**Multivariate Procedures with R**

**Prof. Shalabh**

**Department of Mathematics and Statistics**

**IIT Kanpur**

**Week – 11**

**Lecture – 50**

## Hierarchical Classification and Analysis with R

Hello friend, welcome to the course Multivariate Procedure with R. So, you can recall that in the last two lecture we had talked about the cluster analysis and in the last lecture I had given you the basic concept about the hierarchical clustering. Now in this lecture we are going to implement those rules and those concepts inside the R software using a very small dataset. Well, why small dataset? Because in this lecture I want to explain you how the things look and how do they work. So, if you have a small dataset then you can actually see that what is really happening, but if you have a big dataset then you will get only the outcome. So, in order to understand the outcome of a big dataset it is very important for you to understand that what is really happening when you try to execute a command in the R software over a given dataset which you know.

So, if you try to implement the same technique on an unknown dataset then you will have confidence that okay whatever I have understood, whatever I had verified the same thing is going to happen in this case, but because the data is so huge so I cannot see from my eyes, but I can understand the graphic and the results correctly and I can make the correct statistical inference. So, with this objective let us begin our lecture. In this lecture we will consider a small dataset so it is very important for you to understand what I am doing and what is the interpretation and in the next lecture I will try to take a bigger dataset, but there I will not have an opportunity to explain you in more clarity that what is happening. So, this lecture is going to help you in understanding the next lecture with a bigger dataset.

So, let us begin our lecture and try to understand this implementation in the R software. So, in this lecture we are going to understand the hierarchical classification and its analysis with R software. So, there are three packages which are depending on what you

want to use in R to implement the hierarchical clustering. One package here is say hclust. This is found in the stats package and this is used for agglomerative hierarchical clustering.

We have another command here you will see agnes which is found in a package whose name is cluster and this is also used for agglomerative hierarchical clustering. So, another command here is diana, d-i-a-n-a and this is also found in cluster package and this is used for divisible hierarchical clustering. And then we also have an option to find this agglomerative coefficient and divisible coefficient in this agnes and diana function respectively. Actually, these coefficients measure the amount of clustering structure found in the data and these are actually computed. So, they are the coefficient.

So, you will see that there are couple of things which I would like to illustrate during the implementation of data in the R software for this concept. And these coefficients values if they are closer to 1 then this suggests a strong clustering structure. This is how you have to interpret it and I will try to show you how are you going to do it. So, in order to understand the contents in the lecture I would request you that you try to please install these three packages cluster, factoextra and dendextent using this command install.packages.

And sometime you will see that in many books and different places they are also using the package tidyverse. So, that if you wish you can also install because that may be needed for further data manipulation. But anyway, we are going to use only these three packages. So, I request you that you please install them in your computer or laptop and you just load them so that it is comfortable for you when you try to understand this lecture. So, first I try to take here different type of commands which are used and first you try to understand them and then I will try to show their output in the R software.

I will try to show it to you. So, first command here is hclust. So, this is about the hierarchical cluster analysis on a set of dissimilarities and method for analyzing it. This is for use for using in this job. So, the command here is hclust.

And then the argument here is here d which is a dissimilarity structure as produced by the distance function that you will see later on that distance is a command to measure the distance of the objects. And then you have to decide for the method that what type of method you are going to use. For example, I had described about single linkage method, complete linkage method, but there are several other methods also. So, I have not

explained you here, but they are not difficult. If you wish you can use other type of methods also and you can see here, we have different types of methods here ward method of coo type, single linkage, complete linkage, average linkage, mcquity, median method, centroid method and so on.

And then you have to give here about means option for here numbers which is equal to here NULL or a vector which is the length size of d. So, that I will try to show you how are you going to use it. And similarly, there is another command here agnes. So, this computes the agglomerative hierarchical clustering of the data set and this you are going to use as agnes inside parenthesis you have to write here x which is a data matrix or a data frame or this similarity matrix depending on the value of this argument which is I will try to show you which is the next option. So, this here is a logical flag and if this is TRUE the default is actually default for this command or this similarity objects then x is assumed to be a dissimilarity matrix and if this is equal to FALSE then x is treated as a matrix of observation by variables.

After that you have to define here metric which is Euclidean method and then you have to define here the method which is here average and so on. So, I will say that beside these things here there are many more options, but I would strongly request you and recommend you to look into the help menu. Similarly, this here metric this is the characteristic specifying the metric to be used for calculating the dissimilarities between observations. And at this moment we have two options Euclidean and Manhattan. So, Euclidean distances are the roots of the square of the differences and Manhattan distance are the sum of absolute differences.

If x is already a dissimilarity matrix, then this argument will not be used and it is going to be ignored. The next command here is DINA and this actually computes a divisive hierarchical clustering of the data set and its usage is like this you have to use the command here diana, d i a n a, all in lower case then here x which is the same as earlier data metrics or a data frame or dissimilarity metrics or object depending on the values of this argument. And this is also here the same as earlier and this is a logical flag if this is TRUE then x will be considered as a dissimilarity matrix if FALSE then x will be considered as a matrix of observation by the variable. And similarly, here the metric also have the same interpretation which we have done earlier that it can take the value either Euclidean or Manhattan. Now I try to use this command over this artificially created data set.

We can see here I am trying to take here three variables y, x1 and here x2. So, y is value

like here replicated of 1 2 3 4 5 each equal to 4, c will be replicated value 11 to 20 each is equal to 2. So, I am trying to show you here that some values are going to be similar because in 1 to 5 if you are using here the command each that means each of the number is going to be repeated four times. And then in the x2 I am trying to take here some data vector of the character like T F T F and so on. Then I try to create here a data frame.

So, data frame is created as a datacl which is created by the command data dot frame and y, x1, x2 and then I try to save this data as a matrix. And let me call it here as d and then I try to use here the command hclust over this d. So, this d will find out the distance matrix and then hc here will apply the hierarchical clustering and then if you use the command here plot h c then it will plot the dendrogram. So, let us try to first understand how it will look like and then I will try to show it on the R console. So, you can see here this y here is rep 1 to 5 so you can see each equal to 4, you see 1 is replicated 4 times, 2 is replicated 4 times, 3 is replicated 4 times, 4 is replicated 4 times, 5 is replicated 4 times.

Then 11 to 20 each equal to 2 you can see here this is 11 11, 2 times, 12 12, 2 times 13 13, 2 times and so on and then your here x2 here is T F T F and so on. So, this is my here data c l right so this is my data frame. Now from this data frame I am just trying to show you here this is here d command dist(as.matrix(datacl)) it will give you here this thing yes it is difficult to read this value at the moment for you but I will try to show you it in the R console you can see it very clearly. The only thing is that you have to understand what it is trying to show you. So, if you try to see here suppose if I try to take here this value so this is coming to here 2 and here 3 so this is the distance between the second and third observations.

So, this is the distance right and so on all these values they are giving you the distances and you can see here this part is actually blank because this is a symmetric matrix in the distance between x and y is the same as distance between y and x right. I will try to show you but I would strongly recommend you that you try to use this command and you would say otherwise this matrix is large so the computation will roll. So, then I try to use here the command here hc and I try to do the hierarchical clustering using the command hclust over this distance matrix d and then I try to plot it here plot(hc). So, you can see here if you try to do it then it will try to give you here this type of plot but before that let me try to show you about the coefficients which I just talked right. So, if you want to use the agglomerative coefficient then this can be obtained by that first you try to do the hierarchical clustering using agglomerative approach using the command agnes over this datacl using the complete linkage method.

Method is equal to linkage and whatsoever is the outcome you try to save it in this

hcdataclagnes() this is how I have given the name so that one can understand it. Say hierarchical clustering data of cluster and then agnes method right and then after that it will give you different thing out of which I want to extract the value of the agglomerative coefficient. So, I use here that the command here dollar ac with this hcdataclagnes here and you will see here it will come out to be some value which is 0.89. So, as I told you earlier that if this value is more closer to 1 that means this clustering is pretty strong right and you can see here this is the screen shot of the same thing.

Now if you try to see here this is here the dendrogram. This dendrogram is obtained by the command here pl3 right. The so the same output which you have got here and then you have seen here command here cx = 0.5, hang = -2. I will try to show you that what are these things otherwise I would strongly request you that you please try to look into the help also and try to take different value of this parameter cxn hang and try to see what happens right.

If you try to see here this is the dendrogram which you have obtained for this data set. This data set. You can see here these two are similar, these two are similar, these four are similar, these four are similar and so on. So, that is why you have this dendrogram right and this is here the heights which I was talking earlier. You can see here this height or this height or this height.

You can see that they are different heights and you can interpret it in the way I told you in the last lecture right. Now, similarly you can have the divisive coefficient measures also just like the agglomerative coefficient measures. You can also have a divisive measure coefficient and they are trying to measure the amount of clustering structure found in the data. For example, if you try to see the value of the divisive coefficient is closer to 1, this means it is suggesting a strong clustering structure right. So, now you can recall that in the earlier lecture I have given you this slide that how are you going to work with the dendrograms right.

So, here each leaf in the dendrogram corresponding to one observation that is displayed here. You can see here this is one observation like each observation right. And then as we try to move up the trees observation similar to each other are combined into branches. So, you can see here that as we are moving here this thing they are joined by here this branches and so on. And if you are moving here up that at every stage for example, this is now club 1, this is club 1 and now both of them are club here at this point and so on right.

So, the height of the fusion provided on the vertical axis indicates the dissimilarity between two observations. The higher the height of the fusion the less similar the observations are. This is how you have to now interpret. Now, the next thing comes here that how are you going to measure that how many clusters are needed. If you try to recall earlier, I had made here these types of horizontal lines to indicate that this will indicate you how many clusters can be created when you try to cut the dendrogram.

So, if you have this idea that how many clusters would you like to have, then you can identify the subgroups using the command here cut tree. So, this cut tree c u t r e e you have to just be clear it is not cut tree this is cutree() which is a sort of say cut tree like this. So, this will cut the dendrogram which is resulting from this say command hclust into several groups either by specifying the desired number of groups or the cut heights right. So, for example, if I say suppose I want to have only here four groups four clusters. So, I try to use here the command here cutree(), c u t r e e and then I have to give here this output which is what is this output if you try to see here this was the output which we had obtained here right of which you had computed the agglomerative coefficient.

This is the same thing which is obtained by here this command here agens. Now if you try to use it and if you specify here t = 4, then you can see here these are the units and these are the here cluster number. It is trying to see here that cluster number 4, cluster number 3, cluster number here 2 and cluster number here 1. So, it is indicating now that unit number 2, 3, 4 they are in cluster number 1, 5, 6, 7, 8 they are in cluster number 2, 9, 10, 11, 12 they are in cluster number 3 and remaining are in cluster number 4, but it is difficult to count. So, what I can do that I can create here at tabular data of these values.

So, I try to use here the command table over this subgroup 4 in which I have stored the data and you can see here that it is trying to say that in cluster number 1 there are 4 values, cluster number 2 there are 4 values, cluster number 3 there are 4 values and cluster number 4 there are 8 values. So, this is the cluster number and this is the number of values. So, that will give you an idea that how the things are working. Now suppose if you want to have 6 clusters then you have to use the command here cutree() and the same data set which I have used here, but now k = 6. So, you can see here that it will give me the outcome and then if I try to make a table then or the frequency table then I can use the command table over this outcome subgroup 6 and then you will see here this outcome will come.

So, here we have here 1 2 3 4 5 6 these are the cluster number and these are the number

of units. You can see here first cluster has or say cluster number 1 has 2 units, cluster number 2 has 2 units, cluster number 3 has 4 units and so on. And now you can see here this is the screenshot of the same outcome which I have shown you. So, now let me try to show you this outcome in the R software.

So, that you can be more confident. So, first let me try to upload these software or this library. So, I already have installed these libraries. So, if you can see here that a library cluster fact extra and then extends this I have uploaded here. Now, I try to create my here this data set. So, if I try to take here this data set and I try to this thing here.

So, you can see here this is we can see here now this datacl is here like this right. And now I try to use first create the distance matrix. So, distance matrix if you try to see here it is here like this d if you try to see let me try to say you can see here d will be here like this right. But if I try to make it here it will be here like this because it is a long value. So, you can see here it will go up to 20 and then I have only had up to 10.

So, this 10 and then the remaining part is here right. So, you can see here, but these are the different values between for example, this value where I am highlighting this is the distance between this here third and here fifth value you can see here this is 1.73 and so on. So, similarly other distances are computed. So, you can see there it is not difficult to do them.

Now, I try to apply here the hierarchical clustering. So, you can see here this outcome is here like this and then I try to plot here this one. So, you can see here this will hit like this you can see here this is the dendrogram which I had shown you right. So, you can see here that it is now what a very difficult thing and you can do it very easily. So, now let me try to use here these agglomerative coefficients here if I try to show you here.

So, you will see here this is here the when I am trying to use the command agnes over this data set you can see here this outcome looks here like this right. You can see here this is agglomerative coefficient is equal to 0.89 and so on. Then order of the object and then there is here is summary of the observations and the different types of other information. And if you want to have here similarly here if you want to know only the agglomerative coefficient then you have to use the command and then you have to use here the command dollar ac and you will get here the same value which you have obtained here like this one right.

And then if you want to create here this tree here. So, there another command which has more option it is given by here like this you can see here this is created here or because it was not there earlier. So, I will try to create it once again. So, that you can see that the it is created here now it has many many options which I would like you to look into the help menu and try to do it here right. Now, if you want to have here if you want to create here the groups.

So, you can see here you can see here like this. So, the value of here subgrp4 it is here like this and if you want to make it here say here 6 groups here. So, you can see it let me call it here as a sub groups 6. So, that you can store it this one sub grp 6 it is here like this and if you if you want to suppose you make it here actually say here 12. So, you can see here 12 will be here like this. So, just you have to make k = 12 and then let me call it subgrp say 12 this outcome here is like this.

So, you can see here that the number of elements in the clusters they are decreasing and that is expected also right and yeah. So, this is all about this clustering in this lecture and. So, now, we come to an end to this lecture. Well, there are many more things which we have to do.

So, let us see. But my objective in this lecture was basically to give you the idea about these commands and these packages. Now, because I have taken a very small data set. So, here using clustering does not make any sense, but as I said in the beginning my objective is something else. I wanted you to understand that how the outcome is coming and how the things are happening. When you have only here 10 values in one variable possibly you can see how the things are happening, but when you have say this million values and also values in 1000s you can see this structure, but you will not be able to look inside the data to understand it.

In that case you will have to depend on the outcome. Looking at the outcome you have to understand what is happening inside the data. So, once I try to take here the very small data set which you can see from your eyes then looking at the values then you can have an understanding that okay this value is like this then this is going to happen. For example, when we try to look at the thermometer reading by looking at the level of the mercury or the reading of the temperature we decide whether the person has got a fever or not. So, this is the same job which we are trying to do when we are trying to understand the outcome of the R software. So, my request to you all is that please try to use this command try to look into the help menu because honestly speaking I cannot

explain you here all possibilities which can be done with these commands and unless and until you try to experiment with different choices different parameters it may be difficult for you to understand the whole process.

So, my request is that please explore this please practice this and I will see you in the next lecture till then goodbye.