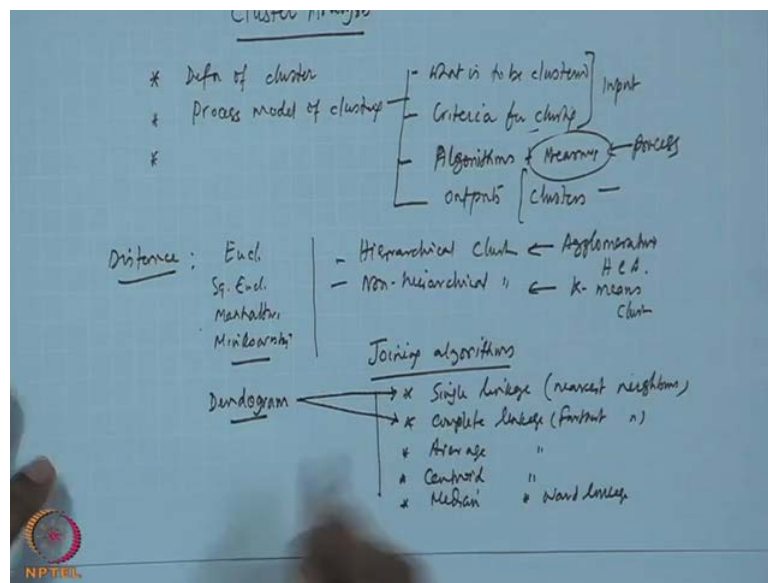


Applied Multivariate Statistical Modelling
Prof. J. Maiti
Department of Industrial Engineering and Management
Indian Institute of Technology, Kharagpur

Lecture - 37
Cluster Analysis (Contd.)

Good morning, today we will continue cluster analysis.

(Refer Slide Time: 00:24)



Last class what we have discussed? We have discussed the definition of clusters, and then we have gone for some process model of clustering. Under process model, what are the things we have known that what is to be cluster and criteria or criteria for clustering. Then, what you have done, and then we say that the partitioning algorithms and also we have seen outputs are nothing but the clusters. This is input, this one algorithm plus the measure some distance. These are basically process, this is output, and then under criteria distance measures distance measures particularly for continuous variables submitted data where I have given some of the measure like Euclidean distance square, Euclidean distance square.

Then, your Manhattan distance, then your Minkowski distance and then we have discussed the two types of algorithms, one is hierarchical clustering and second one is non hierarchical clustering. Under hierarchical clustering, we have discussed about agglomerative hierarchical clustering algorithm. And under this that we said that we will

be discussing k means, clustering and again under hierarchical clustering particularly for regular monitor clustering.

Then, we have we have discussed different joining algorithms, for example we have discussed single linkage which is well known as nearest neighbour, complete linkage which is also known as farthest neighbour. Then, we have gone for average linkage, then centroid linkage, and then there is median linkage, then ward linkage. We have also discussed something called dendrogram particularly ageing single linkage and ageing complete linkage we have described. Now, we will start with this single and complete and then we will go to have centroid and all those linkages.

(Refer Slide Time: 04:22)

Hierarchical joining algorithms

- *Single (nearest-neighbour)*: distance between two clusters = distance between two closest members of the two clusters.
- *Complete (furthest neighbour)*: distance between two clusters = distance between two most distant cluster members.
- *Centroid*: distance between two clusters = distance between multivariate means of each cluster.

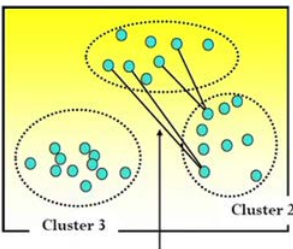
NPTEL

Now, let us see this slide what I have shown you in the last lecture, that single linkage where the nearest neighbours are joined this distance measures are considered for joining complete linkage. For this neighbour, distance is considered for the two clusters joining clusters, centroid we have not discussed, we will be discussing today.


(Refer Slide Time: 04:50)

Hierarchical joining algorithms (cont'd)

- *Average*: distance between two clusters = average distance between all members of the two clusters.
- *Median*: distance between two clusters = median distance between all members of the two clusters.
- *Ward*: distance between two clusters = average distance between all members of the two clusters with adjustment for covariances.



Mean/median/adjusted mean of all pairwise distances




Then, average, median and ward are also not completely discussed.

(Refer Slide Time: 04:55)

Agglomerative Hierarchical Clustering Algorithm

- Step 1: Identify the variables (p) and objects (n)
- Step 2: Collect data ($X_{n \times p}$)
- Step 3: Select similarity or dissimilarity measures
- Step 4: Obtain distance matrix ($D_{n \times n}$)
- Step 5: Start with n clusters where each cluster contains a single entity
- Step 6: Find out the nearest pairs of clusters from $D_{n \times n}$. Let the distance between most similar clusters R and S be d_{rs}




Then, the total algorithm I have already discussed with you.

(Refer Slide Time: 04:57)

Agglomerative Hierarchical Clustering Algorithm

- Step 7: Merge clusters R and S and label the newly formed cluster as (RS) . Update the entries of the distance matrix D by
 - Deleting the rows and columns corresponding to clusters R and S and
 - Adding a row and column giving the distances between cluster (RS) and the remaining clusters
- Step 8: Repeat the steps 6 and 7 for a total of $(n-1)$ times. When all of the objects will be in a single cluster the algorithm terminates
- Step 9: Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place



(Refer Slide Time: 04:59)

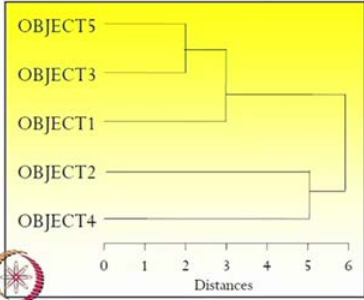
Single linkage (nearest neighbour)


Object	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Distance matrix

Object	5,3	1	2	4
(5,3)	0			
1	3	0		
2	7	9	0	
4	8	6	5	0

Distance	Cluster
0	1,2,3,4,5
2	(5, 3), 1, 2, 4
3	(1, 3, 5), 2, 4
5	(1, 3, 5), (2, 4)
6	(1, 3, 5, 2, 4)

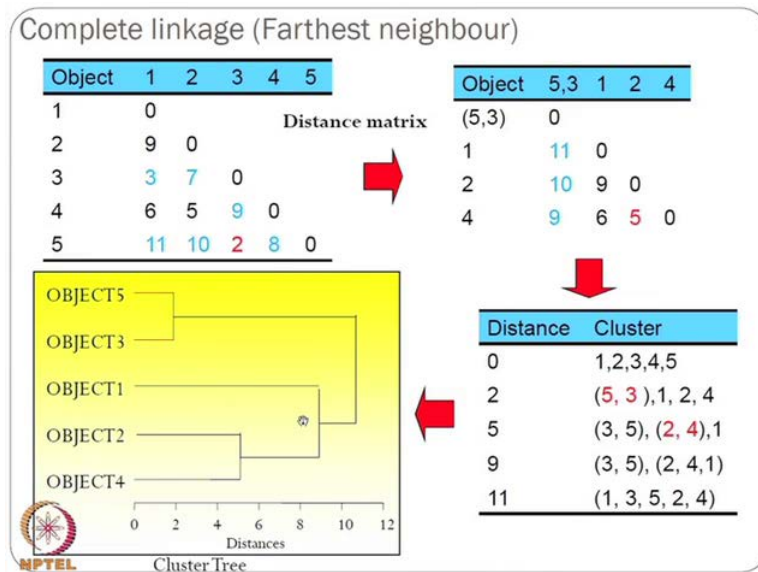




Cluster Tree

I have also given this example in the last class.

(Refer Slide Time: 05:08)



You have seen that how dendrogram was formed.

(Refer Slide Time: 05:10)

Joining algorithms

Single linkage

$$d_{RS} = \min \{d_{rs} / r \in R, s \in S\}$$

Complete linkage

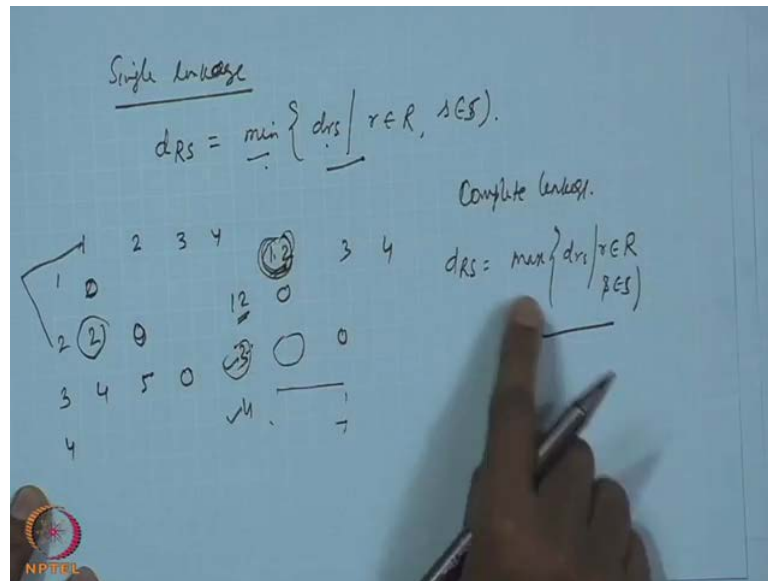
$$d_{RS} = \max \{d_{rs} / r \in R, s \in S\}$$

Average linkage

$$d_{RS} = \frac{\sum_r \sum_s d_{rs}}{n_R n_S}$$

So, today let us see the mathematical formulation of all those things.

(Refer Slide Time: 05:20)

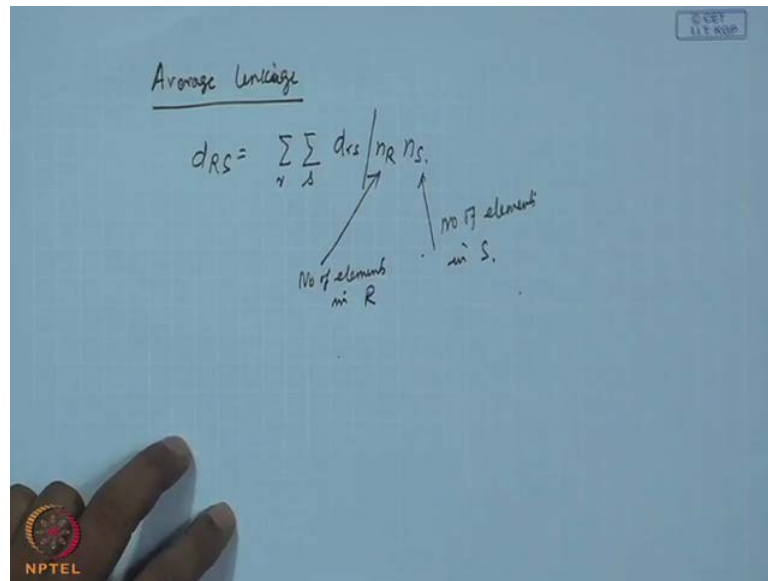


When we talk about the single linkage, basically we talk about distance between, suppose we are talking about two groups to clusters R and S, then the distance will be calculated like this minimum that d_{rs} given r belong to R and s belong to S. So, we have seen earlier also suppose there is 1, 2 and 3 on this side, 1, 2 and 3 and there is certain distance. Let it be this one, 1, 0, 2, and this one, let it be 4, then one let it be 2, 2, and 0. So, then 3 this will be 0, so we first started with grouping, this is the minimum we have grouped, so one 2 is grouped then 1, 2, 3 this will be this will be 0.

Then, one more thing is coming, but if you add one more object here, then ultimately you will be getting some more values and you have to find out the distance between these two, that this group vis a vis this, this group vis a vis this. Then, from this group vis a vis this, from this point of group the distance minimum distance between 1 and 3 and 2 and 3 will be consecutive.

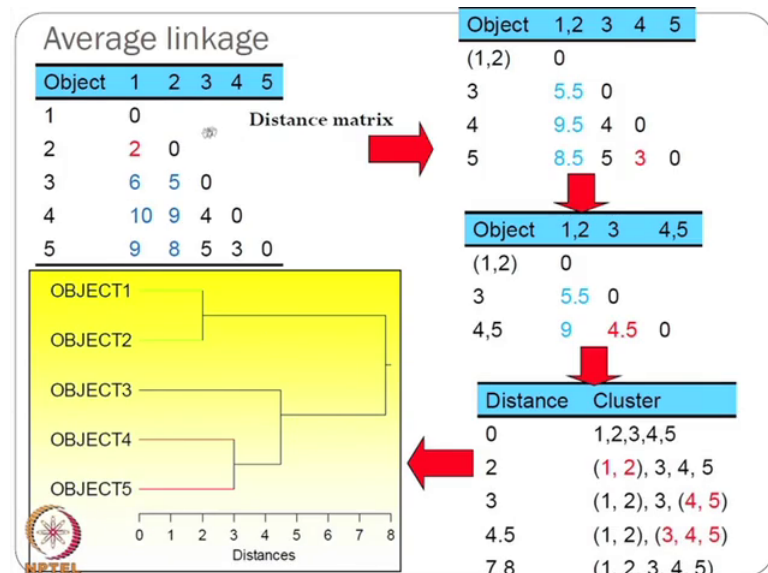
That we have discussed earlier and your complete linkage case what is the formula, complete linkage case that d_{RS} , it will be maximum of this model d_{rs} giving r belong to R and s belong to S. So, here when we join, we compare or we measure the distance between two groups or two clusters like 1 and 2 and 3. Then, the maximum distance between 1 and 3 and 2 and 3 will be considered.

(Refer Slide Time: 07:43)



In case average linkage, what it is done that is basically taken the average distance like this. So, we need talk about average linkage, then we talk about d_{RS} , that is the distance between two clusters, this will be sum total of s all s all r and d_{rs} by $n_R n_S$, where this is the number of elements in cluster R and this is number of elements in S.

(Refer Slide Time: 08:25)



Now, how it will work if I go back receive you see we are saying that average linkage here, so this is an example where five objects are to be grouped and the distance measures 2, 6, 5, 10, 9, 4, 9, 8, 5, 3. These are computed using certain matrix, let it be

equivalent distance, then your first step is what the minimum distance is, here that is 2. So, that is between 1 and 2, you have grouped 1 and 2 and then other matrix you have computed distance matrix like this.

Now, how do we get this 5.5, 9.5, 8.5 values, this other values 4, 5, 3 are basically between 3 and 4, between 3 and 5 and between 4 and 5 and that will be straight way getting from this previous matrix. Now, in this matrix how 5.5 is what we are using is average linkage, we want the average distance between the member, members of this of the two clusters that you are going to join.

So, here I want to see that the cluster 1 and 2 and cluster 3 what is the distance, now the average distance between 1 and 3 and 2 and 3, you find out 1 and 3 is 6 and 2 and 3 is 5. Now, you take the average, what will be the average? Average will be 6 plus 5 divided by number of elements in this cluster into number of elements, in this cluster number of elements in this cluster is 2 and number of elements in this cluster is 1 so 2 is to 1. So, 6 plus 5, that is 11 by 2, that is 5.5. So, in similar manner you will be calculating and finally this dendrogram can be prepared.

(Refer Slide Time: 10:25)

Centroid linkage

$$\bar{x}_r = \begin{bmatrix} \bar{x}_{r1} \\ \bar{x}_{r2} \\ \dots \\ \bar{x}_{rp} \end{bmatrix}, \quad \bar{x}_s = \begin{bmatrix} \bar{x}_{s1} \\ \bar{x}_{s2} \\ \dots \\ \bar{x}_{sp} \end{bmatrix}$$

$$\bar{x}_r = \frac{\sum_{r=1}^{n_r} x_r}{n_r}, \quad \bar{x}_s = \frac{\sum_{s=1}^{n_s} x_s}{n_s}$$

$$d_{RS} = \|\bar{x}_r - \bar{x}_s\|^2$$

The centroid of new cluster T

$$\bar{x}_t = \frac{n_r \bar{x}_r + n_s \bar{x}_s}{n_r + n_s}$$

$$x_r = \begin{bmatrix} x_{r1} \\ x_{r2} \\ \dots \\ x_{rp} \end{bmatrix}, \quad x_s = \begin{bmatrix} x_{s1} \\ x_{s2} \\ \dots \\ x_{sp} \end{bmatrix}$$

NPTEL

Now, come to centroid linkage, centroid linkage is little bit different where here we will first find out the centre of the cluster and then we see the distance between the centres. For example, let consider R is one cluster and S is another cluster, hypothetically assume that these many members are here, these many members are here in this total structure.

Definitely, each members are measured in terms of corrupt p variables x_r that $x_{r1}, 2$ and p and x is also x is 1, 2, p and p variables are there.

Now, you are finding out the mean of the variable vector here and as well as you are finding out the mean of the variable vector here and then what is the mean value how you are finding out. Here, r equal to 1 to n R, R is the number of elements in this cluster and S is the number of elements in this cluster. Then, you are finding out for each of the variable the mean value that is some total of individual variable values up to n R observations and divided by n R.

Similarly, for these you are getting this one, so the \bar{x}_r and \bar{x}_s , these are the vector quantity of p cross 1, then you find out the distance between this and this \bar{x}_r and \bar{x}_s and this will definitely this one to, so $\bar{x}_r - \bar{x}_s$ that is square. So, it will be $\bar{x}_1 - \bar{x}_1$ minus \bar{x}_1 square plus this minus this square plus this minus this square. In this sense, it will be calculated, so this is the distance when you are now what will be the, suppose if we combine these two r and s. These two if we combine, this cluster versus this cluster, this combine means one cluster will be created and this new cluster is T then what will be the centroid.

For this new cluster, the centroid for this new cluster will be that centroid that waited centroid of the two clusters joined together, so $\bar{x}_T = \frac{n_R \bar{x}_r + n_S \bar{x}_s}{n_R + n_S}$. So, then these two clusters once they joined, the number of clusters will reduce by 1, again you repeat the process for remaining clusters so long you are not reaching to only one cluster.

(Refer Slide Time: 13:38)

Average Linkage

$d_{RS} = \frac{\sum_r \sum_s d_{rs}}{n_R n_S}$

Centroid

1

2

n-1

n

No. of elements in R

No. of elements in S.

$d_{RS} = \|\bar{x}_R - \bar{x}_S\|^2$

So, you are starting with n cluster finding out these values, then you are going to n minus 1 cluster like this, then ultimately you will be going to two clusters, then finally we are going to one cluster. Everywhere when you are comparing, you are comparing the distance between clusters by this formulation not this one, this is your average we are talking about centroid, one centroid case we are saying this is s square.

(Refer Slide Time: 10:48)

Ward linkage

$SSE = \sum_{i=1}^n (x_i - \bar{x})^T (x_i - \bar{x}) = \sum_{i=1}^n \|x_i - \bar{x}\|^2 = T$

$SSE_r = \sum \|x_r - \bar{x}_r\|^2, r \in R$

$SSE_s = \sum \|x_s - \bar{x}_s\|^2, s \in S$

$SSE_t = \sum \|x_t - \bar{x}_t\|^2, t \in T$

$\bar{x}_t = \frac{n_R \bar{x}_r + n_S \bar{x}_s}{n_R + n_S}$

Join those two clusters for which $SSE_t - (SSE_r + SSE_s)$ is the least.

Let us see the word linkage, word linkage takes into consideration as I told you in my earlier class, the covariance structure. If you see this formulation, what we are saying

here, we are saying here SSE some square arrows, these arrows are nothing but we treating in this way that as if they are the distance between objects within clusters as well as between clusters. They are the major of heterogeneity, so when everything is belonging to one group, that means that is completely lipid, it is completely homogeneous, the distance will be 0, so in that sense the arrow quantity will be 0.

So, it is possible when we say that all are basically forming one cluster and then you will find out this is what is the total arrow in terms of the distance. So, then we are going on to what you want to find out that how many clusters that can be made. It is one cluster that means lot of heterogeneity will be there and when you when you go for more number of clusters, then the heterogeneity will be reduced. The homogeneity will increase, but when you go for n clusters, then it will make completely homogeneous because every cluster is having one object.

So, how it works the word linkage, works like this first you think that you have n clusters, so here are arrow quantity, that distance measure is absolutely 0. Now, you will then find out if you join any two objects into one cluster and then what will happen, the clusters you formed. It will have certain distance and as a result, the resultant arrow also will increase, so in this manner ultimately will move and the criteria in the movement upward are that. Suppose, at any distance let it be you are here that this is your R cluster, then this one there will be at this point, this is 1, 2, 3, 4.

That four clusters we have formed and in the first cluster what we are saying there are r clusters, that are 2, 3 and 4, 2, 4 and 3. There is another cluster, here this is one cluster, then this one is second cluster, this is third cluster, then this three again are forming one cluster, four clusters and then five clusters are there.

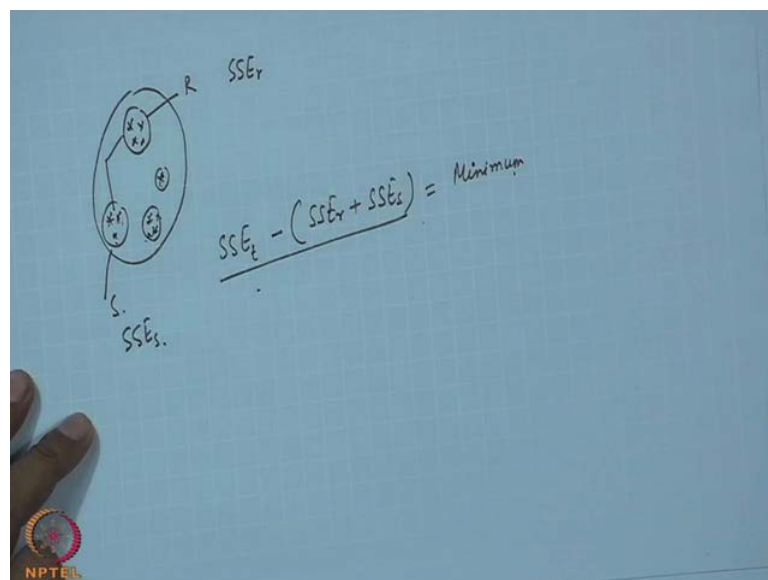
Now, what do you require to, you want to join that which two clusters will be grouped together so as we have seen that how SSE is measured by this distance measurement that sum total of distance. So, then you find you can find out for the earth where there are three points, three objects that is SSE that each point minus its centroid value. Similarly, for other if you say s is this one, and then you can find out what is the SSE value, it all depends if there is only one object in a cluster that is assuming it to be 0.

Now, if I join r and s ultimately what will happen, for example if I consider s is this one fifth one, and then the resultant cluster will have 2, 4, 3 and 5. These observations

resultant clusters if I say cluster is T, then what will happen? You will find out s T considering 2, 3, 4, 5, these 5, 4 objects distance sum total of this.

Now, what will be the centroid of this, that means what I am saying you are joining r and s two clusters, then you are getting this r SST, SST values and also the centroid of this cluster can be calculated like this. We have seen earlier also, then what we will do, basically we have to find out that how many clusters are there, then you will find out for every two clusters there are n k clusters. Then, k 2 combinations you will get for every combination you find out the SSE T, then you see the join those two clusters for which SSE T minus this is the least.

(Refer Slide Time: 20:11)




This means, what we are trying to say here I have, so for example, I have my this is my dataset, now I am creating one cluster here, another cluster here, like this. So, if I say this is my R cluster this is my S cluster, then I have a SSE r, here I have SSE s. So, if I join these two, I will get SSE t for these two clusters joint. These two, now you can join this versus this, this versus this, so there will be several 1, 2, 3, 4 combination will be there. When you find out for this value, you join those two clusters together for which this is the minimum.

(Refer Slide Time: 21:31)

$$d(R, P+Q) = \delta_1 d(R, P) + \delta_2 d(R, Q) + \delta_3 d(P, Q) + \delta_4 |d(R, P) - d(R, Q)|$$

Here $n_p = \sum_{i=1}^n I(x_i \in P)$ is the number of objects in group P

Name	δ_1	δ_2	δ_3	δ_4
Single linkage	1/2	1/2	0	-1/2
Complete linkage	1/2	1/2	0	1/2
Average linkage (un-weighted)	1/2	1/2	0	0
Average linkage (weighted)	$\frac{n_p}{n_p + n_Q}$	$\frac{n_Q}{n_p + n_Q}$	0	0
Centroid	$\frac{n_p}{n_p + n_Q}$	$\frac{n_Q}{n_p + n_Q}$	$-\frac{n_p n_Q}{(n_p + n_Q)^2}$	0
Median	1/2	1/2	-1/4	0
Ward	$\frac{n_R + n_P}{n_R + n_P + n_Q}$	$\frac{n_R + n_Q}{n_R + n_P + n_Q}$	$-\frac{n_R}{n_R + n_P + n_Q}$	0

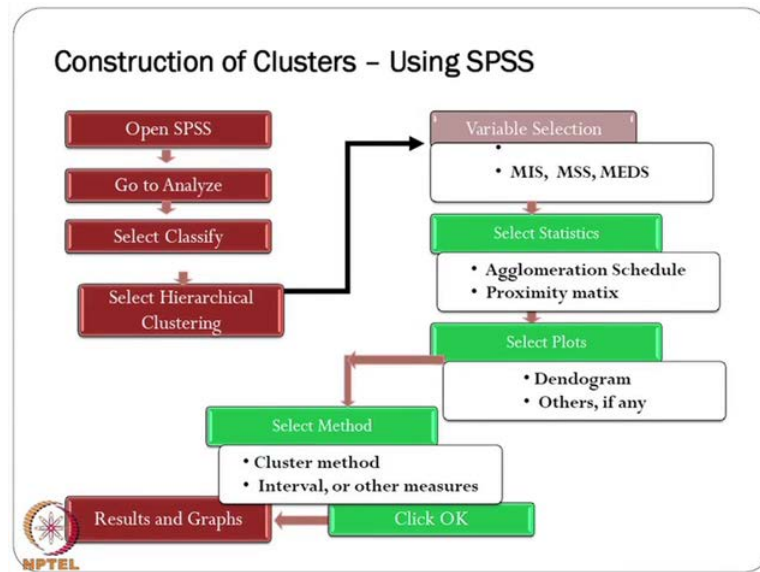
 Dr. J. Maiti, IEN, IIT Kharagpur

Then, we go to the totality, what I mean to say there are you see we have seen single linkage, complete linkage, so different way of joining for hierarchical clustering. These all combined together, there is a formula this $d(R, P+Q)$, that means R and P plus Q, they are joining here $d(R, P)$. Suppose, that means R P Q 3 clusters you are joining P and Q cluster, then do you get this distance for R and P plus Q joined together, that is another cluster, this is the case.

So, choosing appropriate delta 1, delta 2, delta 3 delta 4 values, so you will be able to apply for any of the joining algorithms what we have discussed so far. If you want to use single linkage delta 1, delta 2 will be half delta 3 will be 0 and delta 4 will be minus half. If you want to use complete linkage half only delta 3 is 0, if you go for average linkage this is the case delta 1 is half delta 2 is half. Then, 3 and delta 4 is 0 and similarly if you want to centroid linkage the same n_p by $n_p + n_q$, this one is this one is n subscript p, then n_q by $n_p + n_q$ like $n_p q$ this these are the values for determining this one.

So, similarly, word linkage you can find out, so if you anyone is interested to develop, then this is a very good formula, and then just by changing the delta and all this values. You will be able to get the get the clusters using all the Joining algorithms for hierarchical clustering.

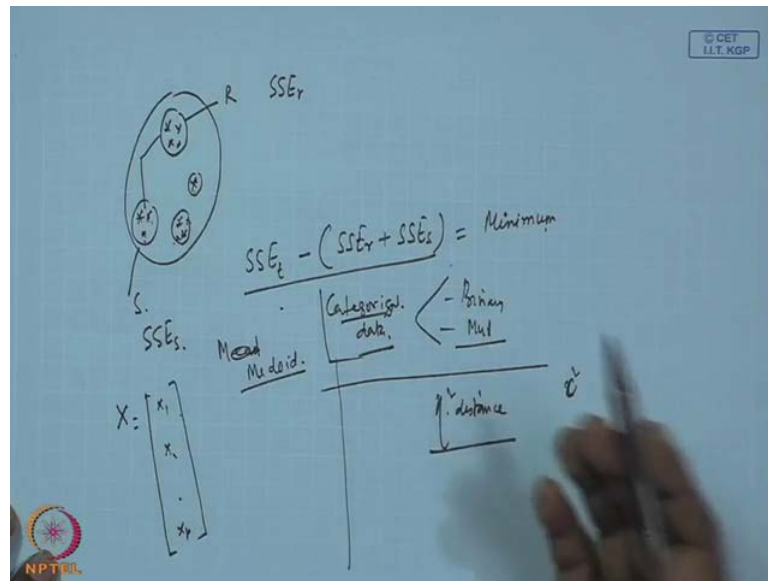
(Refer Slide Time: 23:37)



Now, I will show you that suppose you want to use these in s use this in the sense in 1 to run SPSS proton, SPSS for clustering. So, how do you go about in last but one, I think in few classes I have shown you how to run SPSS for different models and this, this, this, is the flowchart for running SPSS. For clustering, you open SPSS, go to analyse, then select classify, select hierarchical clustering, then you select variables that we have given, one example that MIS, MSS and MEDS, that mean incidence score, mean severity score and mean equivalent score.

Then, there are many statistics available agglomerations Sidole approximating matrix starting from n cluster to 1 cluster, how it is moving of that is the agglomeration schedule. What are the proximity matrix and then you can select plot dendrogram and then there are other plot any if you want to available. Then, you can do this, then you have to select the cluster method what type of measure it is, internal or other measures. If you are data in non metric data, suppose you are going for non metric data, then what will happen to your distance calculations will be different.

(Refer Slide Time: 25:20)

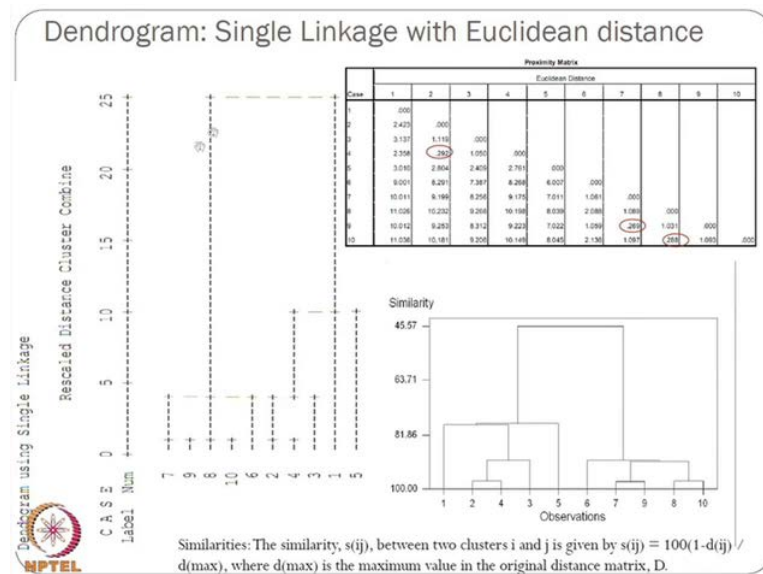


If there is binary data, the distance will be different; there will be maybe multinomial data, so I think there are certain techniques for handling categorical data. Then, there is one other technique is there multidimensional scaling they are categorical data handling mouse is done, but in clustering it is possible.

There is one I think that medoid, each one algorithm is there only the mean linkage median, but simply the medoid linkage is also there. Similarly, other linkages are there for example, if there is there are contingency tables then your chi squared distance you can use chi squared distance. So, cluster analysis requires distances, so if you can measure you first have to know that what is this variable vector and individual variables, how they are what is they are how they are measured, what is the there scale of measurement.

If it is continuous, fine if it is non matric, then non-matric data like your binary, your multi category. In that case, suppose you are using contingency table, so you can go for chi squared table. So, to find out some way or other the distance once you get the distance, then these algorithms are applicable, then you once you click ultimately I think in a second probably, depending on the memory available you will get the result.

(Refer Slide Time: 27:37)



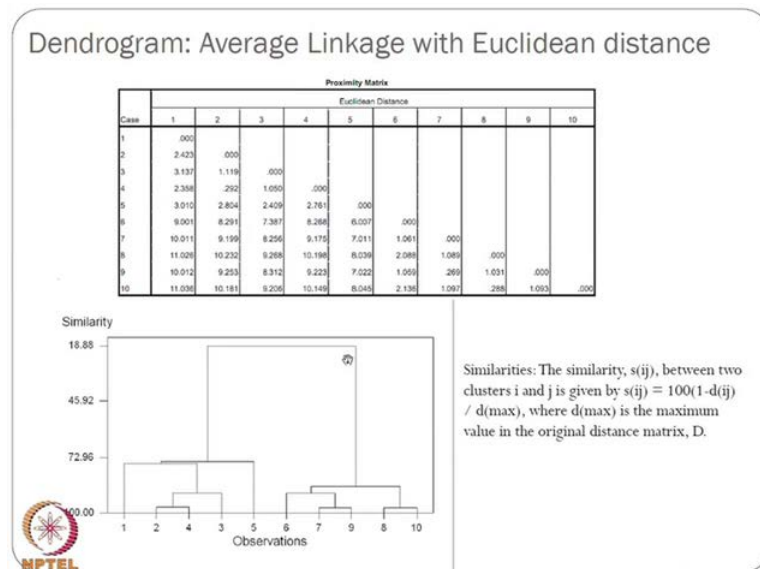
I will show you the results here, see what we have done, that screenshot clusters analysis we have gone to analyse them; we have gone to classify then hydrogen cluster. There are two steps hydrogen clusters, there are 3 and discriminant and there are different technique then the nearest labour will also. Then, once final click is made and we got this type of the one this is your Euclidian distance matrix, this is what is dendrogram, this is also dendrogram. The difference between this and this is here, distances are little distance are given the scale here and here we are talking about.

Similarly if I see that at any instant here, suppose I take this point 81.86 then if I draw a horizontal line with respect to this x-axis, then you will find out some intersection for example, with particle lens, this is one, second, third, fourth, fifth. So, this is the border, so four one, two, three, four, so that means at 81.96 percent, similarity level you can keep four clusters from dendrogram.

You are able to find out this four clusters but, if you are not happy with four clusters may be interested to have your two clusters, then what will happen two clusters were two clusters will form this is the case. Then, if you want x 1 cluster, so this is the kinetic 45.57, so how the similarity is calculated, here similarities is calculated in this formula.

that is the difference 2, 4, 3 joint, then 1, 5 I go like this, then 2, 4, 3, but 5 joint next then 1 you see this is the joining.

(Refer Slide Time: 30:44)




Now, if you go for average linkage, is there any difference here, yes again difference 2, 3, 4 are coming together, here means one after another, but then 1 coming and 5, that means that linkage this side and this may be joining may be different. Whatever may be the linkage method we have used, ultimately the clusters number of clusters and selective measures, two important criteria you cannot go for one cluster, if the similarity measure is drastically reduced. If we have there, you have to put a cut off that, how much similarity that means homogeneity you are going to accept.

(Refer Slide Time: 31:32)

Number of clusters

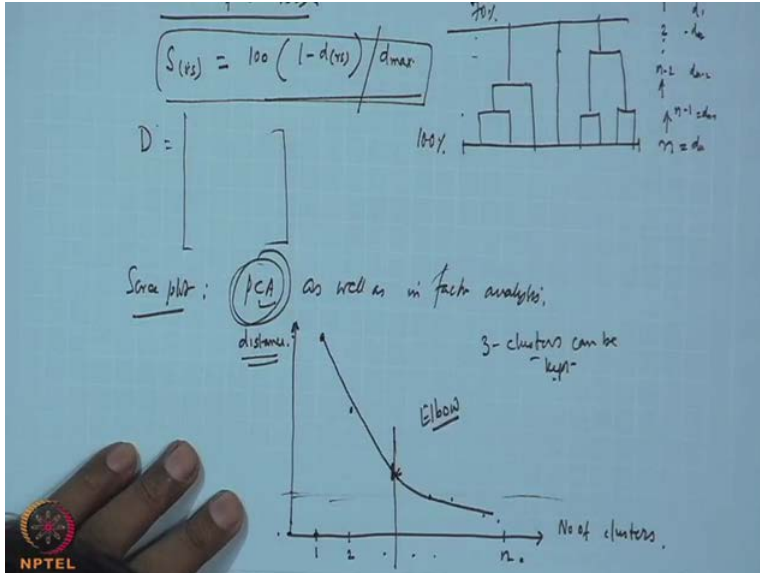
- % similarity
- Scree plot
- Incremental R-sq
- Pseudo F

Similarities: The similarity, $s(rs)$, between two clusters R and S is given by $s(rs) = 100(1-d(rs)) / d(\max)$, where $d(\max)$ is the maximum value in the original distance matrix, D.




Now, I will explain little bit of that similarity part as well as scree plot, incremental R square and pseudo R square ,what i can pseudo F also.

(Refer Slide Time: 31:49)



The handwritten notes include the following content:

- Equation: $S(rs) = 100 (1 - d(rs)) / d_{max}$
- Distance matrix: $D =$
- Scree plot: (PCA) as well as in factor analysis; distance.
- Elbow: A graph showing a sharp drop in distance followed by a gradual decline, with the point of sharp drop labeled "Elbow".
- Conclusion: 3-clusters can be kept.



These are the criteria for finding, keeping or finding out that what will be the number of clusters that you want to keep. As I told you that similarity measures is this, so definitely we want 100 percent similarity, but that is possible only when you treat each individual object as a cluster and 100 percent will be there and then you are basically slowly

joining. Then, you are making your cluster like this, maybe at this level you are having maybe 70 percent similarity, are you happy with 70 percent similarity, is it ok, it is fine.

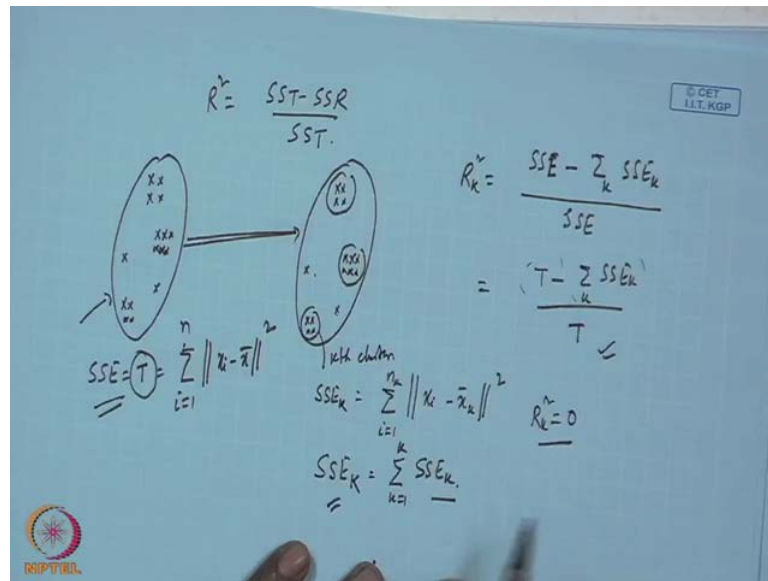
So, the scree, second one is used in scree plot, now scree plot I think you all understand at this point in time the scree plot we have used in PCA as well as in factor analysis. So, here also, the structure almost say that the concept is same almost same that we want to hear in PCA. We try to find out what will be the number of components in factor analysis what will be the number of factors in in clusters analysis what will be number of clusters.

Then, how do you do it, you do find a scree plot like this side you write the number of clusters and this side you write on the distance. Now, your cluster can be 1, cluster can be 2, like this you have n clusters, when you have n cluster, the distance becomes 0 because one cluster when what you are doing, suppose someone is n cluster, then you are creating n minus 1 cluster. Then, you are creating n minus 2 clusters like this, you are creating circus. This cluster, so if I say this is my d_n , this one is d_{n-1} and this one is d_{n-2} like this d_2 and d_1 , definitely d_1 will be the maximum.

Now, what will be this distance, it all depends on which linkage method you are using, if you use the single linkage that means the minimum distance amongst the member of the groups. If it is complete, that is the farthest linkage that will come here in this side. So, let it be there are ten, there is one cluster definitely this then what you do your cluster number two for example, let it be here cluster number 3, 3 cluster here. Maybe, after that it is like this, so I can say here maybe the point which is the yellow, so if I have appropriate distance measure, then only we will be able to do this.

Then, elbow this says that when as per the distance is concerned, now at this point what happened after that the distance is almost parallel means not improving much. So, I can say 3 clusters can be kept and you have seen in my the distance part, I will tell you that there is the issue. I think here you can understand now when you are going for these that, no cluster that 5 cluster 0 distance. So, what it is from the single linkage, from complete linkage, the distance is changed, so in this manner you have to remove the distance while you are joining and then you find out. Then, next one is incremental R square, what do you what is happening there.

(Refer Slide Time: 37:08)



Basically all of you know in R square regression this one is SST minus SSR by SST, that is sum square total minus sum square regression by sum square total. Here, what we are doing here, we are assuming like this, suppose I have group of objects like this, like this and when you are treating one cluster, then what is these means if I say the total. Then, this one is i equal to 1, 2 x i minus \bar{x} R square, now suppose you have you are going for generating k clusters like these 1, 2, 4, 5 clusters circuit clusters.

So, this is my suppose this is the k clusters, then I want to know what is the SSE for this k cluster, my SSE for this k cluster will be sum total of i equal to n_k the number of items. Then, you write down x_i minus \bar{x}_k or you can write x_i minus \bar{x}_k here, so as there are there k clusters. So, ultimately if i say that SSE, the total capital k , then this will be k equal to 1 K SSE small k .

So, as you have created k clusters 1, 2 k clusters, so ultimately what will happen this will reduce this one, now from this one change to this one and using this feature that as if the regression the what way. It will reduce the difference between the predicted versus original, in the same manner we can create r k square which is SST minus k SSE minus k .

I think we will write k SSE the way we have written here SSE t divided by definitely SSE, then what you can write on this one, this is basically T we are saying this as T minus this something k SSE k by T . So, what will happen ultimately, ultimately your R k

square will be 0, when this quantity equal to T and as you as you go on increasing the cluster, that this value will change, if you decrease this value will change.

(Refer Slide Time: 40:58)

Handwritten mathematical derivations on a blue background:

- Top left: $R^2 = \frac{SST - SSR}{SST}$
- Diagram: A large oval containing points x_1, x_2, \dots, x_n is shown on the left. An arrow points to a smaller oval on the right containing a subset of points, labeled "kth cluster".
- Bottom left: $SSE = T = \sum_{i=1}^n \|x_i - \bar{x}\|^2$
- Bottom middle: $SSE_k = \sum_{i=1}^n \|x_i - \bar{x}_k\|^2$ (with "kth cluster" written above the sum)
- Bottom right: $SSE_k = \sum_{k=1}^k SSE_k$
- Top right (circled): $R_k^2 = \frac{SSE - \sum_k SSE_k}{SSE} = \frac{(T - \sum_k SSE_k)}{T}$
- Bottom right: $R_k^2 = 0$
- Bottom center (boxed): "Incremental change" $\Delta R_{k,k-1}^2 = R_k^2 - R_{k-1}^2$

So, ultimately what you want at here you have to find out the incremental change, if I say if I go from k to k minus cluster, then this change which R k square minus r, this where R k square or k minus 1 square will be computed even in this formulation. If there is sharp change, sharp decrease let it be, and then what will happen when such change that means that maybe the point where we can stop that this many clusters.

(Refer Slide Time: 42:00)

Handwritten mathematical derivations on a blue background:

- Top: $F = \frac{MSR}{MSE} = \frac{SSR/dof}{SSE/dof}$
- Bottom: Pseudo F = $\frac{(T - \sum_k SSE_k) / (k-1)}{\sum_k SSE_k / (n-k)}$


Now, using this concept, similarly you can use F statistics I think in regression, we have used F is MSR by MSE and that one we have used that SSR by degree of freedom by SSE by degree of freedom. Here also, you can say that pseudo F which will be T minus sum total of k SSE k by k minus 1 divided by sum total of SSE k by n minus k.

This is your pseudo F similar to statistics used in regression equations, why pseudo because ultimately the distributions is not known, you are in a loop, there are several individuals of that what distribution, joint distortion. Then, we will follow that is not clearly is it multivariate or not if not then how can we say that this will be followed in distribution. So, why we are talking about the pseudo distribution and the in similar manner you have to decide the number of k.

(Refer Slide Time: 43:30)

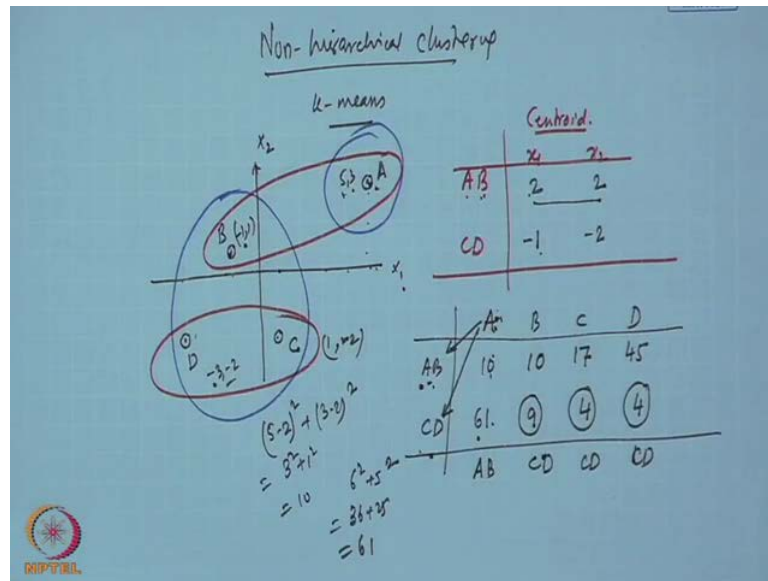
K-means clustering algorithm

- Step 1: Identify the variables (p) and objects (n)
- Step 2: Collect data (X_{np})
- Step 3: Select similarity or dissimilarity measures
- Step 4: Obtain distance matrix (D_{nn})
- Step 5: Partition the objects into K initial clusters
- Step 6: Proceed through the list of objects, assigning an object to the cluster whose centroid (mean) is the nearest
- Step 7: Recalculate the centroid for the cluster receiving the new object and for the clusters losing the object
- Step 8: Repeat step 6 and 7 until no more reassignments take place



Then, we will come to another important algorithm which is known as k means algorithm, which is coming under known hierarchical clustering k means we will be discussing.

(Refer Slide Time: 43:41)



Now, k means if you just read out then you see that the steps, there are eight steps totality identify the variables collect data select similarity or dissimilarity measures obtain distance matrix. This can be said that is the obvious thing that is equal, then partition the objects into k initial clusters what you are doing here, you are basically arbitrarily setting some clusters.

Then, proceed through the list of objects assigning an object to the cluster who centralises the nearest. Recalculate the centroid again and then repeat this process until no more assignment takes place, how do to it I am showing one example, you see this example here there are four objects a, b, c and d and you are measuring using two variables x_1 and x_2 . Now, the values the coordinate values I can say for a 5, 3 minus 1 over 1 minus 2, 3 minus 2.

So, if you see pictorially what is the point is I have x_1 x_2 my x_a is 5 and 3, so 1 2 3 4 5, so this is my a where is my b, b one minus 1, so it is minus 1 then 1 this is you b then c is 1 minus 2. So, I get 1 minus 2, this is my c then d minus 3 minus 2, 2, 3 and it is 3, so I want to create possible clusters what we will do you arbitrarily put the objects into clusters. For example, we can do so we may these one or two clusters, this is another cluster and this still what you are doing you are basically making a b 1 cluster c d another cluster, then you have to find out the centroid for x_1 and x_2 .

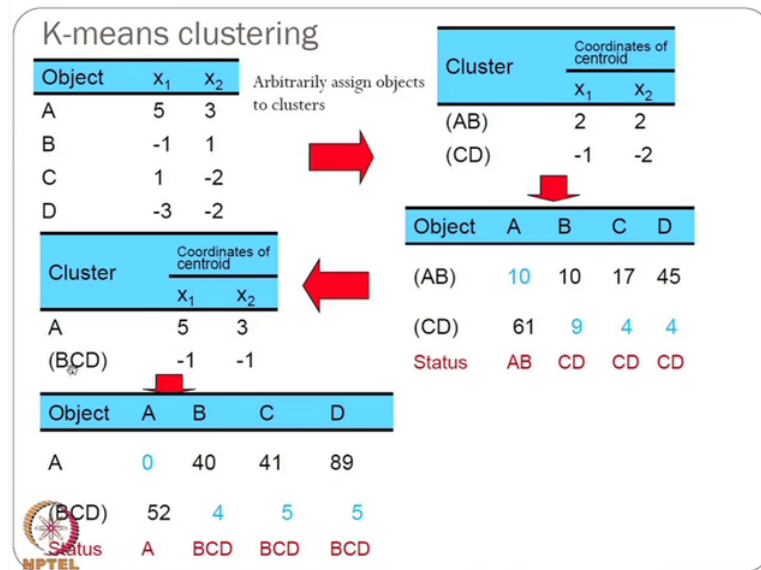
Now, what will be the centroid because a and b these two, so this one is 5, 3, so if I write further this is my 5 and 3, this one is my minus 1, 1 and c is 1 minus 2 and d is minus 3 minus 2. So, 5 minus 1 divided by 2, this will be 2, similarly 3 plus 1 by 2 this is 2, now what is the centroid here, 1 minus 3, that is minus 2 by 1, 2 that is minus 1, then minus 2 minus 2 minus 2 minus 4, y minus 2. This is arbitrarily, then what you have to do?

You have to calculate again that that for all the things that a b c and d where the lie are they truly falling under a b or c d b c d then a b is there and c d is there. Now, find out the distance between a and a b, a and c d, now a a is 5, 3 a b 2, 2, so 5 minus 2, 5 minus 2 square plus 3 minus 2 square, that means 3 square plus 1 square that means 10 square, this one is 10 distance.

Now, if I want this, then 5 minus 1 minus 1, so 5 minus 1 minus 1 is 6, so 6 square 3 minus minus 3 minus minus 2 that means plus 5 square. So, this is 36 plus 25, so that one is 64, so then you compare which one is higher c d is higher, so definitely I am considering comparing a's position where it should go to a b. It will be in this group or in this group as a b less than a b and a distance 10 which is less than the distance between a and c d is 61, so this one it should be a, a b.

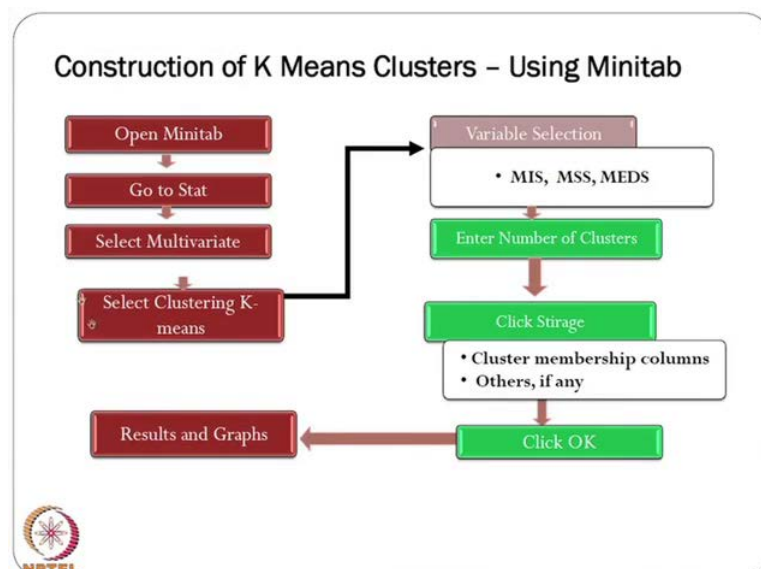
Similarly, if you compute for b versus this, this will be 10, then this will be 17, then this will be 45 this will be 9 this will be 4, this will be 4 and if you compare, now are that this should come under b should come under c d, c should come under c d. These are the lowest values and d should come under c d, so accordingly and it is also it is quite obvious because if you see here this is quite faraway compared to even if I compare b and d it is nearer than a and b. So, our new cluster is like this one and this one you repeat this then find out.

(Refer Slide Time: 50:28)



Let us see what happened here, so new cluster A B C D and you are finding out the centroid here and then again you are comparing that the distance between A and the two clusters formed. Distance between B and two clusters formed and in this case you see, so ultimately A belongs to cluster A, but B C D and if b belongs to cluster B C D and C belongs to cluster B C D and d belongs to cluster B C C. So, no more clustering is required so that for this data set two clusters are enough that what k means cluster.

(Refer Slide Time: 51:13)



Now, if you want to use mini tab for k means clustering, then you go to mini tab go to stat go to multivariate select clustering k means then you select variables enter number of clusters click. This story this will be stories and then click you will get the results and this is the snapshot.

(Refer Slide Time: 51:44)

K means clustering for department clustering problem

Final Partition


	Number of Obs	WCSS	Avg Dist	Max Dist
Cluster1	5	10.013	1.325	1.925
Cluster2	5	2.881	0.66	1.204

Cluster Centroids

Var	Cluster1	Cluster2	Grand Centroid
MIS	5.4	14.2	9.8
MSS	2.32	1.104	1.712
MEDS	1	2	1.5

Distance Between Clusters

	Cluster1	Cluster2
Cluster1	0	8.9397
Cluster2	8.9397	0




This is what cluster one and two for the for the say data this is not the example data what we have just solved here Data safety data MIA's and all those things. So, ultimately you are getting two clusters 5, 5 observation in each clusters and they are distributed with a centroid some square and average distance maximum distance this values all are given. Now, you have to appropriately interpret the values what you are getting for different parameters.

(Refer Slide Time: 51:13)

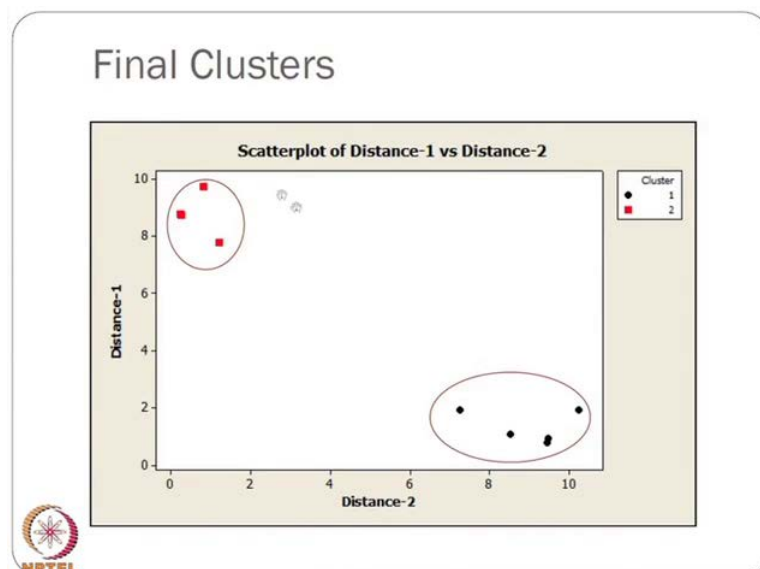
Cluster Membership

S No.	Department	Cluster	Distance-1	Distance-2
1	A	1	1.92	10.25
2	B	1	0.92	9.48
3	C	1	1.07	8.52
4	D	1	0.79	9.45
5	E	1	1.92	7.27
6	P	2	7.78	1.20
7	Q	2	8.72	0.25
8	R	2	9.74	0.81
9	S	2	8.76	0.23
10	T	2	9.71	0.82



Then, I am showing here that the cluster membership a b c d, one cluster, one P Q R S T cluster two, why we have generated data generated from two clusters and we are now finding out that is why A B C D and P Q R S T two different ways. We are now finding the name and ultimately the distance is coming and you are getting, you have got two clusters with the distance like this.

(Refer Slide Time: 52:46)




This is the final clusters this is here 5, this one left so that is why you are not able to see.

(Refer Slide Time: 52:56)

A Case Study

- The case study was conducted on a large integrated steel company. Data was collected for the past one year. In total 175 departments were analyzed based on mean injury potential score (MIPS) and mean equipment damage score (MEDS) to understand the hidden structure of the data.
- The sole purpose of this case study is to group departments with similar MIPS and MEDS so that safety initiatives and resources will be put accordingly.



Now, I will show you within three minutes of time and one case study a case study was conducted on a large integrated still company data was collected for the past one year. In total there are 175 departments and we have analysed based on mean injury potential score and mean equipment damage score to understand the hidden structure of the data. The sole purpose of this case study is group departments with similar features so that safety initiative and researchers will be. Now, we have used multivariate clustering k means using mini-tab and these are the snapshots.

(Refer Slide Time: 53:47)


Cluster Analysis Results

Final Partition

	No. of Obsv.	WCSS	ADC	MDC
Cluster1	26	231.26	2.60	6.20
Cluster2	28	205.26	2.37	6.60
Cluster3	47	113.14	1.39	2.89
Cluster4	15	407.04	4.37	12.28
Cluster5	59	160.71	1.53	2.85

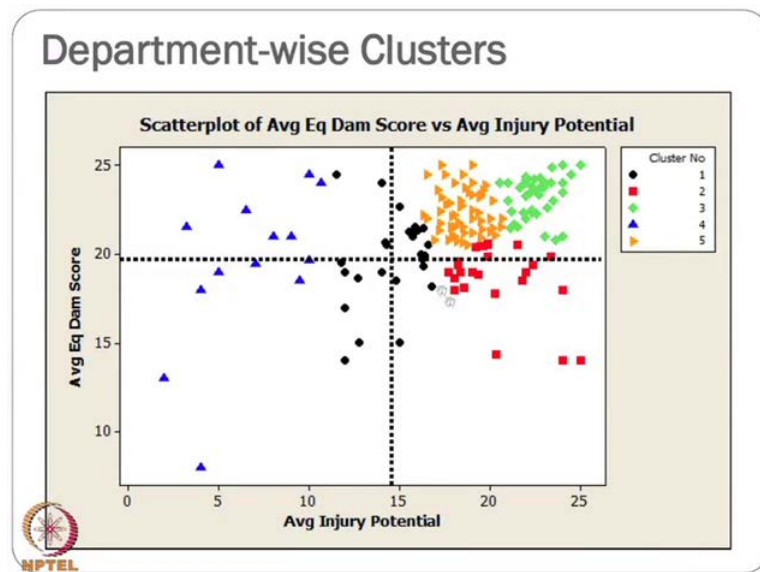
Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Grand
Avg Injury Potential	14.53	20.38	22.63	6.60	18.55	14.53
Avg Eq Dam Score	19.66	18.72	23.34	20.01	22.52	19.66



This is what is result we found that 175 this one can be the departments can be grouped under 59 clusters, not 59 cluster 5 clusters. Number of departments under each of the cluster is given cluster 126 cluster 228 like this. So, we from 175 to 5 reduction is enormous and because of this much reduction definitely there will be similarity problem, but in k mean clusters that similarity that mini tab is not given, but the method procedure we have discussed we can find out.

(Refer Slide Time: 54:39)



Now, see the plot if I see the from average this injury potential vehicle damage score and as per the membership the cluster number concerned you're getting this type of picture, but please find out to see the humour what is happening here. There are so much of overlap and it is quite obvious and because you a may not create that we are clear crystal clear that demarcations unless and the nature is like this. The department are such that there is a distinct difference, but in the safety point performance point of view for these yes these are the departments which are absolutely different from these are the department.


Similarly, these are the department who are which are basically performing; this is a 0 point that is the performing 0 is this one and this. So, this is the 15 that is average performance here average performing here, but here this content this is really a complicated one and very much, what I can say important also because injury potential and equipment damage score both are higher in this.

So, now management may see this result and then out find out that no this cannot can it will be a one clusters. So, then you will reduce the cluster, but when you review the cluster similarity will lost that means homogeneity will reduce so that also you have to consider properly. Then, you have to decide on that what you have to do how many clusters, you will keep some clustering and selection of clusters some multivariate statistics nowadays available, but it is again in the research domain and difficult.

(Refer Slide Time: 56:38)

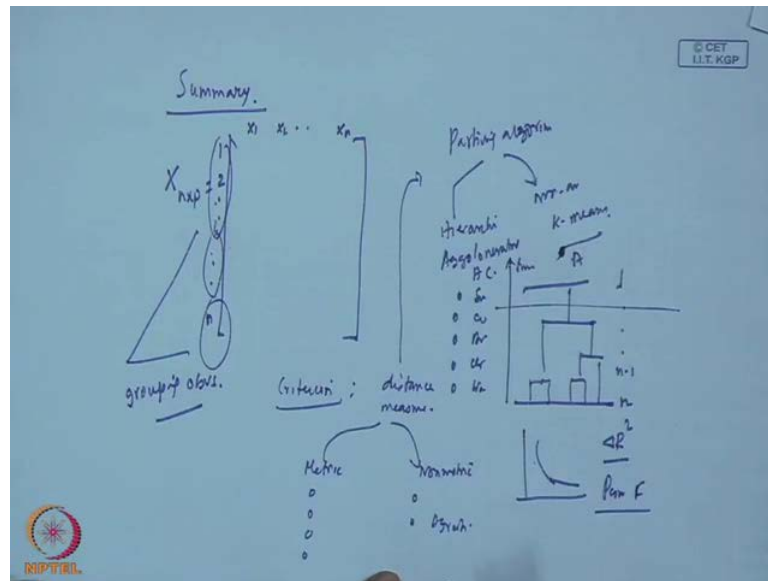
Conclusions

- Cluster analysis is very useful and widely used technique in grouping unstructured multivariate observations. ☺
- Several algorithms are available and so, the analyst must choose the best one suited to the data.
- The case study reveals that although the company has 175 departments altogether, from safety performance point of view these can be viewed as 5 groups.
- The safety initiatives and resources can be accordingly allocated.



Then finally, conclusion that clusters analysis is very useful and widely used technique in grouping unstructured multivariate observations several algorithms are available. So, the analyst must choose the best one for the data the case study reveals that although, the company has 175 departments together from safety point of view, this can be grouped into five groups the safety initiatives and resource can be taken accordingly. So, I will just summarise that what we have learnt in cluster analysis.

(Refer Slide Time: 57:16)



New dataset is like this, you have several characteristics features you want to group the in observations. This is our purpose, then what we've discussed that we require certain criterion to group this is distance measure there are many distance measure. Suppose, you have metric data you non metric data, metric data Euclidian square Euclidian Manhattan inquisitive distance and categorical data chi squared distance, then you are in binary case. Some several combinations of distance several distance measures are possible, follow once you know the distance you require knowing the partitioning algorithm.

This can be hierarchical can be non hierarchical, hierarchical is the agglomerative hierarchical clustering and here k means clustering here there are several linkages starting from single complete average centroid then ward linkages. Here, it is basically k means is realistic one and it one basically arbitrarily grouped and then what like this, now in hierarchical case things started from n clusters then move n minus to 1 cluster.

In this process, you create dendrogram in this process is a very useful pictorial representation which will show you that how many clusters and how the agglomeration process also takes into account in the in this. This side will be the similarity measure that similarity measure will tell you how to go about and what will be the number of clusters.

There are some other ways of doing things I shared that scree plot you can use you can go for that incremental R square, you can go for pseudo F. Ultimately, the cluster

analysis is really very helpful and useful to what all depends on the quality of data and the analyst knowledge for the purpose. It is used keep in mind this because you your brain should not verify the statistics only.

Thank you very much.