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

Lecture – 53
Cluster Analysis - V

In our previous class, I have explained how to handle different types of data for doing cluster analysis. I have started some theory, there is a mixed data type, how to handle that kind of data.

(Refer Slide Time: 00:38)
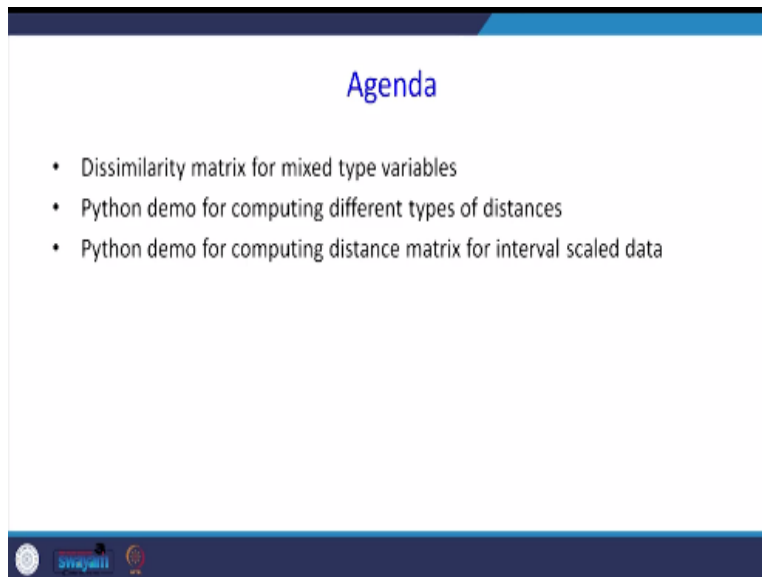
The agenda for today's lecture is how to find the dissimilarity matrix for mixed type variables. And we will do python demo for computing different types of distances which I have explained theory in my previous lectures and also I will tell you python demo for computing distance matrix for interval scaled data.

(Refer Slide Time: 01:01)

Example

Consider the data given in the following table and compute a dissimilarity matrix for the objects of the table

Now we will consider all of the variables, which are of different types

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 an example. This is a mixed type dataset. So consider the data given in the following table and compute a dissimilarity matrix for the objects of the table. Now we will consider all of the variables which are different types. So there are three type of dataset one is categorical, ordinal and ratio-scaled. If this kind of mixed data type is there how to use these kind of dataset for finding dissimilarity matrix and giving as an input for the cluster analysis.

(Refer Slide Time: 01:34)

Example

- The procedures we followed for test-1 (which is categorical) and test-2 (which is ordinal) are the same as outlined above for processing variables of mixed types

- For categorical variable - $d(i, j) = \dfrac{p - m}{p}$,

- For ordinal variable - $z_{if} = \dfrac{r_{if} - 1}{M_f - 1}$

- For interval scale variable - $d_{ij}^{(f)} = \dfrac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$,

The procedures we followed for the test-1 which is categorical data and test-2 which is ordinal data are the same as outlined above for processing variables of mixed type. So what we have done, if it is a categorical variable type we have used this p – m divided by p so where the p is number of variables where m is number of matches, that we have discussed in our previous class.

If it is an ordinal data, we have to standardize into 0 to 1 by using this formula z if = r if that is the current rank – 1 divided by maximum rank – 1. So this after converting into 0 to 1 scale then you can use our simple Euclidean method for finding the dissimilarity matrix. For interval scale variable so d if for f = x if – x jf modulus divided by maximum value of h xf – mean value of h xf.

(Refer Slide Time: 02:33)

First normalizing the interval scale data. First, however, we need to complete some work for test-3 which his ratio-scaled. We have already applied a logarithmic transformation to its values. Based on the transformation values we got 2.65, 1.34, 2.21 and 3.08 obtained for the objects 1 to 4, respectively. We let maximum value of hx is from this among these values the maximum value is 3.08 and the minimum value is 1.34. The normalized value in the dissimilarity matrix obtained in the example and solve for ratio data by dividing each one by this difference that is 3.08 – 1.34 that is 1.74.

(Refer Slide Time: 03:21)

Dissimilarity matrix for test-3

- This results in the following dissimilarity matrix for test-3:

Object Identifier	Ratio scaled Data (x)	Log (x)
1	445	2.65
2	22	1.34
3	164	2.21
4	1210	3.08

$$\begin{bmatrix} 0 \\ 0.75 & 0 \\ 0.25 & 0.50 & 0 \\ 0.25 & 1.00 & 0.50 & 0 \end{bmatrix}$$

- For 1 and 2 = (2.65- 1.34)/(3.08−1.34) = 0.75

This results in the following dissimilarity matrix for test-3. So what happened there was a object identifier was there, there was a ratio-scale was there, I am calling it as x. So we have taken log of that value 2.65, 1.34, 2.21. So we have to standardize this one for that purpose for example 1 and 2 we use this formula that is this formula to standardize. So here 2.65 this value minus this value divided by the maximum value and minus minimum value.

So when you divide this, this is a 0.75. So 2, 1 this was the distance. Similarly, if you want know 4, 1 so 4, 1 so the difference is 2.65 – 3.08 whole divided by this value, that is 1.74 will give you this value. The same way the standardization is done for all the cells.

(Refer Slide Time: 04:23)

dissimilarity matrices for the three variables

- We can now use the dissimilarity matrices for the three variables in our computation of Equation $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{max_h x_{hf} - min_h x_{hf}}$,
- For example, we get d(2,1)= (1(1)+1(1)+1(0.75))/ 3 =0.92

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

Dissimilarity matrix for categorical

$$\begin{bmatrix} 0 \\ 1 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Dissimilarity matrix for ordinal

$$\begin{bmatrix} 0 \\ 0.75 & 0 \\ 0.25 & 0.50 & 0 \\ 0.25 & 1.00 & 0.50 & 0 \end{bmatrix}$$

normalize the values in the dissimilarity matrix for ratio data

Now let us consider dissimilarity matrices for all three variables. What are the three variables? One is categorical, ordinal and ratio data. So we can now use the dissimilarity matrices for the three variables in our computation of equation, so d if for f = x if – x jf divided by max h x hf – minimum h x hf.

(Refer Slide Time: 04:51)

For example, we got between d 2 1. So this is the location, 2, 1. I will explain how we got 0.92. So how we got this one is, so there are three variables is there; for dissimilarity matrix it is 1, for categorical variable, for ordinal variable, for 2, 1 portion is 1, for ratio data it is a 0.75. So we will find out the weighted mean. So weight is we are giving equal weigh for all kind of dataset. So 1 into 1 + 1 into 1 + 1 into 75 so total sum of weight is 3, so 0.92 that is why we got this value.

Similarly, each element you can do that one, for example here 1 into 1 + 1 into 0.5 + 1 into 0.25 divided by 3 so we will get this value. So this matrix is a resulting dissimilarity matrix attained for data described by the 3 variables for the mixed type. So this dissimilarity matrix is given as a input for doing the cluster analysis. So what happened this is a combined matrix, combined dissimilarity matrix for all three kind of variables, so what are that variables it is for categorical, ordinal and ratio data.

(Refer Slide Time: 06:12)

Interpretation

- If we go back and look at Table of given data, we can intuitively guess that objects 1 and 4 are the most similar, based on their values for test-1 and test-2
- This is confirmed by the dissimilarity matrix, where d(4,1) is the lowest value for any pair of different objects
- Similarly, the matrix indicates that objects 2 and 4 are the least similar

If you go back and look at the table of the given data we can intuitively guess that the object 1 and 4 are most similar, based on their values for test-1 and test-2. You see that, because if you look at this one, this is 0, this is 0, this is 0.25. In the given matrix, for example, for, say for categorical data it is 0, for ordinal data it is 0, it look like the position of 4, 1 for both the variables is seems to be very similar. But what is happening here if the third where it is ratio data it is 0.25.

But when you find the average of 3 weighted average by looking at the dissimilarity matrix for categorical and ordinal look in the position of 4, 1. So this seems to be very similar because it is close to 0. But what is happening the 0.25 it is not very close to 0, so this can be verified by when you look at this one, so among the all the dataset in the combined the resulting dissimilarity matrix the value of 0.0 it is very similar, so the position of 4, 1 is very similar to each other, that is the point.

This is confirmed by the dissimilarity matrix, where 4, 1 is the lowest value for any pair of objects. Similarly, the matrix indicates that the object 2, 4 are the least similar. When you look at this 2, 4, so the highest value is 1. When you go there, here also 2, 4 here also it is 1, here also 1, here also 1 so in the resulting dissimilarity matrix also this seems to be 1 so highly dissimilar.
(Refer Slide Time: 07:55)

Distance Measurement using python - Euclidean Distance

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

Now, we will go to the distance measurement using python. In my previous class, I have explained how to find out Euclidean distance, Manhattan distance, Minkowski distance. So this is an example for Euclidean distance, the formula for finding Euclidean distance is xi1 – x01 whole square + xi2 – x02 whole square + xip – xjp whole square. If there are only two variable the distance formula is xi1 – x01 whole square + xi2 – x02 whole square, square root.

(Refer Slide Time: 08:30)

Python Demo for Euclidean Distance

```
In [1]: import scipy
        from scipy.spatial import distance

        #Euclidean Distance

In [2]: import numpy as np
        a = [1,2,3]
        b = [4,5,6]
        dst = distance.euclidean(a,b)

In [3]: dst
Out[3]: 5.196152422706632
```

But how to use python command, I brought the screenshot of that, for that import scipy, from scipy.spatial import distance, so how to find the Euclidean distance? So import numpy as np, so a and b there are two; a is one point where 1, 2, 3; the b is another point 4, 5, 6. If you want to

know the distance Euclidean distance between a, b so we have to write dst = distance.euclidean a, b. So when you type dst, so this is the our Euclidean distance between point a and b.

(Refer Slide Time: 09:07)

The next, the distance is Minkowski distance. So the Minkowski distance is, the combination of both Manhattan distance and Euclidean distance. So d i, j = xi1 – x01 to the power p + xi2 – x02 to the power p and so on plus xin – xjn to the power p whole to the power 1/p. So what is happening, this is small p. If p = 1, then it is a Manhattan distance. If the p = 2 it is an Euclidean distance. Let us see how to find out this Minkowski distance using python.

(Refer Slide Time: 09:46)
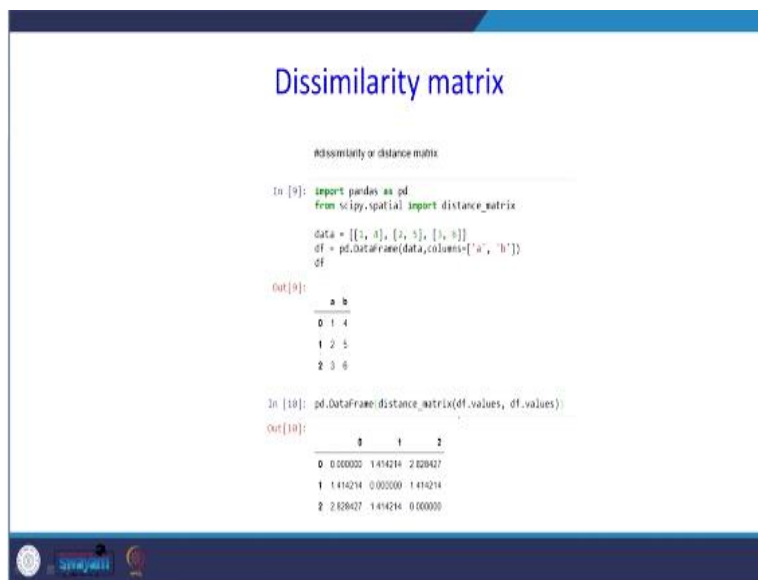
So I have taken two points, Minkowski distance let us see 1, 0, 0, 0; 0, 1, 0. So the comma 1, this represent that we are finding Manhattan distance. So when you enter this the Manhattan distanced is 2. So this same dataset if you type 2, you will get Euclidean distance because the p = 2 will get you formula for the Euclidean distance that is 1.14. So this was our another example just to verifying this 1, 2, 3; 4, 5, 6, 2 so this will represent our Euclidean distance.

Suppose if you take 1, 2, 3 and 4, 5, 6 the p value can be 3 also; if it is 3 then Minkowski distance is 4.32. This formula distance.minkowski 1, 2, 3; 4, 5, 6, 2 this data previously we have used the distance between by using this formula distance.euclidean distance a, b we got 5.19. So by using Minkowski formula also the Minkowski formula the same dataset if you type 2 you will get the same answer. So the distance.minkowski between the two points 1, 2, 3 and 4, 5, 6, 3 where the p = 3 so this our Minkowski distance.

(Refer Slide Time: 11:07)

Now, can we explain how to find out dissimilarity matrix? So for that you have to import pandas as pd from scipy.spatial import distance_matrix. So there are 1, 2; there are 3 points; 1, 4; 2, 5; 3, 6. So there are two columns a, b. So pd.DataFrame data, columns equal to this one, you will get this kind of output. So if you want to know the distance between a, b so there are identifier name is 0 1 and 2. There are two variables a and b. So if you want to know the distance matrix between different identifier 0 0 is 1; 1 0 is 1.41; 2 0 is 2.84 by using this command distance_matrix df.values and df.values.

(Refer Slide Time: 12:05)

Now, distance matrix calculation for interval-scaled variables. For example, there are 1, 2, 3, 4, 5, 6, 7, 8 there is a 8 persons, we can say object identifier, take 8 people, the weight is given in kilograms and the height is given in centimeters, so n = 8, so p = 8. Now how to find out the distance matrix?

(Refer Slide Time: 12:30)

So I have taken the data 15, 95; 49, 156 and so on. So there is a labels A, B, C, D, E, F, G, H. So the data frame is so Weight and Height. So we got this table. So weight and height is there, these are different identifier.

(Refer Slide Time: 12:54)

Suppose if you want to know the distance matrix, so for that pd.DataFrame distance_matrix you take these values. So we are getting the distance matrix between A and A 0, B and A. So you see that the diagonal value 0 because it is a ReplicaOf because the distance between F and F is 0; G and G 0, H and H is 0. So what is happening between A and; B and A it is 69.83; A and B also 69.83 just a mirror value.

(Refer Slide Time: 13:29)

We will round into 1 decimal. So by using matrix; distance_matrix.round decimals = 1, so out = none so we got this matrix. This will run with the help of python.

(Refer Slide Time: 13:41)
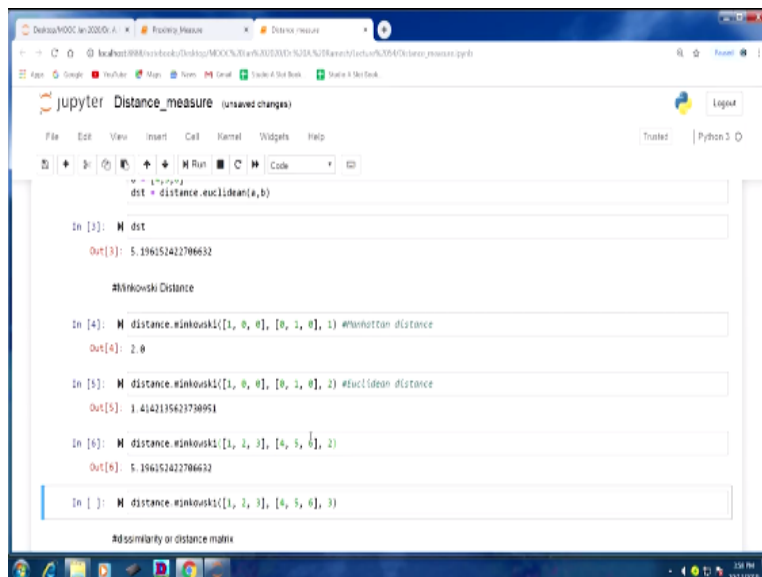
Now let us see how to use python to find the distance between points. So import spicy from scipy.spatial import distance, I will run that one. Now let us know how to find out Euclidean distance. So import numpy as np, there are two point a, b. The position of point a is 1, 2, 3; the position of point b is 4, 5, 6. If you want to know the distance so dst = distance.euclidean then if you want to know the what is the distance so will let us display this, so distance is 5.19 that is the between point a and b the Euclidean distance is 5.19.

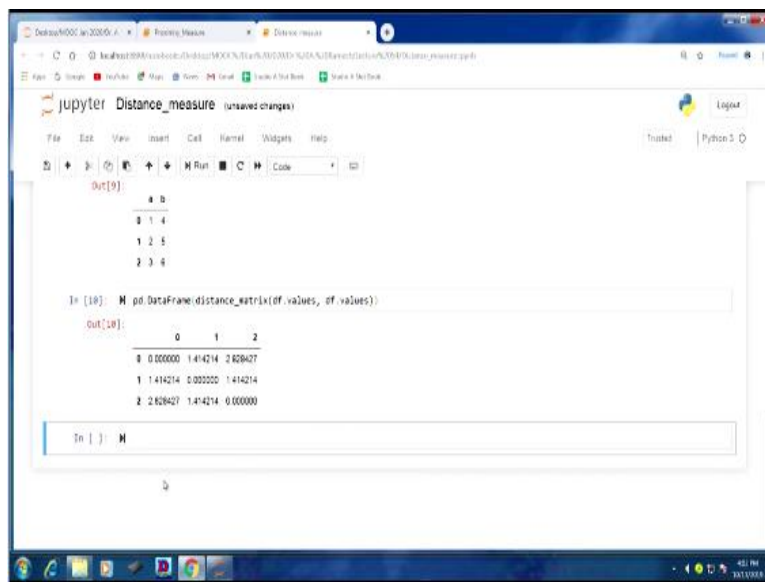**(Refer Slide Time: 14:27)**

Now, let us know how to find out the Minkowski distance. For that, the Minkowski distance is calculated by using this function distance.minkowski. This is position of point A and position of point B. Then if you use 1 you will get a Manhattan distance. So this is Manhattan distance.

Instead of 1 if use 2 you will get here. Instead of 1 if you get 2 you will get a Euclidean distance. This is 1.41. We already got 5.19 as Euclidean distance even the Minkowski function you can get; you can verify that answer.

What happened here I gave distance.minkowski, I taken the same point that is 1, 2, 3 and 4, 5, 6 this one I used 2, so number 2 is used to get the Euclidean distance. Number 1 is use to get Manhattan distance. So we got to see that this also 5.19 when in the function Minkowski for use p, otherwise by using distance.euclidean function we got 5.19, so both are same.

(Refer Slide Time: 15:43)

Now I will explain how to find out the dissimilarity or distance matrix; import pandas as pd from scipy.spatial import distance_matrix. The data equal to 1, 4; 2, 5; 3, 6 so there are three dataset for two variables. If you want to know the, the distance between a and b; now we have three dataset for two variables. If you want to know the distance matrix, so pd.DataFrame in bracket distance_matrix df.values, df.values, so you will get this is distance matrix. So 3 variables and 3 dataset. So this, the distance matrix between 0 and 0 is 0; between 1 and 0 is 1.41, between 2 and 0 is 2.82.

(Refer Slide Time: 16:42)

Now, we will go for a distance calculation for that import pandas as pd import numpy as np (())
(16:50) data matrix is given like this. So there are two variables. There are 8 persons. A, B, C, D,
E, F, G, H. Suppose if you want to know the distance matrix so pd.DataFrame distance_matrix
you follow this command, the distance matrix is this one, right between A and A is 0; between B
and A is 69.83; between C and A it is 2.00.

(Refer Slide Time: 17:22)

So if you want to have 1 decimal with the rounding of 1 decimal we got this distance matrix. So
this distance matrix is given as input for cluster analysis. In this lecture I have explained how to
find out the dissimilarity matrix for mixed type of variables. What is the meaning of mixed type

of variables? When we got for a cluster analysis the dataset may be combination ordinal, interval and ratio data.

When these three types of data come together how to find out the dissimilarity matrix? That I have explained. Then, by using python I have explained how to find out the distance. So I have explained how to find out the Euclidean distance. Then I have explained how to find out the Manhattan distance and then I have explained how to find the Minkowski distance. At the end I have explained with the help of python how to find out the distance matrix for the interval-scaled data, because that distance matrix can be used as a input for our cluster analysis. Thank you.