Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 10

Lecture – 48

Cluster Analysis: Basic Concepts and Definitions

Hello friends, welcome to the course Multivariate procedure with R. Now in this lecture we are going to begin with a new topic which is Cluster Analysis. So, you can recall that in the last couple of lectures we had talked about the discriminant analysis. And when I introduced the discriminant analysis at that time, I had introduced that there are two tools to classify an observation. In one condition, the categories in which the observation has to be classified they are known. So, for example, that is what we have achieved in the case of discriminant analysis.

Whereas in some cases we do not know that what are the different categories. We simply have a population, population of observations and even we do not know that how many categories are inside that data. So, under such a case when an observation come how are you going to classify it into one of the categories. So, this is achieved by cluster analysis.

So, this is the main difference between discriminant analysis and cluster analysis that in discriminant analysis we know that what are my categories in which the observation has to be classified whereas in the case of cluster analysis we do not know that what are the categories in which my observation is going to be classified. Well, I can tell you that in cluster analysis we are mainly dependent on the software because the data is usually so huge that it is very difficult to manage it manually. So, the way I am going to conduct this lecture on cluster analysis that in this lecture I will try to give you some basic concepts and background that what is really happening behind the curtain of cluster analysis. What are the basic fundamentals based on which we try to work and you have to just understand what is happening and then I will try to illustrate them in the R software where you will see that it is very easy to implement but it is more important to understand what is happening. So, let us begin our discussion on cluster analysis and we try to discuss about the say basic concept and definition in this lecture.

So, we are going to talk about the basic concepts and definition in cluster analysis. So, just consider a situation where we have a very big data set, a large data sets. So, dig out the information from such data set we need to classify and group the data into similar groups, right. For example, if you have say students in a college where the classes are from class 1 to class 10, so you understand that there is a significant difference between the ages of students of class 1 and class 10 but the ages of class 1 and class 2 may not be that much different. So, now you have got a population where you do not know that how many students are there from class 1, class 2 and class 10.

So, you would like to divide the entire population for example, with respect to the age of the students in similar groups. Similar group means that all the students in that class, in that group should have similar ages, more or less way which are very close to each other. So, for that we need a classification scheme. So, a convenient method for organising a large data set so that retrieval of information may be made easier can be achieved through a classification scheme which describes the pattern of similarity and differences among the objects under investigation. So, the next question comes here, how are you going to quantitatively compute this similarity? With our eyes I can always say okay these two people are similar but how are you going to quantify because the software understands only the numbers and in statistics unless and until you have the proper numbers, we cannot do a good job.

So, now if you try to see here at this slide what do you really understand and suppose these numbers are written on a very big wall, can you understand anything out of this? You can simply see that there are some number 5, 4, 4, 3, 1, 5, 2 and so on but it is very difficult to understand. So, this is a replica of what is happening in the clustered analysis with that we have got this type of data and we try to classify them into similar groups. For example, you can see here basically I have written here the numbers like 1, 2, 3, 4, 5. Now you would like to classify them into different groups. Now depending on how do you define the similarity there can be 5 groups 1, 2, 3, 4, 5 containing all the numbers or there can be one group containing 1, 2 and another group containing 3, 4, 5 and so on.

So now the next question is how many groups are you going to achieve? So, these types of questions which are going to be answered now in this topic. So, just to repeat the basic concept what we have understood that in the discriminant analysis this is a technique for grouping the individual or objects into known groups and that those groups are known a priori in which the observation has to be classified. But in the case of cluster analysis this is a technique for grouping the individuals or objects into unknown groups and data set that contains only data point is available and is without the class labels. So, now for

example, if I want to take an example suppose if I want to classify the people on the basis of their economic status in groups then we have groups like lower class, middle class, upper class which are known group. So, here I can use the discriminant analysis.

But in another example suppose there are some diseases which looks similar. For example, if somebody is coughing, the coughing can be due to different reason. There can be simply cold and cough or there can be tuberculosis or there can be some other disease. So, these types of diseases which have similar means symptoms, similar causes or conversely different disease from the same cause can be there. So, now if you want to understand the treatment for such a disease, we need to classify the symptoms in a proper group to diagnose the disease, right, whether this coughing is due to tuberculosis or due to cold or something else.

But these groups are unknown. So, in such a cases we use the cluster analysis, right. And this cluster analysis has got different names in different subjects. For example, in biology this is called as numerical. And this cluster analysis has got different names in different subjects in biology.

This is called as numerical, taxonomy in psychology, this is called as skew analysis, in artificial intelligence which is very popular nowadays, this is called as unsupervised pattern recognition and the most common generic name from statistics is cluster analysis. So, if you try to understand this graphically you can see here this is here some data whose scatter diagram is here. But now we try to divide them into different groups based on their different properties. So, I can create here this group of green dots, blue dots and here red dots. So, these will be three clusters and I can say that based on certain criteria that the observations within these clusters are more homogeneous or they are similar to each other or say more similar to each other.

So, this cluster analysis is essentially a data mining tool and the emphasis is on the tools which can handle the large data sets. And it is very empirical, highly empirical. You will see when I try to take a data set then you will see that for example, that how many groups are going to be there. This depends on the objective and experience of the experimenter, right. And there are different methods to find out the similarity, different methods to find out the grouping, etc.

So, different methods lead to different grouping and both in number as well as in the content, right. And since the groups are not known in advance a priori, so it is usually

difficult to judge whether the results make sense in the context of the problem studied or not. But still, we have to do something so that we can decide that whatever we are doing is correct or not. So, in cluster analysis first we have to understand the choice of variable. So, initial choice of a particular set of measurement is used to describe how individual can be constituted in a cluster, right.

That you have to decide that what is your criteria based on which you can say that the different observations inside the cluster are going to be similar. And this choice is the choice of the investigator or the statistician, right, such that the chosen variable is relevant to the required type of classification. That is your choice. So, that is why it is highly empirical once again. And when we are trying to do the clustering then different variables may have different type of measurement units.

So, when we are trying to classify then it is not fair to classify the heights and weight of the version together. So, we would like to make them unitary and for that we would like to do the standardization. And we already have understood what is standardization which can be obtained in the R software very easily using the command scale with different types of options, right. So, variables may have different units, their nature may be different quantitative or say qualitative, they may be a, they may have been measured on different scales. For example, weight is measured in kilogram, height is measured in meters, anxiety level is measured on a four-point scale, etc.

So, this standardization process makes them, makes the observation unitary and then we can employ the clustered analysis over it. I agree that this may dilute the difference between groups on variable which are the best discriminant. But anyway, we have to solve the problem, right. And another option is that we try to use the concept of similarity coefficients, right. And then the next question comes what are these similarity coefficients? So, in very simple way if you try to recall what is correlation coefficient.

Correlation coefficient is simply trying to measure the degree of linear relationship or the strength of the linear relationship between the two variables x and y. So, similar type of idea is implemented and we try to define here a similarity coefficient so that we can measure the degree of similarity among different units or different observation inside the cluster. And if the similarity coefficients are close to each other then we can say that the observation will belong to the same cluster. So, this similarity coefficient actually indicates the strength of relationship between two objects given the values of a set of p variables which is common to both. And this is some function of the observed values.

And the similarity between two objects increases with the increase in the value of say sij. If there are ith and jth objects and if the value of sij is increasing that means that will indicate that the similarity between the two objects is also increasing, right. Now if you try to understand how we can define such a similarity coefficient and what type of properties it should have. So, this sij is defined such that sij = sji that it is a symmetric function. That means if object number 1 is similar to object number 2 then object number 2 will also be similar to the object number 1.

And mostly it is convenient for us to understand the value of sij if it is a bounded function and it is more convenient if it is lying between 0 and 1. Otherwise it can take any other value but it is easier to compare the values which are lying between 0 and 1. So, we prefer to have a similarity coefficient which takes value between 0 and 1. And it scales such that the upper limit is always 1. Once you have a similarity measure, the contrary or the contrast will be dissimilarity measures.

So, this is just complement to the similarity measures. So, we can say either we are going to classify the observation which have got similar coefficients or we can say that we are going to classify the observations into different clusters which have different dissimilarity coefficients. So, associated with each similarity measure which is lying between 0 and 1 indicated by sij there is also a dissimilarity coefficient which is defined here as a dij = 1 - sij. So, that dij will always lie between 0 and 1. So, what will happen if I say that here sij = 0.95. That means the objects i and j they are 95% similar. What if I say here dij = 0.93 that means the objects are 93% not similar, right. So, this is how we try to interpret it. So, the dissimilarity between the two object increases with an increase in the value of dij.

And obviously an object has the maximum similarity to itself. So, that is why the maximum similarity with itself is indicated by sii which = 1 and dii = 0. For example, you are more similar to yourself, right. Okay, so now the question comes here how are you going to define this sij or say dij. Now we have understood that there is a close relationship between sij and dij.

So, if I know any one of them means I can find the other value also. So, in the literature several dissimilarity coefficients have been defined which are the function of here say xi and xj where xi is a say ith observation on a p-variate vector and xj is the jth observation on a p-variate vector, right. And then we try to compute different types of functions to compute this dissimilarity coefficients. And some of them I have explained here. For example, the first one is here is Euclidean distance.

(1) Eucledian distance $d_{ij} = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$ (Most popular)

(2) City block $d_{ij} = \sum_{k=1}^{p} | x_{ik} - x_{jk} |$

(3) Canberra metric $d_{ij} = \sum_{k=1}^{p} \frac{| x_{ik} - x_{jk} |}{(x_{ik} - x_{jk})}$

(4) Angular seperation $d_{ij} = \frac{\sum_{k=1}^{p} x_{ik} x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2 x_{jk}^2}}$

(5) Mahalanobis distance $d_{ij} = (x_i - x_j)' S^{-1} (x_i - x_j)$

So, if you try to see I am simply trying to compute the distance between the ith and jth observation by summation k goes from 1 to p, xik - xjk whole square and this is one of the popular measures of dissimilarity. Similarly, there is another measure here which is called as city block difference which is here like this summation k goes from 1 to p and the summation over the absolute value of the difference between xi and xj that is absolute value of xik - xjk. There is another measure here which is called here say Kanbira matrix which is defined here like this, summation of ratio of the absolute difference between xik xjk and divided by the simple difference that is xik - xjk. So, this Euclidean distance and city block they will always be greater than 0 but this dij can be negative also under the Kanbira matrix, right. Similarly, there is another here angular separation which is defined here like this dij = summation k goes from 1 to p, xik xjk divided by square root of summation of xi x square ik and x square jk, right.

And similarly, there is another popular method which is the Mahalanobis distance. If you remember I had introduced it during the Hotelling t square statistics which was indicated there by here delta square quantity and based on that we are trying to define here xi - xj transpose capital S inverse xi - xj and capital S is the pooled Poisson group covariance matrix estimator, right. So, these are different say this measures and different measures will give you different results in general, right. So, and then there are several dissimilarity measures for the grouped observations also. So, they have been proposed based on the between group distance measures which are obtained by substituting the group means for p variables in the inter-individual measures such as Euclidean distance or city block distance.

For example, if I have the observations in say two groups group A and here group B and then we try to find out here this x bar A which is the mean of observation in each on each of the variable say 1 to p and the same thing I try to do it here in the in obtaining the here the value x bar B which is the sample mean on each of the p variables, right in the second

group. And then we try to find out the square depth distance which is measuring the distance between two groups like here like this. So, dAB now here is the square root of k goes from 1 to p, x bar ai - x bar bi. So, if you try to see that first we are trying to group the observation and then we are trying to find out the distance, right. So, similarly there is another option here which is here the Mahalanobis distance which is given by here D square = x bar a - x bar b transpose W inverse x bar a - x bar b where W is the p cross p pool within group covariance matrix for the groups A and B, right.

$$d_{AB} = \sqrt{\sum_{k=1}^{p} (\bar{x}_{Ai} - \bar{x}_{Bi})^2}$$

Mahalanobis distance $D^2 = (\bar{x}_A - \bar{x}_B)' W^{-1} (\bar{x}_A - \bar{x}_B)$

And then we have different types of way by which we try to do the grouping. For example, we have here the single linkage, complete linkage and group linkage measures. For example, in the case of single linkage measures we try to find out the nearest neighbour distance that is the distance between the closest members one from each of the group and this form the basis of the clustering technique which is known as single linkage. For example, if you try to see here this is my one group, this is my another group and then I am trying to consider this distance between the two closest observations between the two groups and it is indicated by here dAB which is the concept behind the single linkage, right. And similarly, we have here the complete linkage which is just opposite to the single linkage where we try to find out the farthest neighbourhood distance that is the distance between the most remote pairs of the individual one from each group, right.

For example, if you try to see here, I have this one group here, one group here and I am trying to find out the maximum distance between the two farthest point which is indicated by here dAB. So, this forms the basis of clustering technique which is called as complete linkage and it is just opposite to the single linkage method, right. And then we have a group average also. So, in the group average what we try to do that we try to consider the average of all inter-individual members of those pairs of individual when the members of the pair are in different groups. And this forms the basis of group average clustering technique and we have for example; this k means clustering.

So in the k means clustering what we try to do that we choose the number k, number of k clusters. So, in order to achieve it in the first step we randomly assign the item to the k clusters, then we calculate a new centroid for each of the k clusters, then we try to calculate the distance of all items to the k centroids, then we try to assign item to the closest centroid and we try to repeat this until the cluster assignments are stable, right. I

will try to show you when I try to do it but here with a very small example, I would try to show you what do we mean by this linkage schemes and I will try to take an example of single linkage and then based on that you can also develop the concept for complete and group linkage. So, suppose I try to take here a simple example where I have got here 5 observations and I try to create here a distance matrix that this distance has been computed from say any of that technique but symbolically I have written their values that these are the values, they can be Euclidean distance or they can be something else. So, now what we try to do here in the single linkage we are trying to find out the distance between the two closest members.

So, first of all, we try to see what is the smallest entry in this matrix. So, you can see here this entry here is 2, right. So, what we try to do here then we try to join the observation number 1 and 2 and this forms our two-member cluster. And then we try to find out the distance between this cluster which is from two observation 1 and 2 and remaining observation 3, 4, 5. You can see here now if you try to find out here that if you try to find out this distance and try to create a new matrix.

So, the distance between this new cluster which is containing here 1, 2 observation and 3 that is going to be the minimum value between the distance between 1 and 3 and distance between 2 and 3. So, you can see here the distance between say here 1 and 3 and 2 and 3 here is like this. So, 1 and 3 here is actually 6 which you can see here and distance between 2 and 3 which is here say here 5. So, you can see here the minimum between 5 and 6 is here D23 = 5. And similarly, we try to compute the distance between D12 and 4 which is the minimum distance between D14 and D24.

So, if you try to see here this distance D14 is given here as say here 10 and D24 is given here as say here 9. So, the minimum between 10 and 9 is the D24 which is here 9. And similarly, if you try to see the third value here the distance between 12 and here 5 which is the minimum value between D15 and D25 which are here 9 and 8. 9 and here 8 and the minimum value among these two between these two values is here D25 = 8.

And then I try to create here a new matrix. So, you can see here now I have clubbed 1, 2 together and 3, 4, 5 are different. And now I try to make these values here 5, 9 and 8 here like this and the remaining value will come from this above matrix. I simply try to copy them there. Now I try to repeat it. Now in this new matrix I try to find out the which is the smallest value.

So, the smallest value here is 3. So, I can now combine here this here, fourth and here fifth observation together. So, I try to make here now one more cluster of 2 units 4 and 5, right. And I try to combine here. So, now I have here this combined observation in 1, 2, 4, 5 and 3. Now we try to find out the minimum value between the distance 1, 2 and 3 which is the minimum value between D13 and D23 and you can see from the table it is coming out to be D23 which is here the value here 5.

And similarly, I try to compute here the values, the distance between the 2 clusters 1, 2 and 4, 5 and this distance between 4, 5 and 3 and they come out to be here 8 and 4. And then we try to create here a new matrix here, this D3. So, you can see here this I have computed here 0, 5 and here 8 and this is here 0. So, and then now you can see here after this you again try to find out the minimum entry in this matrix which is here now here 4.

So now this 3 and here 4, 5 they will be joined together. So, this new cluster will become here 3, 4, 5 in the group 2 and 1 cluster will remain as here say 1 and 2. Now if you want to go it further then there will be only one cluster containing all the observations 1, 2, 3, 4 and 5. And now the same process I can indicate by this figure also which is called here as a dendrogram. So, if you try to see here these are my stages here say partition 1, partition 2, partition 3, partition 4, partition 5 and we try to see here this observation. So, we try to take the observation number 1 and 2 which have been combined into one group at a stage 1, right.

And this has been indicated here like this. All the observation they are all alone. Then in the next stage I try to combine them 1, 2 and then 3, 4, 5 are separated. Then I try to move further and then I try to combine this 4 and 5 together in the next stage which is here in the third stage. So, you can see here in the third stage I am trying to have this 1, 2 and 4, 5 together and 3 individual. And then if you try to move here further then at this stage here say here P4 what are you trying to do here? That you are trying to have here these many observations and finally at this 5 you are going to have here all the observations together, right.

So, this is how you can do it here. Now similarly all other linkage schemes can also be demonstrated without any problem, right. So, now let us come to an end to this lecture and you can see here we have understood the basics of the cluster analysis in this lecture and these concepts are going to help you when we are trying to implement it over a bigger data set in the R software. So, my request is that you please try to understand these basic concepts and the way I have shown you the single linkage scheme on the same data

set you try to compute the distance matrix and try to classify them using the complete linkage. And I will try to show you these differences in the R software. So, you try to revise your concept, understand them, try to take a book, try to read from there and I will see you in the next lecture till then goodbye. Thank you.