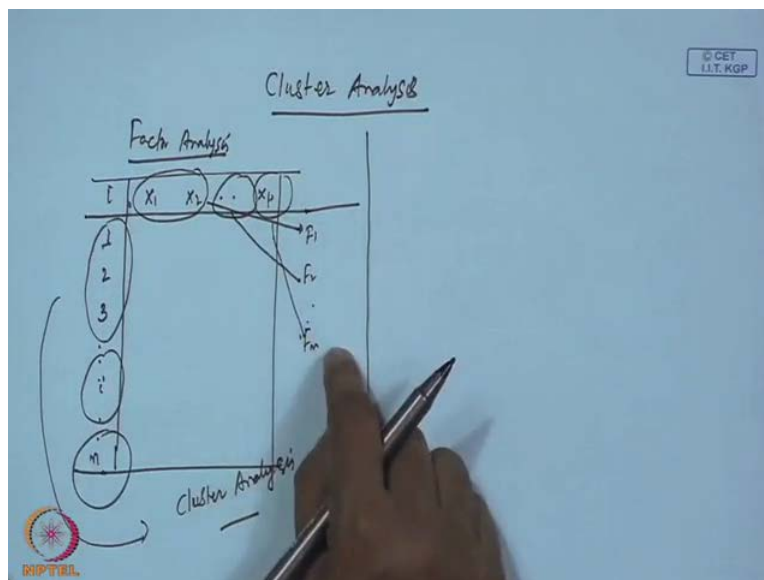


Applied Multivariate Statistical Modeling
Prof. J. Maiti
Department of Industrial Engineering and Management
Indian Institute of Technology, Kharagpur

Lecture - 36
Cluster Analysis

Good evening, our today's lecture is cluster analysis.

(Refer Slide Time: 00:25)

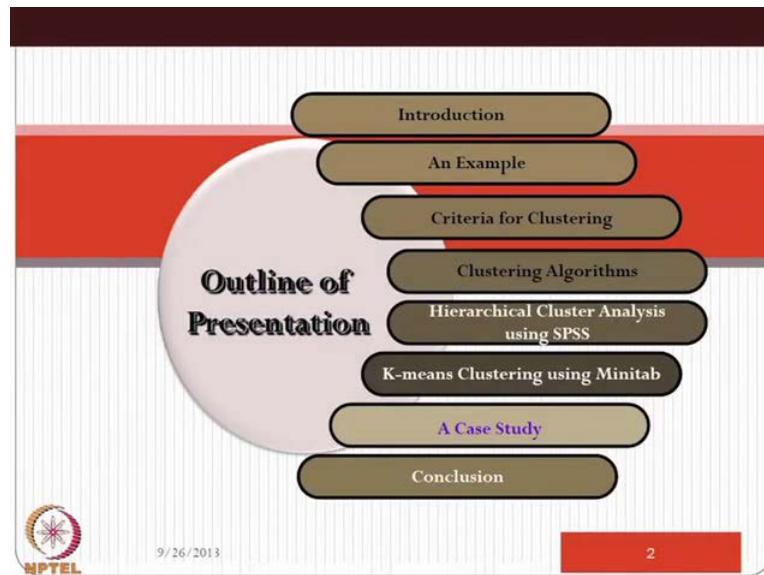


So, let us discuss from factor analysis some analog to cluster analysis and the difference so, if you recall factor analysis then there you found out that, there are n number of observations and p number of manifest variables and you created certain factors. For example, f_1, f_2 or like this f_n factors and you tried to link some of the our group, some of the variables to factor 1, some other variables to factor 2 now, rest variables to factor n like this. So, that mean in factor analysis this is necessarily you have grouped the manifest variables into different groups confectors and you have done that with the help of observations in multivariate observations on p variables.

Now, let us think little differently suppose, I have n number of observations and all this x_1 to x_p , this are characteristic features for each of the observations. Now, based on this characteristic features if you want to group the individuals into several groups, then this is known as grouping the items or grouping the individuals or grouping the observations or grouping anything. So, then

this one is known as cluster analysis. The mathematics behind cluster analysis is different than the mathematics used in factor analysis, the purpose is different, the objectives what we basically sharp they are also difference, but only one analogy is that, that it is also grouping like factor analysis we want to group several variables into different factors. Here, that grouping is also done but not with respect to the variables with respect to the observations.

(Refer Slide Time: 03:28)




With this back ground let me start cluster analysis and today's discussion is while first start with certain example, then we go for what are the criteria to be considered for clustering. Then I will show you some algorithms of clustering that is known as clustering algorithms and two clustering algorithm will be discussing one is hierarchical lumaratic clustering and other one is k-means clustering. And these two clustering how you will be able to run using spss as well as mini tab particularly hierarchical clustering using spss and k-means clustering with mini tab. Using mini tab I will be discussing then I will show you one case study and conclusion that is the totality of the presentation and in this 1 hour lecture. We will try to complete as much as possible if something remains it will be given in the next class.

(Refer Slide Time: 04:39)

Introduction

- ‘Cluster’ is “a number of things of the same kind growing or joined together”(Chambers, 2005)
- A group of homogeneous things.
- The principle (Kaufman and Rousseeuw, 2005):
 - ‘objects in the same group are similar to each other,
 - objects in different groups are as dissimilar as possible’.



Let us define what is cluster so, if you see the chambers dictionary 2005, the definition is cluster is a number of things of the same kind growing or joined together.

(Refer Slide Time: 05:04)

Cluster Analysis

Factor Analysis

i	x_{i1}	x_{i2}	\dots	x_{ip}
1				f_1
2				f_2
3				f_3
\vdots				\vdots
i				f_i
\vdots				\vdots

Species


$x_1, x_2, \dots, x_p \leftarrow$ features

- 1
- 2 - Based on similarity in a group
- 3 - Based on dissimilarity between groups
- \vdots
- \vdots
- n

C1
X X
X X X
X

C2
O O
O O
O

C3
G
G G
G G

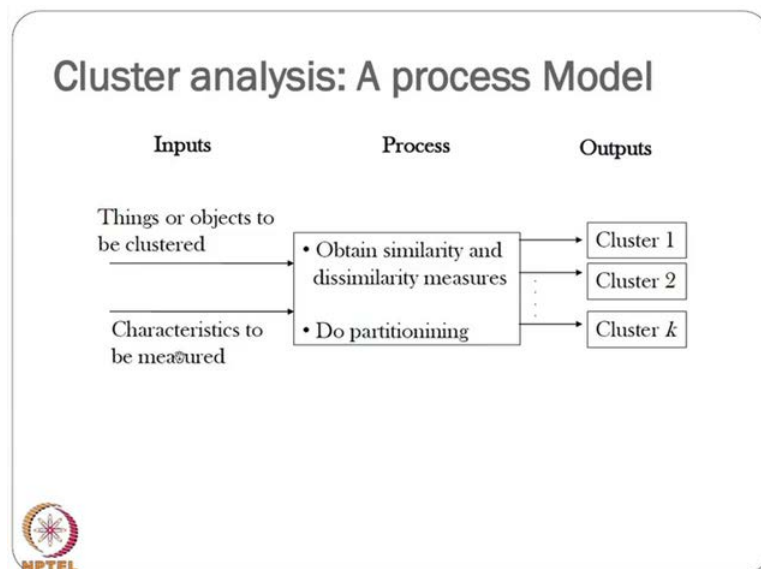


For example, if you all of us know the difference species from the biology different species so, they are grouped together based on certain features and they are grouped basically, because of their natural similarities. For example, if we can say that that from food habit from point of view

the animals can be that herbivores omnivores so, 3 different kinds that can be found out omnivores. So, essentially then a group of homogeneous things are known as cluster. So, the principle in grouping will be like this which is given in Kaufman and Rousseau 2005 that objects in a same group are similar, to each other and objects in different groups are as this similar as possible that is the principle.

So, that mean what we are saying clustering based on certain characteristic features like x_1, x_2 to x_p these are the features, these features will be used to group several items or objects and the grouping is based on similarity in a group and based on this similarity between groups, that is the principle. So, you want to make cluster in such a manner that whatever objects writing will be grouping in a particular cluster say cluster 1, say there is another cluster 2, cluster 1, cluster 2 there may be another cluster, cluster 3. So, there are several items grouped, here also several items group, here also several items group, what I mean to say that the items within a group they as similar as possible. Then so, item between goes these plus these are this they will be as this similar as possible, that is what I can say principle or i can say that is the philosophy in doing cluster analysis.

(Refer Slide Time: 07:54)



Cluster analysis can be thought of a process model and in this case you see the inputs process and outputs model for cluster analysis. Under input what we require, we require two things one is


what is to be clustered or what are the things to be clustered and what are the criteria or may be one criteria or several criteria or several features that will be used to cluster. Then process is you have to find out the wave of clustering that means, you require a major which will tell you that some of the objects or items are similar some of the objects and items are this dissimilar. Hence, the similar items will be joined in group one or in some in a group and dissimilar will be in different groups and that process is known as partitioning. Also, then output of this cluster model will be several clusters or groups, let us start with an example.

(Refer Slide Time: 09:07)

An Example

- The safety manager of an automobile company is interested to group the different departments based on their safety performance scores. Let, there are 10 different departments and variables that are of importance in measuring the safety performance are:
 - I. Incident score
 - II. Severity score
 - III. Equipment damage score

The manager analyzed last 2 years' performance of the 10 departments and arrived at the performance figures given in Table 1.



The example, is the safety manager of an automobile company is interested to group the different departments based on their safety performance scores. Let, there are 10 different departments and variables that are of importance in measuring the safety performance are incident score, severity score, and equipment score. The manager analyzed last 2 years performance of the 10 departments and arrived at the performance figures given in table one. So, what is essentially happening here suppose, you are a safety manager of a company and your duty is to keep people safe while working of the work, on the work.


And you have data all the performance on the safety performance of different departments which are under your control. Over the years your measuring is and then finally what is your interest if there are many departments and there are certain key features which are similar to each of the

department. So, based on given features you want to find out that whether the safety performance of all the departments are same, similar or there are some departments which can be grouped. There are some other departments those can also be grouped and accordingly what will be your benefit, your benefit will be you will you will take safety decisions group wise or cluster wise.

(Refer Slide Time: 11:04)

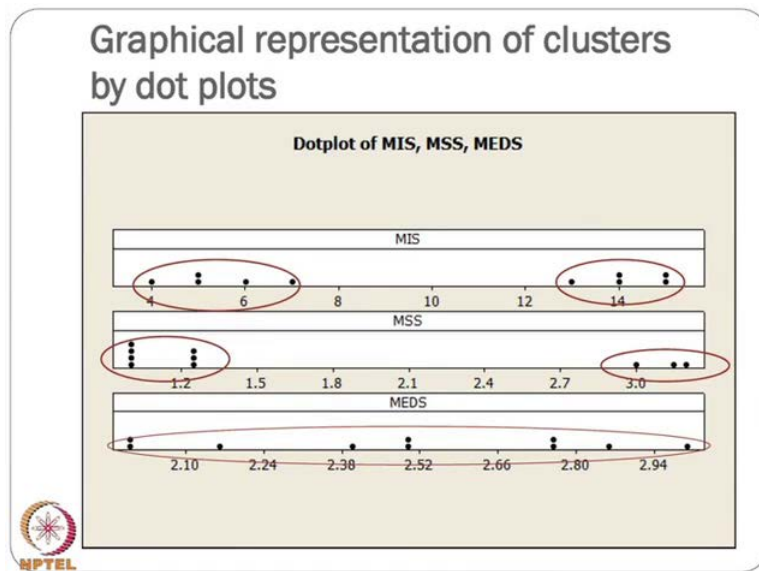
Table 1: Safety performance data

S No.	Department	Mean Incident Score	Mean Severity Score	Mean Equipment Damage Score
1	A	4	1	2
2	B	5	3.15	2.5
3	C	6	3.2	3
4	D	5	3	2.75
5	E	7	1.25	2
6	P	13	1	2.15
7	Q	14	1.25	2.4
8	R	15	1	2.75
9	S	14	1	2.5
10	T	15	1.27	2.85



So, the data collected for example, the 10 department's data mean incident scores so, you are measuring with some measurement system that mean incident score, mean severity score or mean equipment damage score. And you are finding out that A, B, C, D, E, F then P, Q, R, S, T these are the name of the departments and this is mean incident score, mean severity score and mean equipment damage score. So, what is your purpose you want to see that whether all those departments are performing similarly, in terms of these 3 key features so, let us see what we can do this data.

(Refer Slide Time: 12:02)



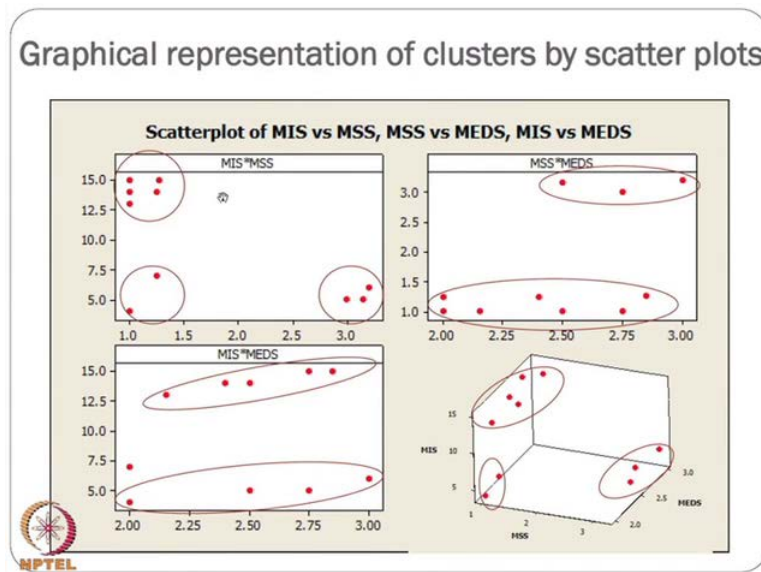
Immediately, what you can do with the data as a safety manager your aim is to see individually the data, this is what is shown here dot plots. So, we have done dot plot for mean incident score, mean severity score and mean equipment damage score. Now, see from mean incident score point of view if you see this dot plot that you were finding, you were finding out the 2 clusters because some are extremely left side, some are extremely right side. So, it is obvious that from incidence move incidence score point of view that means, there are some department which are performing equally and that is 1, 2, 3, 4, 5 departments here and 5 in 1, 2, 3, 4, 5 departments here.

But if you compare any department from here and here you will find out that there is huge difference in terms of their mean incident scores. So, you may be in may be tempted to say that here that ok these are the departments why are I want to actions, similar actions here and here similar actions for these departments, but the actions will definitely different from these group of departments to these group of departments. But here, one mistake is there that if i go buy only MIS, then I am not considering MSS and MEDS so, you are considering individually the performance, but collectively the performance required to be considered.

So, that is possible if we consider all the 3 features at a time we will see little later let us see that from MSS point of view what is the status of the department. From MSS point of view also you

are finding 2 groups, but from equipment damage score point of view it is difficult to tell that there are several groups. If there are several groups then there are many groups this one this, this this, this or there is only one group. Now, as I told you that we have taken here only one variable, one characteristic feature at a time.

(Refer Slide Time: 14:31)

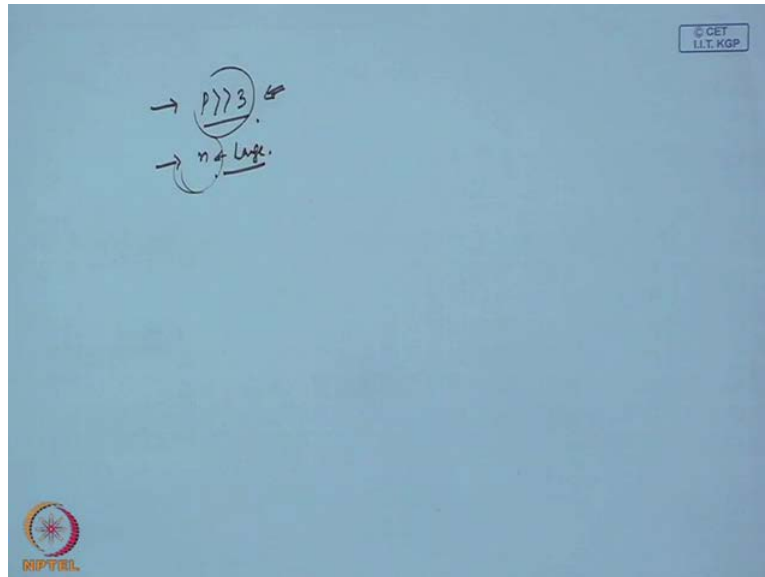


Now, if you take two at a time you will get scatter plot like this and this scatter plot you have seen earlier also, MIS incident score versus a n severity scores you are scatter plots. Now, see from if I consider these 2 characteristic then you are getting 3 clusters, but please keep in mind they are all visually just this clustering the way we have I have done here that circles it is just seeing the that location of all the departments on these 2 dimensional scale, but it is obvious from here that ok there are 3 clusters.

Now, similarly if I go for MIS that is mean incident scores versus equipment damage score you are finding out the 2 cluster, but this one is not able to we are not able to include any clusters then it will be an unique clusters so, it may be again 3 clusters in that sense. Now, if you go by other one MSS and MEDS see 2 cluster, but you may be questioning. Here, that what is the guarantee that these are making 1 cluster, but this is spaded through this so, it is MEDS from MSS point of view they are all at almost at the same level, but MEDS point of view there is a

huge variability. Now, if you combine then together all the 3, then you see you are getting 3 clusters. Now, question comes when there are more than 3 variables.

(Refer Slide Time: 16:10)




So, you number of variable is much, much better than 3 similarly, this is the dimension. Similarly, you will go for large number n will be large definitely, but here as we are working grouping in n based on this, this is the key feature so because it becomes large, what will happen this type of simple plot you cannot make because at 3-dimensional case it is difficult. So, when you go for more than 3-dimension, 4-dimensions onwards you will not be able to visualize just like the pictorial representation what you have seen here. So, we what we mean to say here definitely we are going for more number of variables or more number of characteristics of or more number of features for any items, objects individuals that we want to cluster.

(Refer Slide Time: 17:14)

Criteria for clustering

- Variables to be considered
 - Important variables are to be considered and trivial variables are to be discarded
 - Variables may be of different types based on measurements like nominal, ordinal, interval and ratio
- Similarity and dissimilarity measures
 - It is usually a measure of distance[®] between the objects to be clustered



So, there are 2 criteria first of all you see that what are the variables that you must consider and what will be your similarity or dissimilarity measures. So, under variables to be consider it is clearly mentioned in several books and several researchers also pointed out that, you must be very very careful while finding out what are the variables that you will be considering. Important variables are to be considered and trivial variables are to be discarded variables, may be of different types based on measures like nominal ordinal interval and ratio. This is a very very important one because although we will not be discussing much about the data types, but if your all data are majored in interval of ratio scale then you things become easier.

That means, you will be able to get the similarity dissimilarity majors in much much better manner and in large number of ways, but when you have nominal and ordinal data that getting the distance is little complicated. And you have to go for those data those techniques like i-square continuity table then in case of some other majors because where in the frequency things are coming into consideration that is that will be a different domain from categorical logic data types. But the things become even complicated when you have the mixed data types, some data are nominal, some are interval, some are ratio.

So, when you metric and non metric data both data are mixed in the characteristic features of the variables that will be used in grouping the individuals. So, there is even problem will be

manifold. Today's lecture basically we will be talking about the data the characteristic features which are basically interval of ratio type or metric in nature. Now, then the question comes that similarity and dissimilarity measure it is a usually a measure of distance between the object to be class terms.

(Refer Slide Time: 19:30)

Handwritten notes on the slide:

- $P \gg 3$
- n of large.
- $D dpa = ?$

Distance matrix:

	A	B	C	D	E
A	100%				
B	d_{ab}	100%			
C	d_{ac}	d_{bc}	100%		
D	d_{ad}	d_{db}	d_{cd}	100%	
E	d_{ae}	d_{be}	d_{ce}	d_{de}	100%

For example, if I consider only 5 objects like A, B, C, D, and E then this side A, B, C, D, and E. Now, what is required either you go for distance, if I go for distance between A and A this will be 0 so, this will be 0, 0, 0. Now, this will be d_{ab} , this will be d_{ac} , this will be d_{ad} , this will be d_{ae} . Then this one is d_{bc} , this one is d_{bd} or d_{db} and this is your d_{be} , this one is distance cd this is distance ce , then this will be a distance de , same upper end also same. You must know the distance that is d whether it is in general, if I say d_{pq} you must have a major to get this.


Now, if you go for similarity major then this will be this will changed to differently. Suppose, if I say in 100 point scale similarity then it will be 100, 100, 100, 100 and 100 and here or 1.0. If I say 1 is the most similar case then 1, 1, 1, 1 like this and this values also you require to find out, but actually this similarity is just opposite to this similarity sometimes we may say 1 minus this is also my similarity measure, but there are several measures which can be used and we will see later on.

(Refer Slide Time: 21:32)

Variables representation

Object	Variable	Variable	Variable
	1	2	p
1	x_{11}	x_{21}	x_{p1}
2	x_{12}	x_{22}	x_{p2}
\vdots	\vdots	\vdots	\vdots
n	x_{1n}	x_{2n}	x_{pn}

S No.	Department	Mean Incident Score	Mean Severity Score	Mean Equipment Damage Score
1	A	4	1	2
2	B	5	3.15	2.5
3	C	6	3.2	3
4	D	5	3	2.75
5	E	7	1.25	2
6	P	13	1	2.15
7	Q	14	1.25	2.4
8	R	15	1	2.75
9	S	14	1	2.5
10	T	15	1.27	2.85




Then we I will just formally introduce the like the data matrix in terms of x and this is what is the data matrix for different individual and for the case what we are discussing for this, this is the data matrix. So, this is what the data we are planning to collect is and here this is the data what is already collected. And here, our aim is not grouping the variables like in factorial our aim is grouping the objects with the help of the variables.

(Refer Slide Time: 22:17)

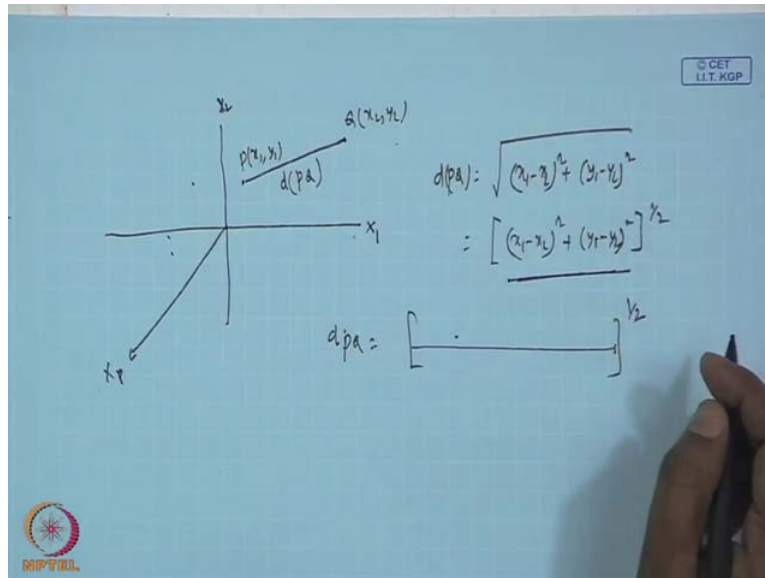
Distance measures

- Euclidean distance
$$d(i, j) = \left\{ (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2 \right\}^{1/2}$$
- Squared Euclidean distance
$$d(i, j) = \left\{ (x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2 \right\}$$



What are the different distance measures, all of us know the Euclidean distance.

(Refer Slide Time: 22:26)




So, if I have x and y , then suppose this is this is my P x_1 and y_1 and this one my Q x_2 and y_2 all of us know that what is the distance PQ $d(PQ)$. So, we can write $d(PQ)$ equal to square root of x_1 minus x_2 square plus y_1 minus y_2 square. Now, if you go for that means this is like this x_1 minus x_2 square plus y_1 minus y_2 square this to the power of half. So, if you go for suppose there are many more dimensions that x_1, x_2 like x_P different dimensions are there so, what will happen at the ultimately. You will get this will be extended to that mean $d(PQ)$ that will be extended to your P dimensions that to the power half.

So, if you see that what we have given here that $d_{ij} = \sqrt{x_i^2 - x_j^2 + y_i^2 - y_j^2}$ like this up to i P square, this is Euclidean distance. Many times you may be interested to prove more weightage to the distance, then you may not go for Euclidean you may go for squared Euclidean distance. Similar, to the first one but only this to the power half is not there.

(Refer Slide Time: 24:14)

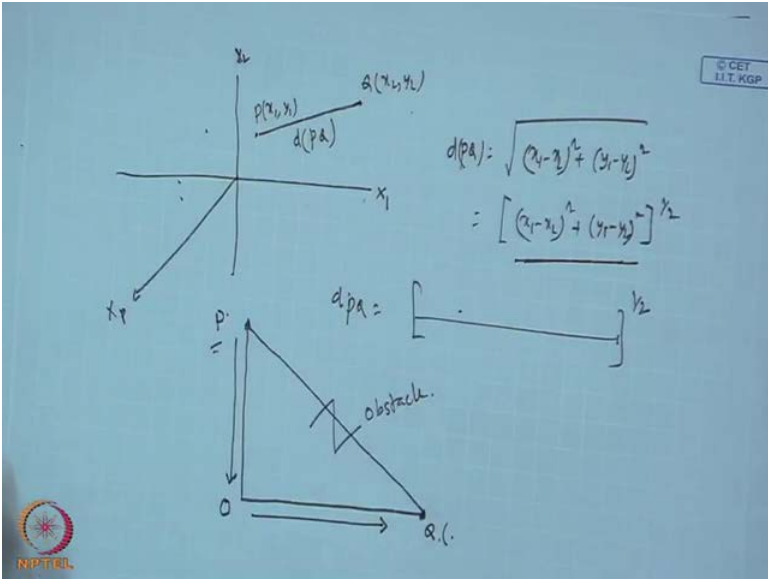
Distance measures

- Manhattan distance
$$d(i, j) = |(x_{i1} - x_{j1})| + |(x_{i2} - x_{j2})| + \dots + |(x_{ip} - x_{jp})|$$
- Minkowski distance
$$d(i, j) = \left\{ |(x_{i1} - x_{j1})|^m + |(x_{i2} - x_{j2})|^m + \dots + |(x_{ip} - x_{jp})|^m \right\}^{1/m}$$



So, there is another distance which is known as Manhattan distance, this Manhattan distance is similar to like this.


(Refer Slide Time: 24:25)



© CET
I.I.T. KGP

$$d(P,Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$
$$= \left[(x_2 - x_1)^2 + (y_2 - y_1)^2 \right]^{1/2}$$

$d(P,Q) = \left[\dots \right]^{1/2}$

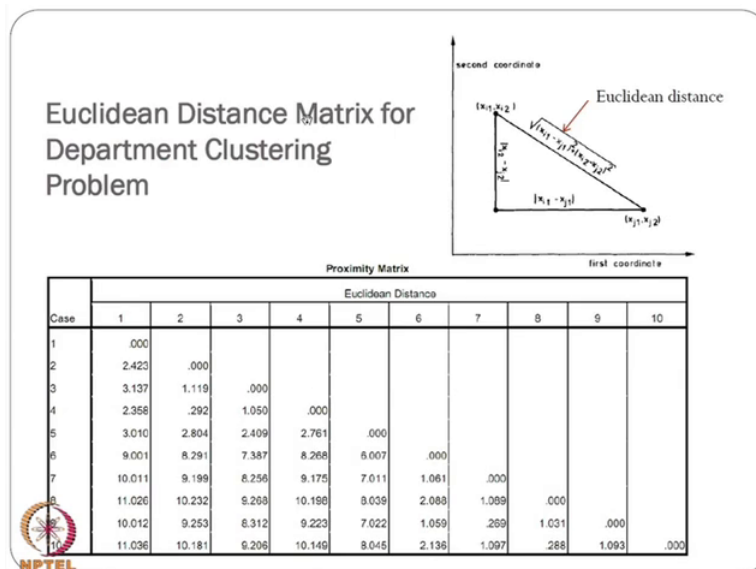


Suppose, you are here you want to come here suppose this is my P, this is my Q let this one some value and this is also some co-ordinate value to dimension case. Now, the shortest one is this, this is the shortest one, but what you will do when you come here suppose, there is obstacle what

you will do and this is the only way to go you come here and then follow this. So, like this Manhattan city and that roads, you just like Manhattan city roots you are coming here. That means if this is 0, you are first covering this distance then this distance.

Now, it is what is given that the you this distance is the mode value of these things this one plus this plus similarly, there are more number of features or variables will be adding up to P features. Minkowski distance is there is a same thing you just seen, but it is every distance is first score power to m and then it is basically like geometric mean. Then what you have done you have basically, again make it in the original scale so all everything will be made that some will be made that root to the power m.

(Refer Slide Time: 26:15)




Now, for the data said that what we have discussed so far that is ((Refer Time: 26:20)). So, if you use Euclidean distance you are measuring this for A to 1 to 1, 2 to 2 like this we have measured the Euclidean distance and these are the values. See for the diagonal elements are 0 of diagonal elements having values some of the values are high, some of the values are low and what is required we want basically to group this 10 departments based on this distance.

(Refer Slide Time: 26:54)

Squared Euclidean Distance Matrix for Department Clustering Problem

Proximity Matrix

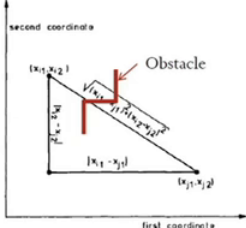
Case	Squared Euclidean Distance									
	1	2	3	4	5	6	7	8	9	10
1	.000									
2	5.873	.000								
3	9.840	1.253	.000							
4	5.563	.085	1.103	.000						
5	9.063	7.860	5.803	7.625	.000					
6	81.023	68.745	54.563	68.360	36.085	.000				
7	100.223	84.620	68.163	84.185	49.160	1.125	.000			
8	121.563	104.695	85.903	104.000	64.625	4.360	1.185	.000		
9	100.250	85.623	69.090	85.063	49.313	1.123	.073	1.063	.000	
10	121.795	103.657	84.747	103.003	64.723	4.563	1.203	.083	1.195	.000



If you go for squared Euclidean distance then every item here, these are basically squared and you are getting this value.


(Refer Slide Time: 27:06)

Manhattan Distance Matrix for Department Clustering Problem



Proximity Matrix

Case	City Block Distance									
	1	2	3	4	5	6	7	8	9	10
1	.000									
2	3.650	.000								
3	5.200	1.550	.000							
4	3.750	.400	1.450	.000						
5	3.250	4.400	3.950	4.500	.000					
6	9.150	10.500	10.050	10.600	6.400	.000				
7	10.650	11.000	10.550	11.100	7.400	1.500	.000			
8	11.750	12.400	11.450	12.000	9.000	2.600	1.600	.000		
9	10.500	11.150	10.700	11.250	7.750	1.350	.350	1.250	.000	
10	12.120	12.230	11.080	11.830	8.670	2.970	1.470	.370	1.620	.000




If you go for Manhattan distance as I told you that obstacle is there you have to come here and then go there. So, ultimately you will be getting this distance values 1 to 1, 1 to 1 in the sense every item to item and or every object to object distance you are getting.

(Refer Slide Time: 27:29)

Minkowski Distance Matrix for Department Clustering Problem

Proximity Matrix

Case	Minkowski (2) Distance									
	1	2	3	4	5	6	7	8	9	10
1	.000									
2	2.423	.000								
3	3.137	1.119	.000							
4	2.358	.292	1.050	.000						
5	3.010	2.804	2.409	2.761	.000					
6	9.001	8.291	7.387	8.268	6.007	.000				
7	10.011	9.199	8.256	9.175	7.011	1.061	.000			
8	11.026	10.232	9.268	10.198	8.039	2.088	1.089	.000		
9	10.012	9.253	8.312	9.223	7.022	1.059	.269	1.031	.000	
10	11.036	10.181	9.206	10.145	8.045	2.136	1.097	.288	1.093	.000




Then Minkowski distance matrix that is Manhattan distance to the power m we are saying it is Minkowski within bracket 2. That means, we are considering m equal to 2 and this is the distance.

(Refer Slide Time: 27:42)

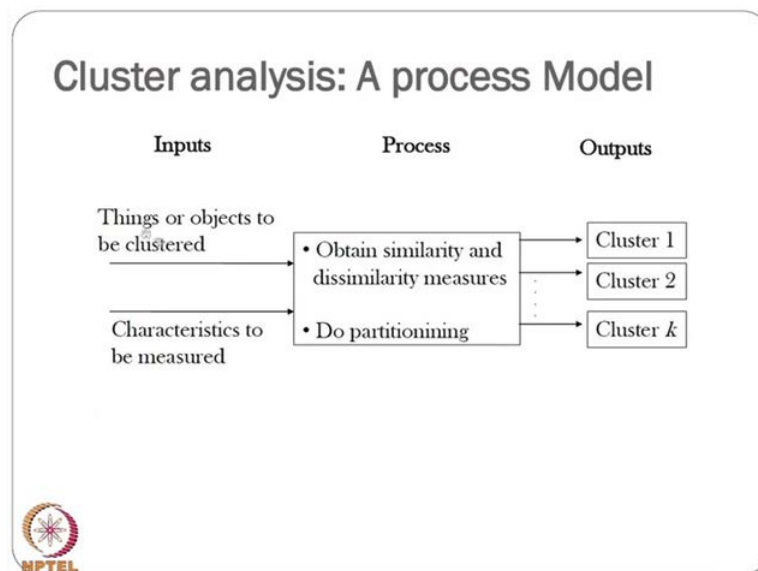
Clustering Algorithms

- Hierarchical joining algorithms
- Nonhierarchical joining algorithms such as k-means clustering



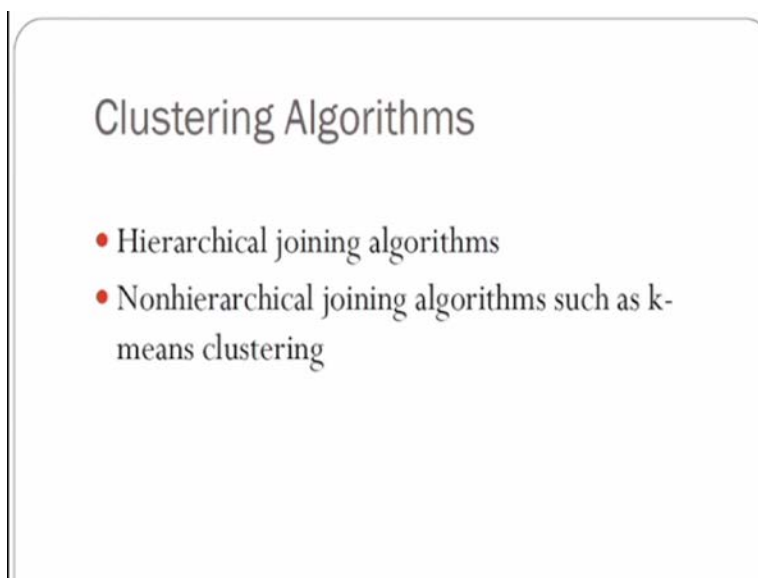
So, now then what we have done ultimately if I go back to the process model you will find out.

(Refer Slide Time: 27:54)



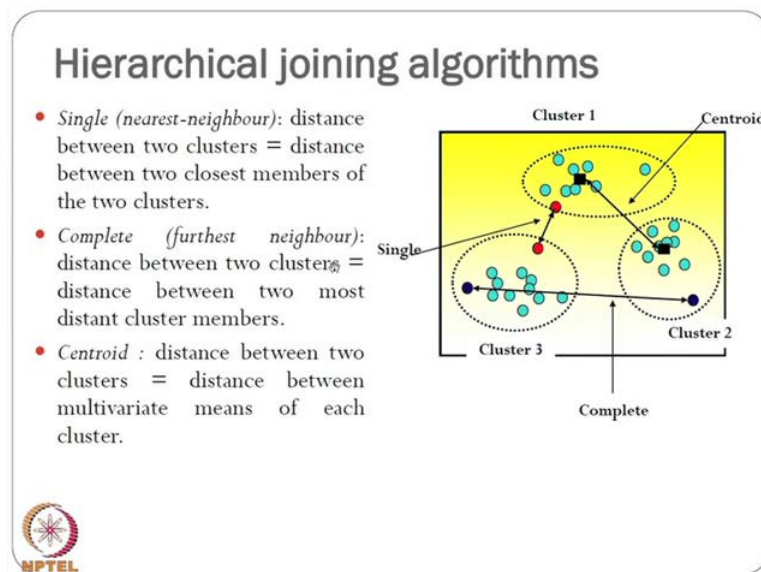
You will find out that we require things or objects to be cluster and we require characteristics to be measured that we have already seen for the safety case, data case. Things are dependent to be cluster characteristics are 3 characteristics MIS SS MSS and MEDS and then we want some similarity or dissimilarity measure, we are saying that we are going to the distance measure either Euclidean or squared Euclidean or Manhattan distance or Minkowski distance. And then what we require we require a partitioning algorithm.

(Refer Slide Time: 28:38)



So, there are hierarchical joining algorithm, nonhierarchical algorithms are hierarchical algorithm are ((Refer Time: 28:46)) that cluster algorithm and nonhierarchical algorithm like k-means clustering algorithm all those things are there. So, first we will discuss about the hierarchical algorithm, joining algorithms.

(Refer Slide Time: 29:04)



There are many ways to group objects in hierarchical joining algorithms, the names are single linkage or nearest neighbor algorithm, complete linkage or furthest neighbor, centroid linkage then average linkage, median linkage and ward linkage. So, these are the different types of algorithm for joining the objects into different groups. Let us first understand what is single linkage algorithms?

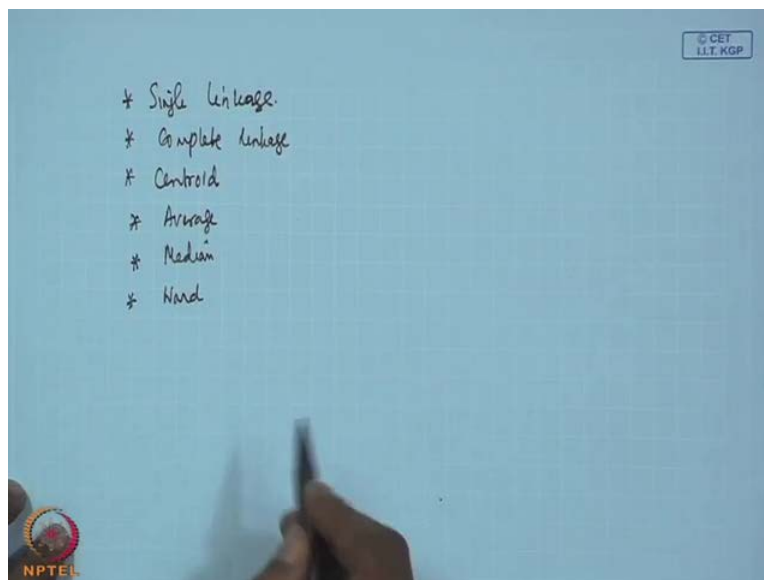
Single linkage algorithm is here distance between 2 clusters or distance between 2 closest members of the 2 clusters. For example, you just think of a 2-dimensional case where all the objects here and the items are put in the appropriate location based on the characteristics features of that 2-dimensional, 2-dimensional data matrix. Now, you see that arbitrarily I have given some groups this is one group this is another group this is another group so, we are saying this is your cluster 1, this is your cluster 2 and this is your cluster 3. Now, question comes how do I make this groups or how do you, you can make this group here, if I were talking about single

linkage then you were talking about the distance between cluster 1 and cluster 3 is the distance between the 2 nearest objects of the 2 clusters so, obviously this one.

Now, you may be wondering that there is no cluster started then how suddenly you make this cluster and ultimately how these things are coming. So, that I will discuss little later, but you first you just understand here little abstraction you make that ok it is, there is cluster different clusters possible. And the distance in single linkage means the distance between the 2 clusters is the nearest neighbors, distance between the nearest neighbors. Now, when we talk about the complete linkage you are talking about the furthest neighbor.

That mean the you find out the member in cluster 3 for example, in this case and member in cluster 2 who are the furthest point it not they are not nearest they are little furthest the maximum distance where they are. Then when we were talking about centroid you see the distance between 2 clusters is distance between multivariate means of the clusters. So, there are several variables here actually 2 variables are there you are basically, considering and the data mean that value for each of the variables and then accordingly we are making the centroid and then finding out where it is.

(Refer Slide Time: 32:40)

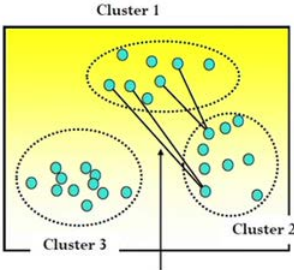


So, let me write down then that we have joining algorithm like single linkage, complete linkage, centroid then you see that average linkage, then median linkage, then ward.


(Refer Slide Time: 33:15)

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.

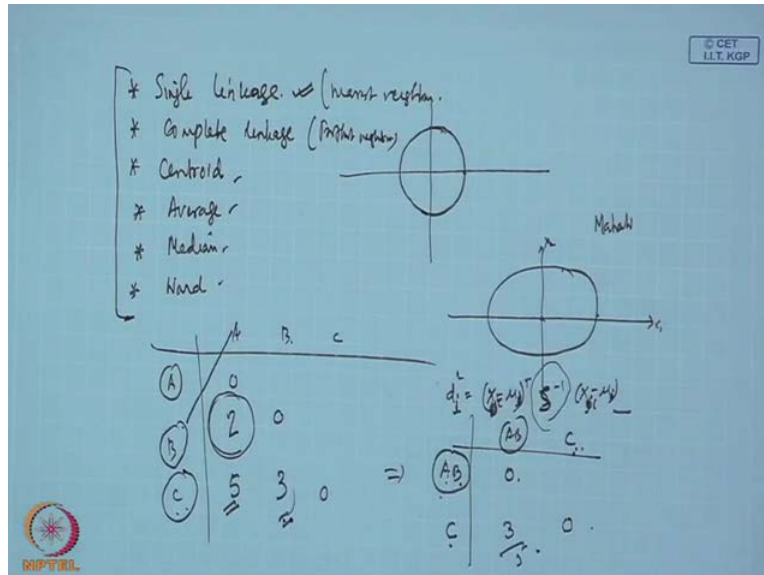


Mean/median/adjusted mean of all pairwise distances



Now, if you see the average then you come here you see the distance between 2 cluster is average distance between all members of the 2 clusters. So, there are so many members I can see here so many items here you find out the average of these items, average of these items and then find out the distance. And median case it is median distance between all members of the clusters so, all of you know median so, find out the distance and then get the arrange from smallest to largest and get the median value. What this tells that average distance between all members of the two clusters with adjustment for covariances.

(Refer Slide Time: 34:02)



Actually, if you can recall that when we have discussed the statistical distance I say when if you consider equivalent distance that all points on the circle are similar, in terms of equivalent distance, but if there is suppose an ellipse. The data resemble ellipse you cannot say based on Euclidian distance where they are equidistance, but if you see that mahalanobis distance, mahalanobis distance. Then what is happening here ah that all points on the ellipse are equidistance because the variability across x_1 and the variability across x_2 these are weighted.

And that is why you got like this that $x - \mu$ transpose sigma inverse, I think here we will use mahalanobis symbols $x - \mu$. So, if I write that for one variable this then this, this is my d_j^2 that is i, j then j, i then j, i then j , just similar manner that i, j or d_i^2 you write x_i . That is straight way the multivariate of derivation then you write μ here i and μ that is better. So, here in warling case what happened that the 2 cluster with the adjustment of covariance's here, it is the adjustment of variances and covariance's in terms of the symbols. A symbol it is just what distance is similar to this because of you see that their adjustment of covariance's between one variable to another variable.

So, these are nothing but to find out distances and these are applicable only when you have made groups. Suppose, you have not made any group there are some observations 1 to n and you know

the distances for example, A, B, C and A, B, C so, you found out the distances this is 0, 0 and 0. Suppose, this distance is 2 this is 5 and this distance is again 3 here, what you will do first, you first find out the smallest distance, this is my smallest distance so, this two can be grouped A and B can be grouped. So, if I do this then what happen you are now deter set it initially deter set when it is not group everything is a cluster, A is cluster, B is cluster, C is cluster. When it is grouped A B becomes one cluster and that C remains, then A B and C that remains.

Now, A B to A B will be 0 and C to C will be 0 what will be the distance between A B and C there you have to choose you cancel any one of this method and then find out. For example, if you use complete linkage, single linkage let it be if you use single linkage which is basically, known as nearest neighbor, nearest neighbor then what is required you find out the distance between A and C, distance between B and C then find out what is the minimum value. If I go for this A and C this is 5, B and C this is 3, then definitely 3 is the minimum so, your distance is 3. if go for single linkage if you go for complete linkage that is furthest neighbor, furthest neighbor then what will happen ultimately we will find out that the maximum of these two will be taken. So, similarly for centroid average and there are other criteria given and we have to do this now we will I will show you on this.

(Refer Slide Time: 38:10)

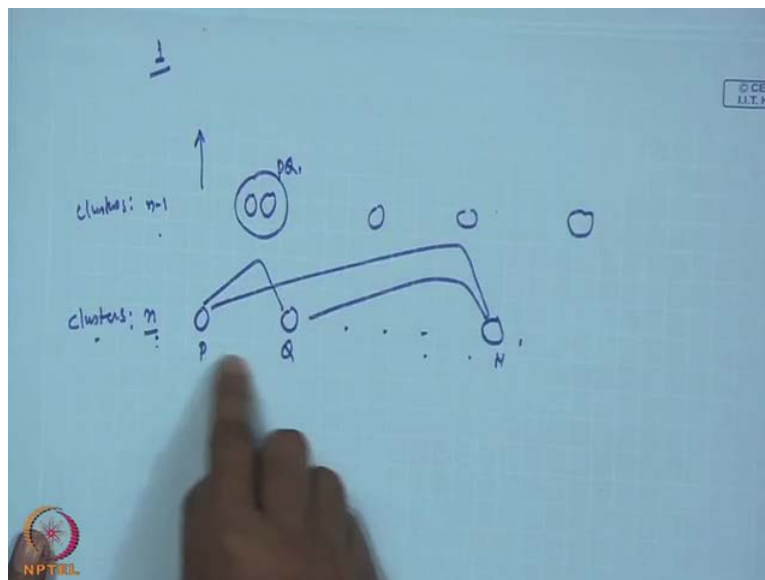
Agglomerative Hierarchical Clustering Algorithm

- 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 P and Q be d_{PQ}



Now, what are the steps, steps to follow in hierarchical clustering, step one; identify the variables and objects then you have to collect data, then select similarity or dissimilarity measures what you want usually go for dissimilarity measures distance. Then obtain the distance matrix start with n clusters please keep in mind we are saying that individual items or the objects or what I can say things that you are trying to group they are unique. So, if there are n objects in clusters starting point if n objects in clusters n clusters.

(Refer Slide Time: 39:11)



Then you see the distance from here, for every, every distance you are seeing you find out the minimum distance and then you group so, if this 2 are minimum then this will be grouped then things will be coming like this. So, when we are talking about hierarchical algorithm team clustering what is happening initially there are n objects so, when you group one only the two of the original n they are grouped. Now, there will be n minus 1 cluster so, if I say the number of cluster here is n here, clusters are n minus 1. So, in this manner your things will be reducing n minus 2 like this finally there will be 1 cluster where all will be grouped so, that is what is we required to be done here in hierarchical study so first you find out the distance.


So, you take each one a cluster find out the distance between the cluster and then you take every pair of cluster, pair of clusters from that things and then you think that the most similar clusters are P Q and this can be grouped that like this. If this is P, this is Q suppose this is your m what is

happen this is most similar so, this two are grouped, this two are grouped. When you group 2 similar objects so, your number of cluster reduce to n minus 1 that is what is given here.

(Refer Slide Time: 41:04)

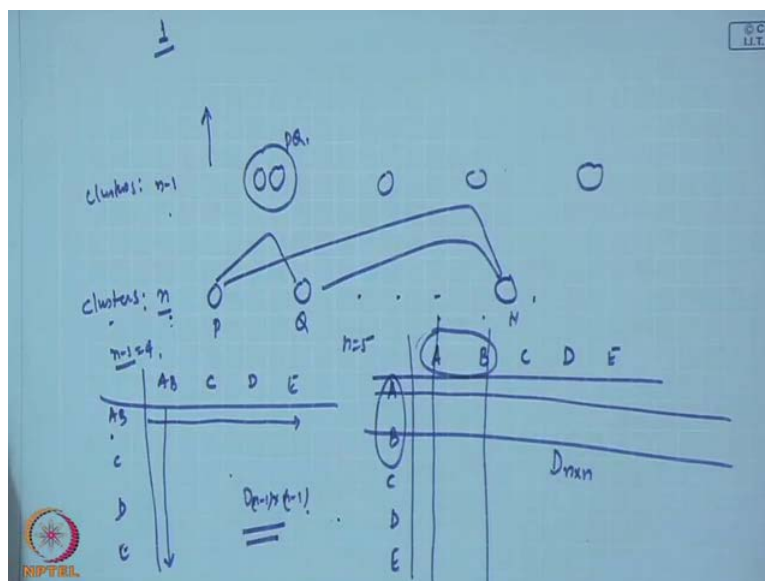
Agglomerative Hierarchical Clustering Algorithm

- Step 7: Merge clusters P and Q and label the newly formed cluster as (PQ) . Update the entries of the distance matrix D by
 - Deleting the rows and columns corresponding to clusters P and Q and
 - Adding a row and column giving the distances between cluster (PQ) 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



Merge cluster P and Q and label the newly formed cluster as PQ update the entries of the distance matrix by matrix d by doing these things. So, deleting the rows and columns corresponding to cluster P and Q .

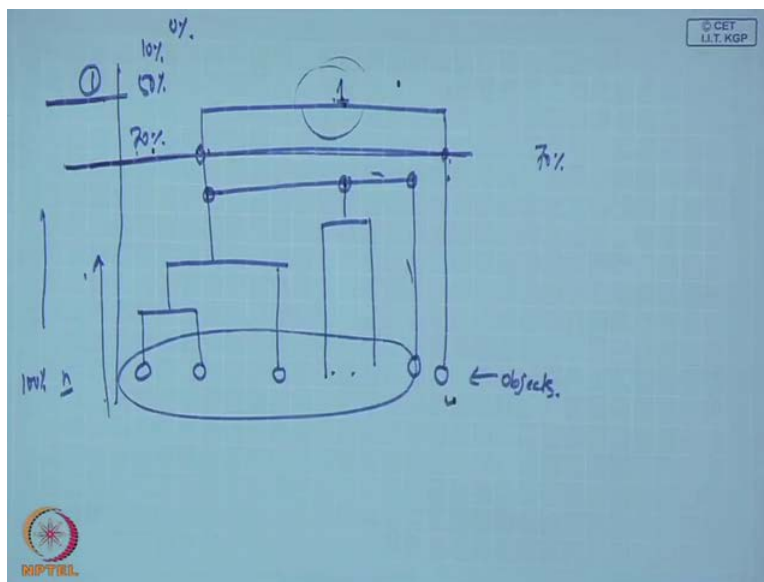
(Refer Slide Time: 41:25)



Basically, if I have 5 items so and you find out that A B are most similar at the first level distance is less, then your number of cluster initially n equal to 5 now, these two will be grouped then C D E then A B C D and E. So, now cluster is here n minus 1 that is 4 now, what will be the entries that means the distance values original distance is $d \ n \ cross \ n$. Now, it will be $d \ n \ minus \ 1 \ cross \ n \ minus \ 1$, we have to find out this that is what is said here that deleting the rows and columns corresponding to cluster P and Q here A and B.

So, A rows you delete then adding a row and column giving the distance between the P Q so, like A B A B you are adding 1 row and column, 1 row for this you are adding 1 column also you are adding. So, you are deleting 2 here 2 rows and 2 columns adding 1 rows 1 columns and ultimately $n \ minus \ 2 \ plus \ 1$, that is $n \ minus \ 1$ is coming here. Then you repeat what we have done in step number six and step number seven for all $n \ minus \ 1$ times when all the objects are grouped is when you are coming you have to go up to this 1 cluster level, when you are reaching there so you stop. Then final one is record the identity of clusters that are merged at the levels at which the merger takes place.

(Refer Slide Time: 43:36)



So, from the n level what will happen ultimately we will find out a situation that from here n levels you will basically, suppose first you merge these two then let me merge this one so, then let you merge another two here. Then finally suppose you are merging this two may be another

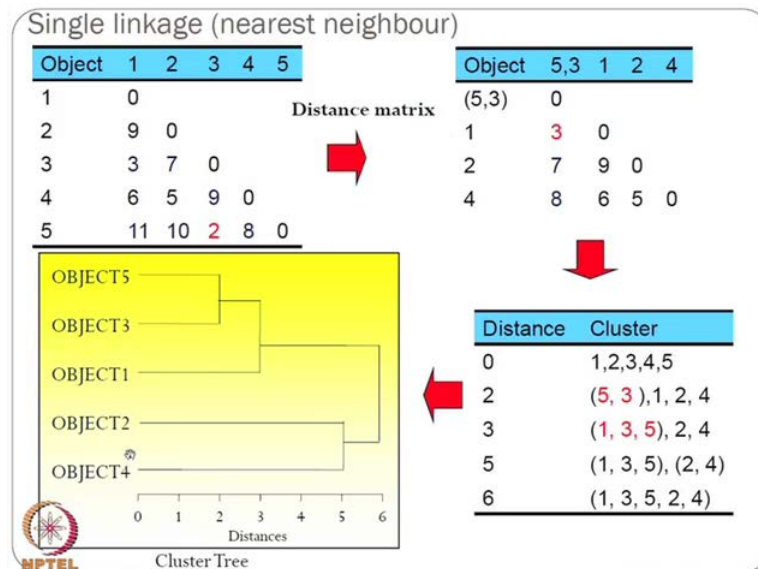
one here, then let the final merger is coming to place here this is the final one, 1 cluster this is n cluster. So, long you are not coming to this you are not stopping these are objects then object to groups then final this is happening. Now, where you will be stopping that all depends on what is the distance between the objects in a group, the maximum distance between the objects in a group or the single linkage that single linkage will come between groups, but what distance between in the same group the maximum distance you say that you are accepting or not.

So, then what will happen as you are going up to lower number of clusters actually distance between the objects within a cluster, within a cluster is increasing for every cluster it is increasing, but after certain distance you may not accept that distance. So, this is little difficult to understand, but if we go by similarity measure what I mean to say that if I make cluster here, then this is made one cluster, second cluster, third and four. One I think this one is, this one is taken care of if I come here this 3 are taken care of you are making 2 clusters. So, then you see that what is the similarity measures of all these objects within this cluster and here, it is 1 then there is 100 percent similarity within.

If similarity, reduces from that means here it is 100 percent similarity now slowly it will reduce it may be you have to find out here how much it is, is it 50 percent is it 10 percent is it 0 percent is it 70 percent what you want to keep. Suppose, at this level if it is 70 percent are you happy with 70 percent similarity between the objects within a group if you are happy in this case you can keep 2 clusters. All items here, here things are such that all the items except this one are grouped under cluster 1 and this is cluster 2. So, that is step nine record the identity of clusters that are merged at the levels at which the merger takes place.

Now, based on this value you can give some identity that is cluster 1 it may be give some other interpretable identity and may be this groups are of some type this group is of another type. For example, from safety point of view example I have given may be these are low accident prone situation I have are departments and this is highest high accident prone situation. So, that mean the 2 cluster low accident group department, high accident group department low medium high hence many ways you can find out.

(Refer Slide Time: 47:17)



Now, I am explaining here how this actually single nearest neighbor or single linkage is working same example similar example, I have already given you, but it is more formal here you just see the example and try to understand fully that there are 5 objects and this is distance measure. At the first instance what you required to do you have to find out which one is the smallest which pair of objects are having the least distance. Here, the 2 this 2 is the least distance value among this values forget about 0 because we are talking about item to item distance not the same item because same item distance is 0 always, same item distance to that item it will be 0, but we are talking about the other things so it is 2.

So, 2 is 3 and 5 the distance between 3 and 5 is 2 so, see 5, 3 are grouped here 5, 3 grouped here what we have done you deleted these column as well as this column and you added 1 column 1 row and 1 column here. So, 1 row here these column as well as this row you have deleted and similarly, that means what I can say the 3 and 5. So, this 1, 3 this 3 and 5 then 3 and 5 this 2 will be deleted and one more will be added here. Now, question is what is the distance between the cluster 5, 3 and 1 so, that nearest neighbors is that you find out the distance between 1 and 5 and 1 and 3 take the minimum 1 then 1 and 5 is 11 and 1 and 3 is 3.

So, minimum 1 is 3 you have taken 3 and rest of the items the distance similarly 5, 3 versus 2 similarly 5, 3 versus 4 you have to find out. So, 5, 3 versus 2 that mean 2, 5 and 2, 3 now if you

go by 2 and 3 this is 7 then you required to find out 2 and 5 the 2 and 5 same by 10 so, the minimum 1, minimum of 7 and 10 is 7 so it is coming seven. So, similarly 8 rest of these things are all ready available here just because this is 2 to 1 and 2 to 4 and 1 to 4 this distance already it is here so, if I say 1 to 2 that is 9, 9 is even here like this.

So, your new distance matrix is this now we have what you want here you have to find out that what are the members or objects that can be grouped further so here 5, 3 is 1 object or 1 group so you will be treating as 1 object. So, then what we will do we will again find out which one is the minimum value, distance value so here you see out of these 3 is minimum and this 5, 3 and 1. So, here you are that is why in this case 5, 3, 1 is grouped and 2, 4 remains so that mean, how many clusters you are making here, you are making here 1, 3, 5 one group 2, 1 another group, 4 another group so, ultimately it is 3 clusters.

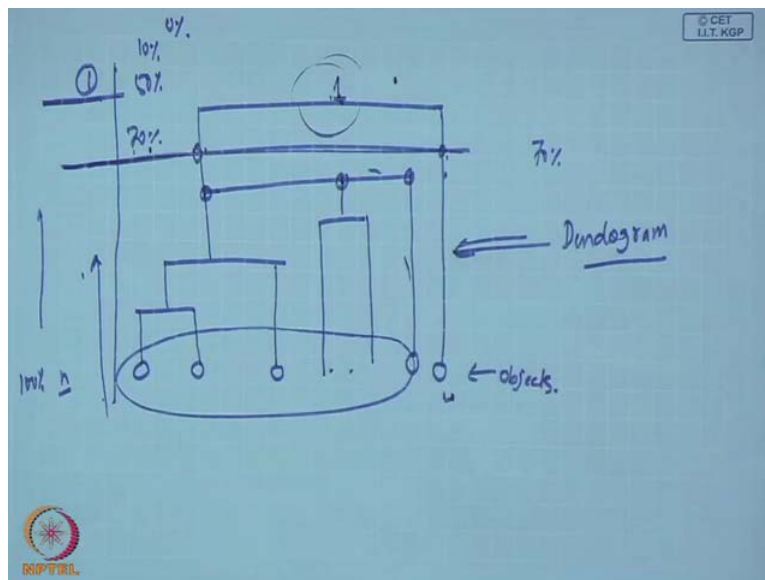
In cluster if I say this is my cluster 1 then in this case 1, 3, 5 of these 3 objects are grouped then cluster 2 is having only 1 object that is 2 number 2 object and class 3 is having only 4 objects. Now, distance if you see here see we are using single linkage so, in the first case when there is no groups in the sense that no 2 objects are in one group here, every object is representing one group in that case there is no distance the 0 the diagonal elements. When I come to this 4 cluster the second this table when you find out that what is the minimum distance between what is these is distance 3 and 5 the distance between this. So, there are two items so you are getting one distance that is 2.

Now, come to the third one when the 3 members are grouped under one cluster 5, 3, 1 then the distance we are talking about the 3 because it is coming from this 5, 3, 1 so, it is not that we are taking the maximum 1 it all depends on which linkage you are using. So, earlier I said the maximum 1 and that also possible when you go for complete linkage. So, 3 then what will happen here you will find out again the similar table will come and you find out that 2, 4 will be grouped so, the two cluster 1, 3, 5, 1 cluster and 2, 4 another cluster and the distance is 5 and finally, 1 cluster and distance is 6, this is the minimum distance between the members of these.

Now, see that same thing can be represented in a diagram like this here, when distance is 0 all the objects are forming one cluster each there are 5 objects so 5 clusters. So, at the distance level one that no grouping possible only at the distance level two 5, 3 are grouped so, at these level

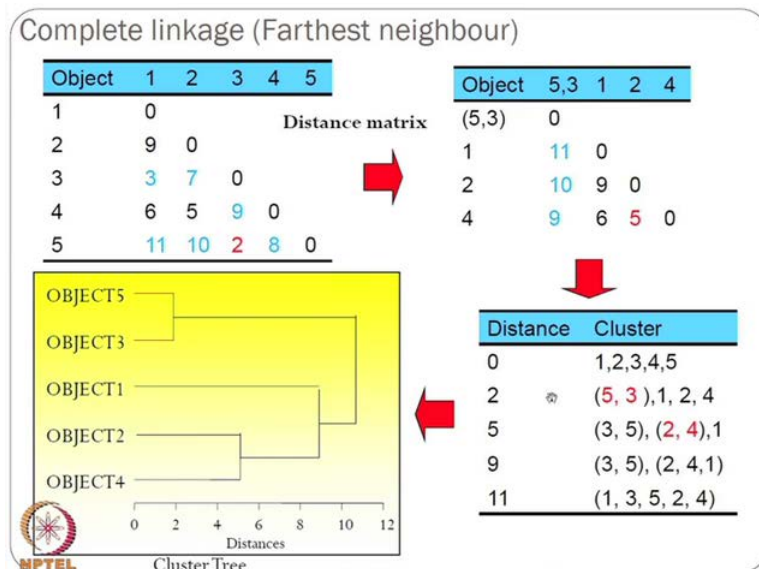
you are sacrificing a distance of dissimilarity of 2 between these two members, but in this process you are gaining one way. That way is your number of clusters reduced from 5 to 4 and at the distance of 3 your number of clusters reduced to 3, but your distance sacrificing distance is 3 and but at the distance of 4 it is not like this nothing possible, but at the distance of 5 you are making 2 clusters at the distance of 6 you are making 1 cluster what is this diagram known this diagram is known as dendrogram.

(Refer Slide Time: 54:02)



Actually, this is what I told you the same diagram it is just 90 degree rotated there so, this is known as dendrogram. So, then you see that we will go for complete linkage with the same example and you see that the object is the process steps are similar except the calculation of the distance major the distance may be here. The maximum distance between the between the between the members of the groups two groups.

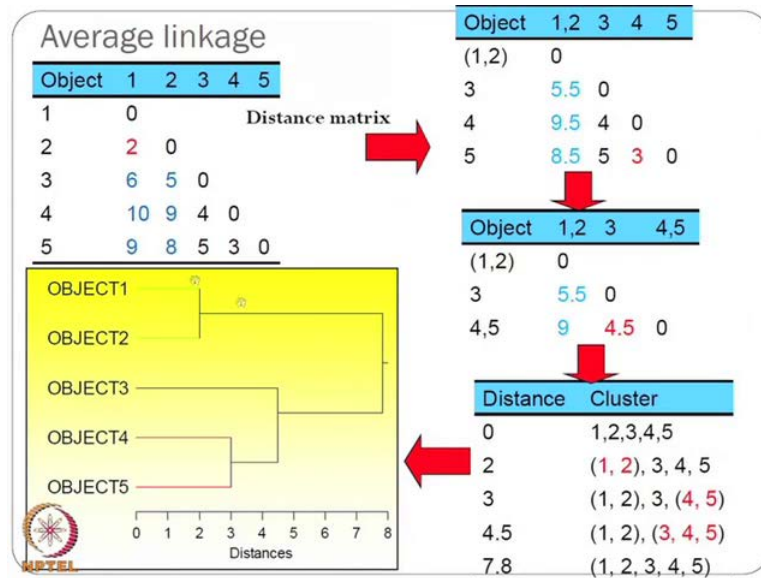
(Refer Slide Time: 54:37)



In that case if you go on grouping your grouping is 5, 3 again first then your 2 and 4 then finally, 1 is coming and finally all will be grouped, but here distance major if you see that you are sacrificing how much distance 11 distance for one group because here you have considered the maximum distance and when you see this table here, is the maximum distance is 11. So, but if I compare this with the earlier one here also we found out that 5, 3 final one if you say that 5, that the second one 5, 3 and 1, 2, 4 here 5, 3 and 1, 2, 4, but that difference is coming.

Here, 1, 3, 5 is grouped and then 2, 4 remain as it is, but here 3, 5 and 2, 4 again grouped and so that will be change in the dendrogram in grouping. And it may so happen that you may not get that exact grouping if you go for by exact grouping what I meant is that same grouping if you go for different linkages, if you go for call single linkage you may get some type of grouping, you go for complete linkage some other grouping.

(Refer Slide Time: 56:05)



If you go for average linkage like this it is a different example, other example you are also getting a different grouping, but please keep in mind the message is that if you can calculate the distance between the objects within a group and between the groups also. That is also required because we required similarity as well as we required highest level of dissimilarity. So, then using this hierarchical clustering algorithm you will be able to find out the dendrogram and what level of similarity or distance you want to keep the, what I can say you want to have this amount of similarity you are satisfied with this amount of similarity then fine you go for that level.

So, what I will do in the next class I will continue this cluster analysis I will go for some other algorithms and then for hierarchical clustering and then we will discuss K-mean clustering. Then we will see one case study and using a spaces and using many step how it can be done.

Thank you very much.