain the following dissimilarity matrix. So this is our dissimilarity matrix. So this is object identifier 1, 2, 3, 4. This is 1, 2, 3, 4. For example, d (2, 1) how we got this one? This is 1, this is 2 so we have to find the difference.

We have to use this formula $(2.65 - 1.35)^2$ whole square, then square root. That value is 1.31. For example, 1 and 3 so this is 1 and 3 so the difference is $(2.65 - 2.1)$ whole square take square root so that will be this value. The suppose 4 1, so suppose this the fourth one, so find the difference $(3.08 - 2.65)^2$, then take square root so that value is 0.43. The same way we can get the other cells.

(Refer Slide Time: 22:05)

Variables of Mixed Types

- So far we have discussed how to compute the dissimilarity between objects described by variables of the same type, where these types may be either interval-scaled, symmetric binary, asymmetric binary, categorical, ordinal, or ratio-scaled
- However, in many real databases, objects are described by a mixture of variable types

swayam 24

Now we will enter into another type of variable; it is not another type of variable where whenever we do the cluster analysis there is a possibility of these variables like categorical, interval binary may come together. So that type of data types we are calling it is mixed types. So far we have discussed how to compute the dissimilarity between objects, described by variables of the same type, where these types may be either interval-scaled, symmetric binary, asymmetric binary, categorical, ordinal or ratio-scaled.

However, in reality, in many real databases, objects are described by the mixture of variable types. So whenever the mixture of these variable types are coming how to use the dataset, how to standardized that dataset for our further analysis of our cluster analysis, that we will see now.

(Refer Slide Time: 23:07)

Variables of Mixed Types

- In general, a database can contain all of the six variable types listed above
- “So, how can we compute the dissimilarity between objects of mixed variable types?”
- One approach is to group each kind of variable together, performing a separate cluster analysis for each variable type
 - This is feasible if these analyses derive compatible results
 - However, in real applications, it is unlikely that a separate cluster analysis per variable type will generate compatible results



In general, a database can contain all of 6 variable types listed above. So, how can we compute the dissimilarity between objects of mixed variable types? One approach is to group each kind of variables together, performing a separate cluster analysis for each variable type. This is feasible if this analysis derive compatible result. However, in real applications, it is unlikely that you separate cluster analysis per variable type will generate compatible result. So we can group the same variables, then you can go for cluster analysis. But sometime that will not be compatible. We cannot follow that approaches.

(Refer Slide Time: 23:52)

Variables of Mixed Types

- A more preferable approach is to process all variable types together, performing a single cluster analysis
- One such technique combines the different variables into a single dissimilarity matrix, bringing all of the meaningful variables onto a common scale of the interval [0.0,1.0]



A more preferable approach is to process all variable types together, performing a single cluster analysis. So in general what we have to do we have to by grouping all the variables we have to

do a single cluster analysis that will give you the meaningful result. One such technique combines the different variables into single dissimilarity matrix, bringing all of the meaningful variables onto a common scale of the interval 0 to 1. So one way to bring all different types of variables into a common scale is nothing but converting all the variables and bringing into this scale, nothing but standardization that is 0 to 1.

(Refer Slide Time: 24:36)


Variables of Mixed Types

- Suppose that the data set contains p variables of mixed type
- The dissimilarity $d(i, j)$ between objects i and j is defined as

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

where the indicator $\delta_{ij}^{(f)} = 0$ if either

- x_{if} or x_{jf} is missing (i.e., there is no measurement of variable f for object i or object j), or $x_{if} = x_{jf}$ and variable f is asymmetric binary;
- otherwise, $\delta_{ij}^{(f)} = 1$


27

Suppose that the data set contains p variables of mixed type, so the dissimilarity $d(i, j)$ between objects i and j is defined as $d(i, j) = (\sum_{f=1}^p d_{ij}^{(f)})$ divided by $(\sum_{f=1}^p \delta_{ij}^{(f)})$, for f where the indicator $\delta_{ij}^{(f)} = 0$, if either x_{if} or x_{jf} is missing that is there is no measurement of variables f for object i or object j or x_{if} equal to x_{jf} , equal to 0, then $\delta_{ij}^{(f)} = 0$ or the variable f is symmetric binary. Otherwise when the option is not there then the $\delta_{ij}^{(f)} = 1$.

(Refer Slide Time: 25:33)

Variables of Mixed Types

- The contribution of variable f to the dissimilarity between i and j , that is, $d_{ij}^{(f)}$, is computed dependent on its type:
- If ' f ' is interval-based:

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}},$$
 where h runs overall non missing objects for variable f
- If ' f ' is binary or categorical: $d_{ij}^{(f)} = 0$, if $x_{if} = x_{jf}$
 - otherwise $d_{ij}^{(f)} = 1$

The contribution of variables f to the dissimilarity between i and j , that is, $d_{ij}^{(f)}$ is computed depending on its type. If ' f ' is interval-based so $d_{ij}^{(f)} = \text{modulus of } (x_{if} - x_{jf}) \text{ divided by } (\max_h x_{hf} - \min_h x_{hf})$ where h runs overall non-missing objects for variable f . If f is binary or categorical so $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$ otherwise $d_{ij}^{(f)} = 1$.

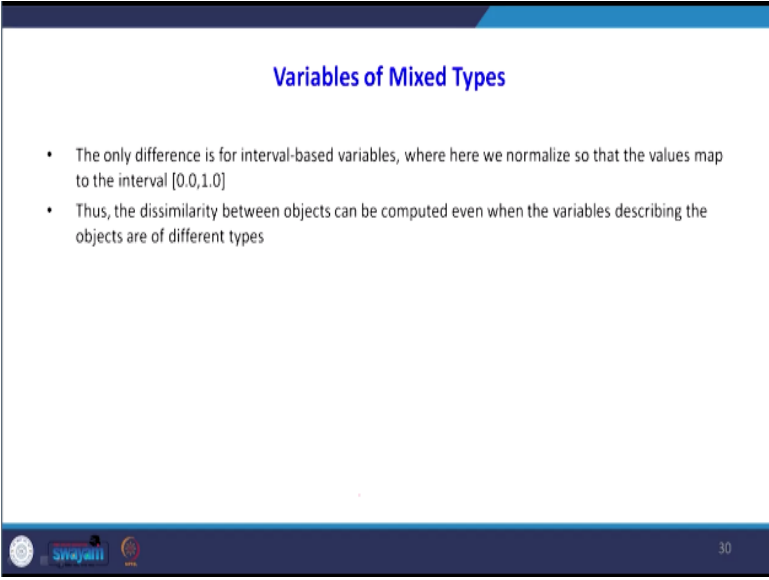
(Refer Slide Time: 26:15)

Variables of Mixed Types

- If ' f ' is ordinal: compute the ranks r_{if} and $z_{if} = r_{if} - 1 / M_f - 1$, and treat z_{if} as interval scaled
- If ' f ' is ratio-scaled: either perform logarithmic transformation and treat the transformed data as interval-scaled; or treat ' f ' as continuous ordinal data, compute r_{if} and z_{if} , and then treat z_{if} as interval-scaled
- The above steps are identical to what we have already seen for each of the individual variable types

If f is ordinal compute the rank r_{if} and z_{if} using this formula, $z_{if} = (r_{if} - 1) / (M_f - 1)$, and treat z_{if} as interval scaled. If f is ratio-scaled either perform logarithmic transformation and treat the transformed data as interval-scaled or treat f as continuous ordinal data, compute r_{if} and z_{if} and then treat z_{if} as a interval-scaled. The above steps are identical to what we have already seen for each of the individual variable types.

(Refer Slide Time: 26:58)



The slide is titled "Variables of Mixed Types" in blue text. It contains two bullet points: "The only difference is for interval-based variables, where here we normalize so that the values map to the interval [0.0,1.0]" and "Thus, the dissimilarity between objects can be computed even when the variables describing the objects are of different types". The slide has a blue header and footer. The footer contains logos for "University of Delhi" and "Department of Computer Science" on the left, and the number "30" on the right.

Variables of Mixed Types

- The only difference is for interval-based variables, where here we normalize so that the values map to the interval [0.0,1.0]
- Thus, the dissimilarity between objects can be computed even when the variables describing the objects are of different types

The only difference is, for interval-based variables where here we normalize so that the values map to interval 0 to 1. Thus, the dissimilarity between objects can be computed even when the variables describe the objects are different types. The summary of this lecture is, I have explained how to handle the different types of variables. For example, if it is a categorical variable or ordinal variable and ratio-scaled variable and variables of mixed type, how to find out the dissimilarity matrix.

So what I have done in this lecture, I have taken one example in that using that example I have explained how to find out the dissimilarity matrix. But only for the last that variables of mixed type I have explained only the theory portions. The next class I will take another example which are mixed in nature then I will tell you how to find out the dissimilarity matrix. Along with that, we will start a new topic in the next lecture that is a K means algorithm.