Indian Institute of Technology Madras
Presents

NPTEL
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING

Pattern Recognition

Module 03

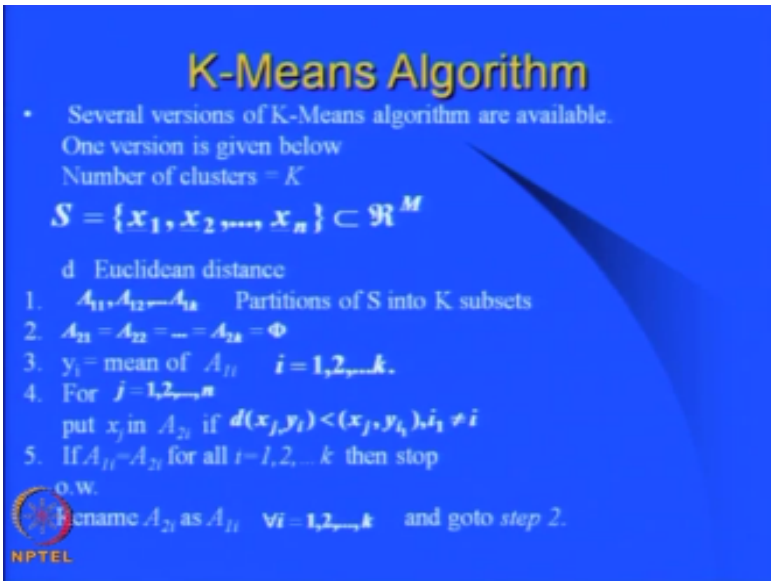
Lecture 02

K-Means Algorithm and Hierarchical Clustering

Prof. C.A. Murthy
Machine Intelligence Unit
Indian Statistical Institute, Kolkata

We were discussing about the properties of a minimum within cluster distance criterion in the last class then a way of implementing the criterion in practice though women are we are now we may not be assured of getting optimal clusters is by using k-means algorithm as I mentioned there are several versions of k-means algorithm the version that I am going to give you is by 4G 1965 FORGY 1965.

(Refer Slide Time: 00:46)



So this is the version the word K it denotes the number of clusters and we are going to assume that number of clusters is known to us K is known to us and we have got these n points in M dimensional space and D is Euclidean distance basically what we do is that we generally we consider a partition of the whole space into K subsets the partition is I have just denoted by $A_{11}A_{12}$ A_{1K} it partitions S into K subsets now I am going to assume A_{21} A_{22} and A_{2K} is equal to null set.

What are these A_{12} I will come to it later then what I do is that I consider mean of Y I as I mean of A1i as Yi, so y1 is mean of A_{11} y2 is mean of A_{12} and YK is mean of A_{1k} okay, so I calculate all the K means then what I do is that I take points from J is equal to 1 to N the points in s every point I put it into I take the first point x1 calculate the distance of x1with this K means whichever means the distance is minimum I will put it into that particular A_{2i} up goes with x1 the minimum distance is occurring for the mean y2.

Then I will put it into A_{22}okay put xj in $A2_i$ if distance between xj + y I is less than here D is missing I should have written here D, so whichever mean the distance is minimum will put into corresponding A_2 cluster so like that we have now a new partition A21 A22 A2K, now check whether A1i = A2i for all I if it is if those two are same you stop it otherwise you rename a twice as A1i and go to step 2 that means now you have a nice a twice are again null sets, now you have to find the mean of A1 a that means you are going to repeat this process I suppose the algorithm is clear to you probably you have seen this algorithm at a few places.

So given this algorithm there will be many questions the first question is does it converge the first question is does it converge the criterion is A1i = A2i for all I then you stop it otherwise you should go on and on doing it does it converge well I would like to give a slightly we concern the answer is that it were just converges I would like to put it within quotes in the sense that people have not found a data set where it is not converging but the proof for convergence they are not exactly satisfactory.

So and there is also another problem the problem is that suppose it is converging after say 10 to the power of 10 iterations and you will surely not go to up to those many iterations, so usually what people do is that they fix the number of iterations beforehand maximum number of iterations they fix the maximum number of iterations beforehand and if it converges before that that is fine otherwise they will go up to the maximum number of iterations and then they stop it okay.

So now this is one issue and there is another issue note that I have taken a partition here now I have taken one partition maybe someone else can take some other partition and let us say the algorithm is converging in both the cases with my partition and with his partition do you think the final results will be seen the answer is they need not be same, I am repeating with two different initial partitions it is not necessarily true that the final results would be same they can be different that is one.

And it also basically provides you convex shaped clusters and so naturally if you have non-convex clusters you may not be able to get those clusters using this method and this was given in 1965 in 1967 Mclean came out with another algorithm that is known as Mcqueen's k-means algorithm the paper was published in fact was presented in Berkley symposium on statistics mathematics and probability, okay it was presented in the symposium1967 but class symposiums they are supposed to be the best symposiums in the world in mathematics and statistics but play symposium, so this was a McQueen's paper was given in 1967 he made a small modification is that suppose take point x1 say x1is there in the original cluster say A_{11} okay suppose x1 is there in the original cluster say A_{11}.

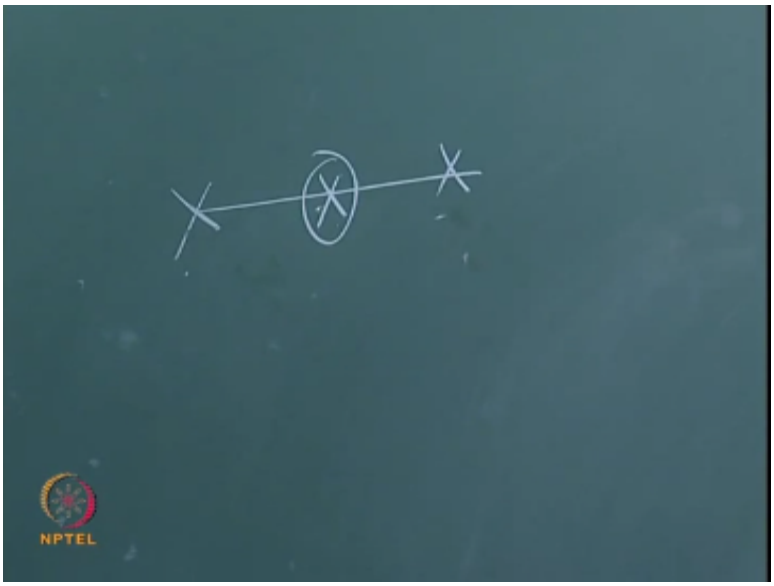
Now when we have considered all the distances with the means and then we found that x1 should be should go to the second cluster then what McQueen had done is that immediately he removed x1 from here and he had put x1 here and since one point is removed this mean is

changed immediately and one point is added here that mean is also changed immediately is it clear to you when ever a point changes its membership from one cluster to another cluster then he immediately changed the means of the corresponding clusters.

Then his termination criterion is then slightly different his termination criterion is he starts with x1 x2 up to xn and again x1 x2 up to x and then again x1 x2 up to xn and, so on cycle like this for n consecutive runs if the clusters are not changing then you stop it okay if um consecutively unconcern consecutive runs are not changing the clusters then you stop it this n consecutive runs the starting may happen with 3 X 3 X 3 X 4 X5 X 6 up to xn then x1 and x2 that is fine then you have n consecutive runs okay, so this is what he had done.

McLean's k-means and there is some other method that is called as john seesk-means and in fact there are several versions say.

(Refer Slide Time: 09:04)



The mean of the first cluster is say here at this place then using this algorithm say this cluster mean has changed it to this place using this one then what this one this method is doing is it is taking this as the new cluster mean and what this Jhansi has said is that this mean is changing from here to here then maybe I will take the new mean somewhere in the middle and I take this as the mean of the that particular question many modifications there are many many modifications of k-means algorithm.

But this algorithm assumes that number of clusters K is known please any question yeah yes it is computationally expensive it is agreed, but then if you want to get see intuitively speaking why do you need to go to the whole range of n points to change the means that is the basic point that is bothered about I do agree that if you have really large number of points then at each time you will change the mean one may not like it, I say I mean there is no doubt about that that prowl I mean people may not like it so that is fine.

But then his basic intuition is that when a point has already changed why do not you change the mean immediately why do you wait till the completion of all the endpoints and then just look at it that is basic concern but your point is quite appropriate, it is quite appropriate any other question okay, so this is assuming that the number of clusters K is known and it has a problem regarding non-convex clusters and it also has some more problems not only non-convex clusters it also has some more problems.

Any method which is based upon mean it is very much susceptible to outliers I will repeat any method which is based on mean is very much susceptible to outliers well what is the meaning of this suppose you have some points here and then just say these two points are here and all these points suppose they should belong to the same cluster, then what will happen to the mean the mean will probably be somewhere here what probably one would have wanted to do is to remove these two points and get this mean in the middle somewhere here.

If you have extreme values in your data since these extreme values have to be put in one of the clusters that cluster mean and in effect then the whole clustering process of the data would be suffering because of these extreme values because of these extreme values the clustering process may suffer these are basically outliers, so then how does one tackle this problem before I go before I start talking about tackling this problem what many persons do is somehow try to decide what outliers are remove them and do the clustering after the removal of outliers.

This is a process but there is a very basic question do you actually remove all the time all the outliers is it good for the experiment is it good for the science is it good for science, if you remove outliers always any other reply I am reminded of the code from a Nobel laureate I think he got Nobel Prize in medicine I think two three years ago so that scientist commented that probably most of the research should be done on outliers okay that was his comment something which is completely against the phenomenon.

If it is coming to you probably that contains lot of information it is not the case that you just removed and then since all other things are following your own ideologies or principles or whatever you try to do them as a result that probably is not good okay this is a basic philosophical point you may take it or you may not take it that is up to you okay you may take it or you may not take it, but what many people would like to do is that they would like to remove the outliers they would like to attach a definition to the word outliers.
Remove them and do the clustering after that so whether you like it or not that is up to you, so how people go about doing it they might go about doing it in very many ways one way is do the clustering do apply k-means algorithm okay K is the number of clusters then what you do is that after you get all the clusters for each cluster and for each variable you measure variance for each cluster and for each variable you measure the variance find out where you have the maximal variance find out the place.

Where you have the maximal variance if you have a progression if the variance value is I am just giving you the inclusion variance value is slowly and slowly you have many values of variance, if they are slowly and slowly increasing like this and this is your maximum variance that is fine this is your maximum variance value then that is fine, but if it is increasing like this and then the next one is say here you have the variances are slowly unfold including at some place there is a very big gap then you take the cluster and consider that variable for which you have got this variances.

Take the cluster consider the variable forward you have got this variance okay and that cluster you may break it into two parts that cluster you may break it into two parts according to the mean of that specific variable, that specific variable you take that specific variable whatever you that variable consider every point in the cluster, and further specific variable look at the value if

that value is less than the mean that falls into one cluster it is greater than the mean it goes into another cluster this.

You can sort of take it to be a way of getting outliers you can also have another way instead of looking at variances for each cluster you can calculate its diameter the diameter of a cluster is the maximum distance between you for every pair of points you calculate the distance and find the maximum of all these distances.

That you call it as the diameter that you call it as the diameter okay if the diameter is diameters of all the other clusters are of say one type and diameter of one particular cluster it is very large and you take that cluster again whatever process, that I mentioned you can just do it and you can remove the outliers, so this basically is saying that if the cluster is found by some looseness is there in the cluster here in this whole cluster there is some sort of a looseness the points are not very close.

They are not compact, compact I am using it in an ordinary sense since they are not compact they are loosely attached they are loosely attached then you would like to remove that cluster which has smaller number of points and keep that cluster would has larger number of points, so this meaning of loosely attached this there are a few ways in the literature you will find where these things are discussed I gave you one or two methods just now arm there is also another one where people have talked about split and merge algorithms for clustering.

And you would see the meaning in the literature split and merge initially when you do k-means and then afterwards you find this diameter and then find that cluster that has a maximal diameter and then split it okay, then first you have done the merging now then you are doing the splitting and then remove those few points and then you may want to do clustering again of this whole thing sometimes you do it sometimes you do not and basically many of the splitten much algorithms.

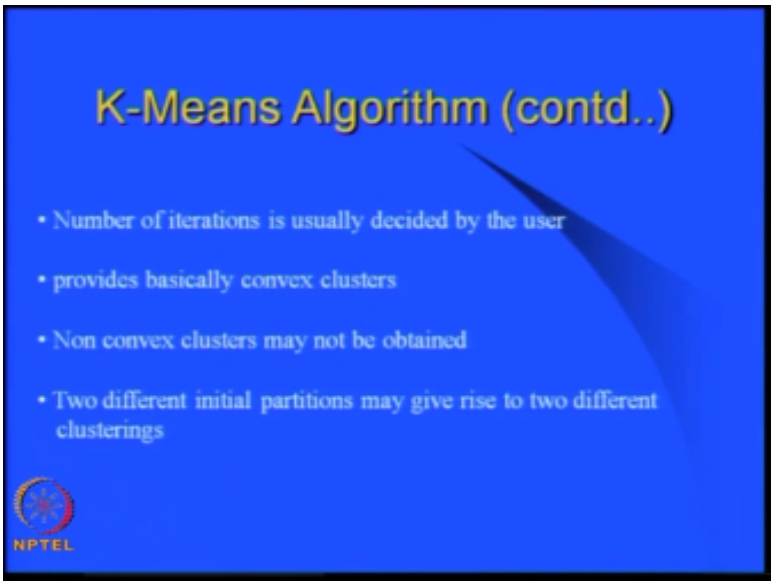
They are based on one or two basic principles you will merge them you will merge two clusters if they are somehow very close and you will split a cluster if the points in the cluster are sort of loosely attached you will split a cluster if the points are loosely attached, you will merge two clusters if somehow they are very close this is the basic principle using this principle you will

find very many algorithms in the literature where people have done both splitting and merging initially they may split and then again.

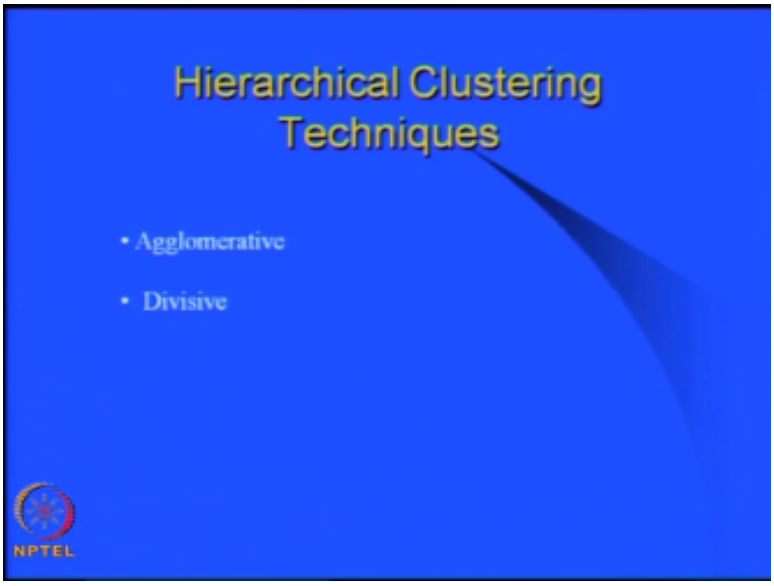
Match then split and then merge can split and then much you go on doing it till you want some conditions to be satisfied and one such method is one such algorithm is a very famous algorithm ball and Hall isolate algorithm, which they talk about it is basically a split and merge technique, so outliers is a problem the other one is whether they are loosely attached or not that is one problem ball and Hall they also try to somehow get the idea of the number of clusters, but if you implement that algorithm it is extremely complicated number one it takes simply too many too many calculations okay, it takes simply too many too many calculations first you do some sort of k-means then you remove outliers.

Then you do split and merge and again you do k-means it just goes on and on and on you increase the number of clusters, if you split then you are increasing the number of clusters if you are merging then you are decreasing the number of clusters, so when you have decreased or increase then sometimes you may need to again do the clustering and it just goes on and down and till there is a termination criterion, so and that has you need to do too many calculations you need to do too many calculations now let me talk about the non convex clusters.

(Refer Slide Time: 23:30)

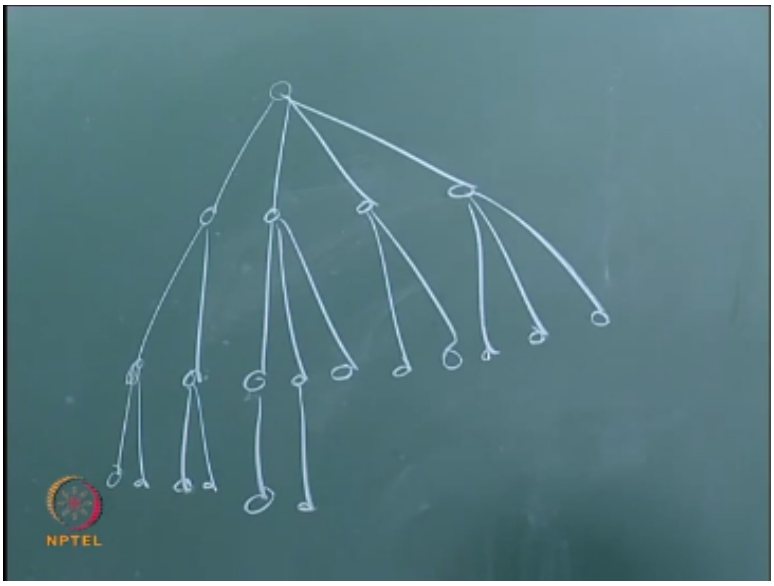


(Refer Slide Time: 23:36)



This hierarchical clustering the k-means sort of algorithms they are all nonhierarchical here are key I suppose you know the meaning of the word here are K do you know the meaning of the word here are k there is.

(Refer Slide Time: 24:01)

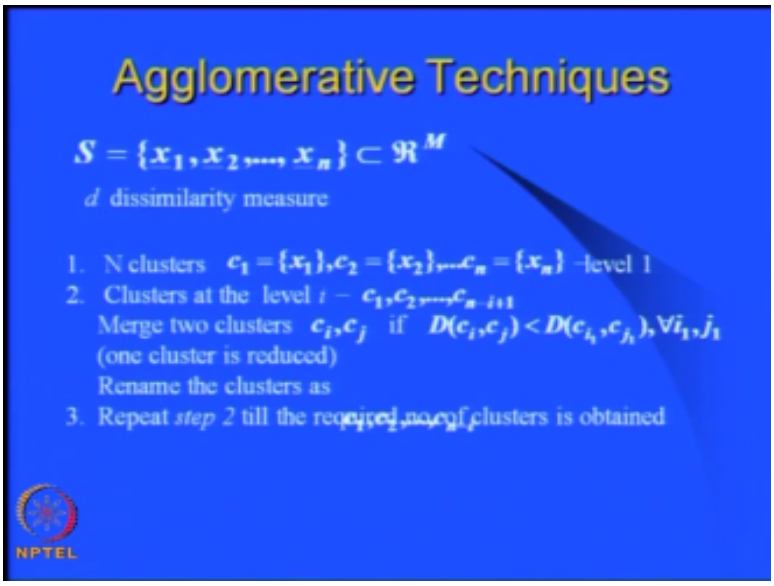


It is basically a tree structure and you might be having something like this you might be having like this there is basically some sort of a tree structure you might be having something like this there is some sort of a tree structure maybe you can have this, I will give you examples of hierarchical clustering techniques, but let me discuss this there are two types of hierarchical clustering techniques one type is known as agglomerative another one is known as division in agglomerative.

What you are going to do is that if you have n number of points you are going to assume that you have n number of clusters and you will go on merging them in the first one you will have n - 1 clusters in the next one you will have n – 2 clusters in the next one you will have n - 3 clusters you will go on and on doing it and in divisible you assume that you have a single cluster then you break, it into two parts then in the next iteration you choose one of the existing clusters and that you break it into two parts.

So basically again it is a tree structure so both of them are tree structures agglomerative and division.

(Refer Slide Time: 25:50)

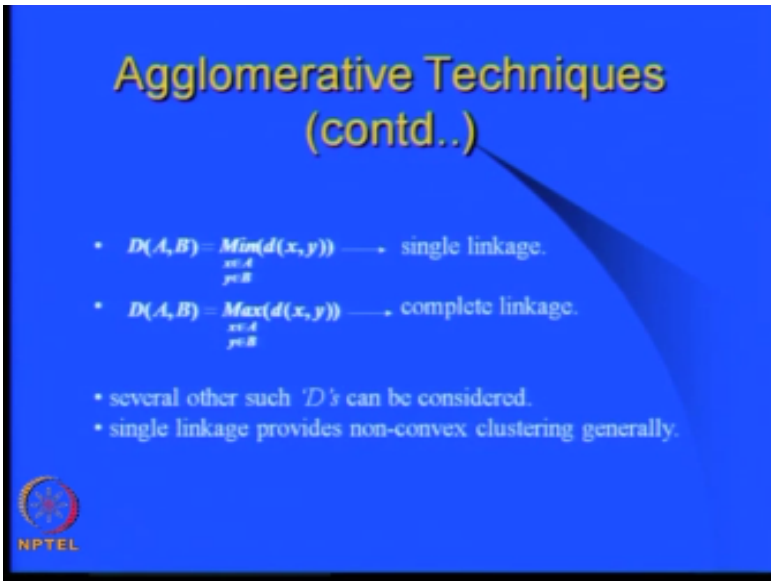


So agglomerative techniques so this is small and this is small and this is not capital n this is not capital n is small n, so if you have smaller number of points you will assume that you have smaller number of clusters C1 C2 Cn, so in the level 1 you have n clusters so in the level I you are going to have n – I + 1 clusters then you will merge two clusters at level I if capital D from the distance between the cluster Ci and cluster Cj is less than the distance between clusters Ci 1 and Cj 1for all I 1 is even.

So in that way you will reduce 1 cluster then you again you will rename and you are going to do this step till either you have, if you have the number of clusters to be obtained you will do that otherwise you will go up to the number of clusters 1 and look at all these things and somehow you decide on the number of clusters looking at all this and I mean all the clusters that you have obtained at each place, but let us assume that you are given the number of clusters and you have got these clusters.

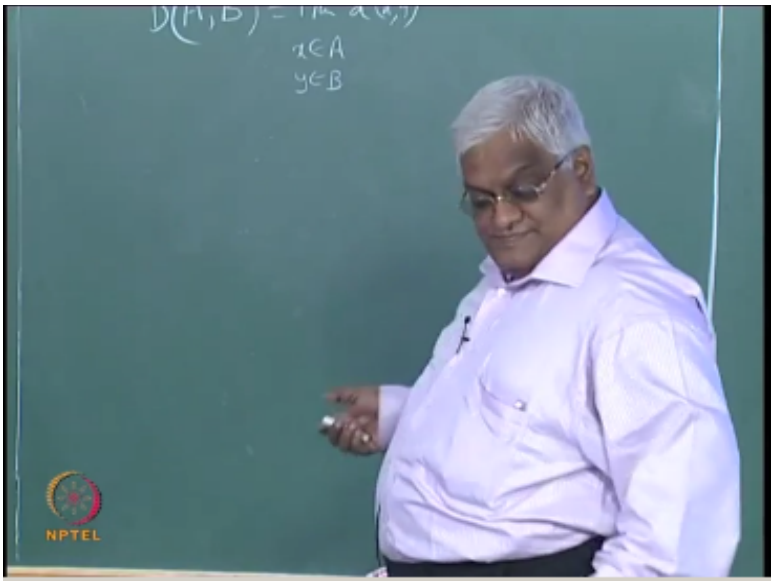
So you will go on repeating this one till you get the required number of clusters, now the question is how to define that capital D.

(Refer Slide Time: 27:29)



How to define capital D.

(Refer Slide Time: 27:46)



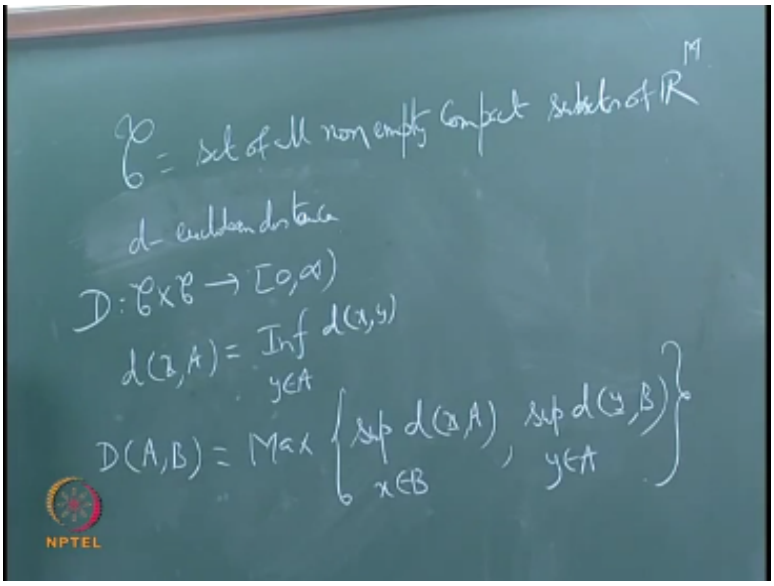
This is a definition that one can follow that is take a point from here and take a point and find the distance between them and that you do it for every point and every point from here and find that pair for which the distance is minimum actually that pair for which the distance is minimum that you call it as capital D note that till, now we have calculated distance between points this distance between two sets this is distance between two sets this is not a metric this is not a metric if you use this definition the clustering that you will get is what is known as single linkage.

And if you use this definition the clustering that you get that is what is known as complete linkage and unfortunately this is also not a metric none of these two DS that I mentioned they are matrix I will tell you why they are not matrix assume that there is a point common to A and B then this distance will be 0, so distance is 0 means the sets the points have to be same you look at the definition of metric so only one point is common the sets are not same, so sets are different but the distance between them is 0.

So that is not a metric now this one where you are taking the maximum of x belonging P and Y belonging to B that is also not a metric because take B as A then distance between a with itself that must be equal to 0, but here you are going to get a positive quantity here you are going to get a positive quantity, so this is not a metric and that is not a metric then the next question is can we actually define a metric the answer is yes you can define a metric there is a metric that is known as how stops metric.

There is a metric that is known as house torque metric that is defined between sets which are compact non empty compact subsets this word here compact is again topology it is coming from topology how staff metric that, if you use that definition then that is going to be a metric you want me to give the definition of how store metric.

(Refer Slide Time: 31:41)



Okay let me just give you suppose we are in say M dimensional space now small D this is the usual Euclidean distance small D is the usual Euclidean distance and when you define this must come from this to zero to ∞ okay, we will define it like this distance between a point x and they set A is Inf of Y belonging to a D X Y first, let us define this distance between x and they set a now then the distance between A and B this you can have it as maximum sup means supremum and if we are dealing with finite sets supremum is same as maximum infimum is same as minimum.

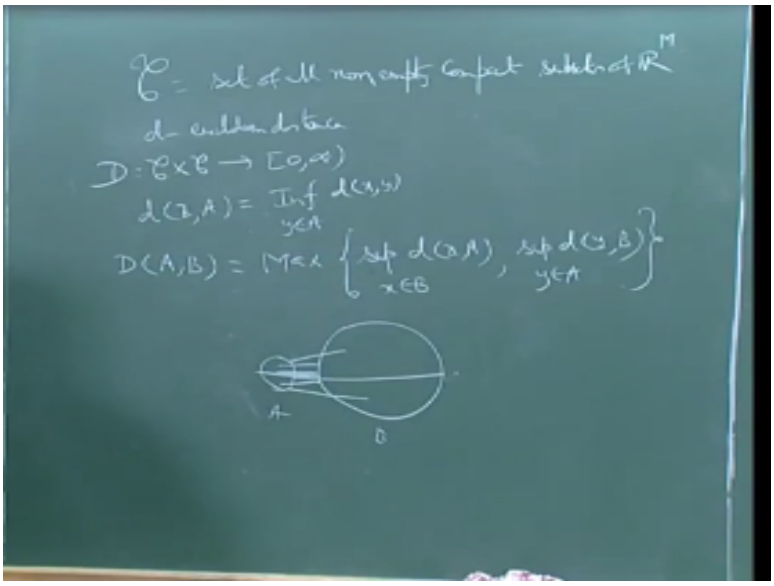
Infimum is same as minimum supremum is same as maximum this is this basically gives you the following suppose this is one set and then say this is another set let me call it A and let me call it B what you do is that, let us look at this take a point Y in a and first you need to find dy B that means further from this point you consider all the distances and find where it is minimum infimum that that distance is this and that you have to do it for every y may be for this one again the distance is this may be for this one again the distance is this.

And the maximum is actually this something like this okay now take this here also for a point here you find all the distance of this point to all the distances here and the minimum the minimum is probably occurring here maybe for this point the minimum is occurring here and then the maximum that probably may occur somewhere here so you have this quantity this quantity then you need to find the maximum of these two that maximum is in this case this to this one that you are going to get.

So this is how store metric by changing the definitions of D you are going to get many different clustering's by changing the definitions of D you can simply get many different clustering's that is single linkage complete linkage you might have something called average linkage the word average linkage you can define it in many ways the word average, now probably you might be having a question the question is minimum X belonging to ay belonging to B this is a dissimilarity between a and B maximum X belonging to ay belonging to B this is also a dissimilarity.

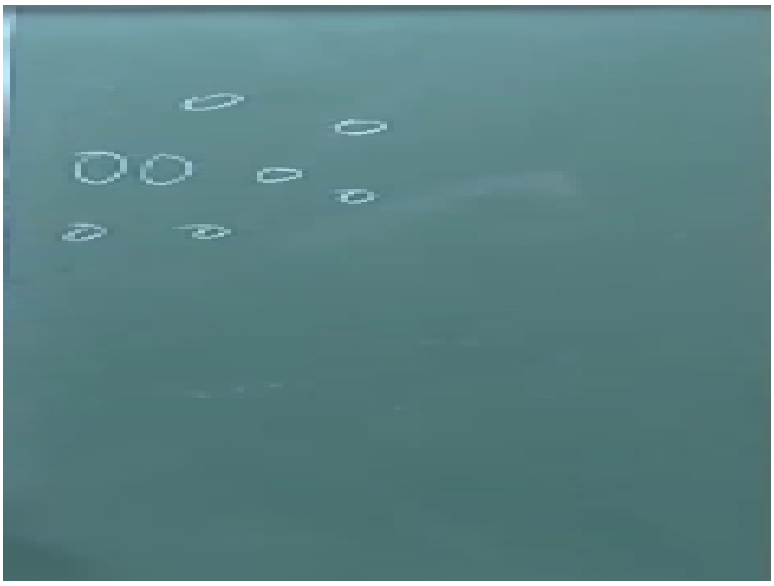
If you consider this dissimilarity as correct do you think this dissimilarity will be correct if you take maximum as dissimilarity then can you take minimum also that is similarity in another one or if you consider this do you think you should consider this, are you understanding my question probably only one of them is one should I mean probably at the same time one should not consider both do you agree to this, but each of them has its own meaning that let me tell you have one cluster here.

(Refer Slide Time: 37:03)



You have one cluster here and you also have some clusters.

(Refer Slide Time: 37:19)



But the algorithm says is that you find let us let us look at the first definition single linkage you find that the similarities between every pair of clusters the one for which the dissimilarity is minimum you will merge those two clusters, now between two clusters here the dissimilarity is measured as minimum of these things that is fine minimum dissimilarity our maximum similarity that maximum similarity wherever it is maximum you are joining those two clusters this is minimum dissimilarity our maximum.

Similarity that wherever for whichever pair that is maximum you are joining those two clusters you are merging those two clusters that is single linkage, now what is complete linkage complete linkage says between two clusters what can be the maximum amount of dissimilarity the maximum amount of dissimilarity between these two clusters is this and that you are minimizing it one is minimizing the maximum disability the other one Is maximizing the maximum similarity both of them are valid only thing is that when you are maximizing.

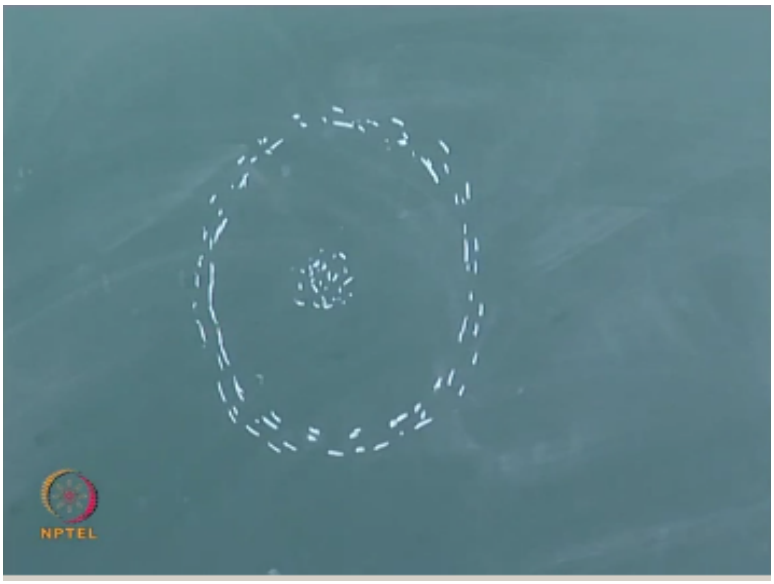
The maximum similarity you have a very positive outlook and you are minimizing the maximum dissimilarity you have sort of a negative outlook it is like saying that people have invented Aeroplanes, but there is another scientist who invented parachutes also both of them are necessary are you understanding what I am trying to say one is a pessimistic way of looking at it another one is optimistic way of looking at it, and you do need both points of use is it understandable.

So one is this and the other one is this it is basically Prince Wei where we have to find that the similarities and then do the joining if you look at the best way of looking at if you are constructing MST you would find Chris Cole's algorithm where it assumes that the edge weights are given to you, but if you need to find the edge weights then Prims algorithm order n square which is what you are asking me this is an order n square algorithm this is what generally people use since you need to know the edge weights between every pair and you need to know edge weights means it has to be an n square algorithm you cannot have anything less than that but you need to consider every pair number of pairs is n into n - 1 /2.

So that is the N square okay and that is n square so you are asking me about for every pair we need to look at the dissimilarity yes then you are going to get this that is also true, if you need if you want to apply single linkage algorithm to satellite images where the number of pixels may be let us just say 512 by 512 then and between every 2 pixels somehow you calculate some features and then you try to get these things then your algorithm your method will collapse, because it is extremely expensive 512 square pixels you have.

It is extremely expensive though MST has very many nice properties this is a property that computationally it is really expensive that is one of the reasons why people do not go in for MST but, now let me tell you some of its good properties I claim that you will get I claim that you will get non-convex clusters how do you get non-convex clusters.

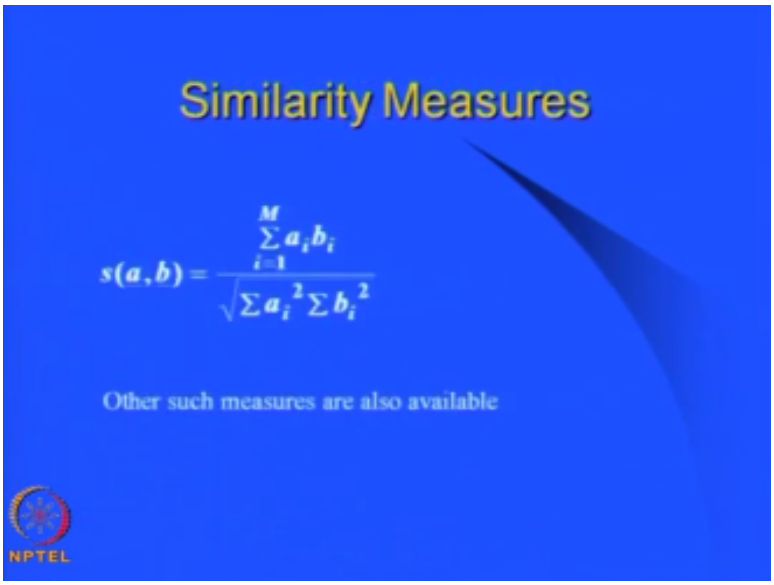
(Refer Slide Time: 42:03)



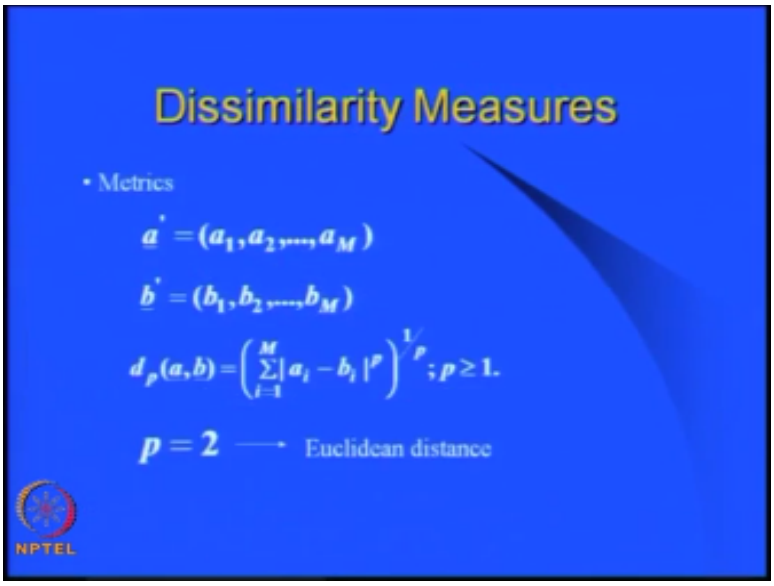
Okay right, so this is the data that is given to you and you have two clusters here one is this another one is this now suppose you know that the number of clusters is 2, now you would like to do single linkage on this that means first you will find 2 points where the distance is minimum okay and like that you apply that whole algorithm, so and you will go on doing it till you get two clusters till you get two clusters and then there you will stop it what will happen is that all the points here they will be connected.

All the points here they will be connected and that will be the case when you will come to two clusters from two to one you are going to have you take a point from here and probably that means something like this and this is a huge value compared to all those small values so you are going to get non-convex clusters and when I started clustering I started with this example.

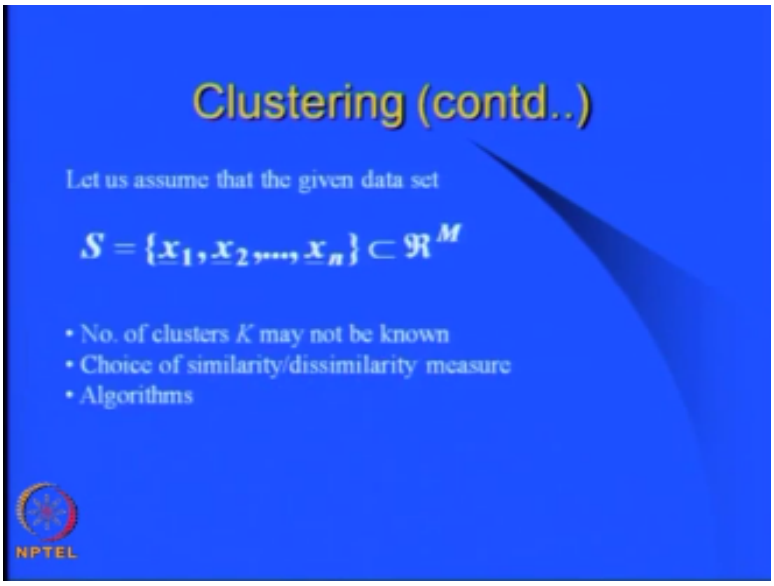
(Refer Slide Time: 43:41)



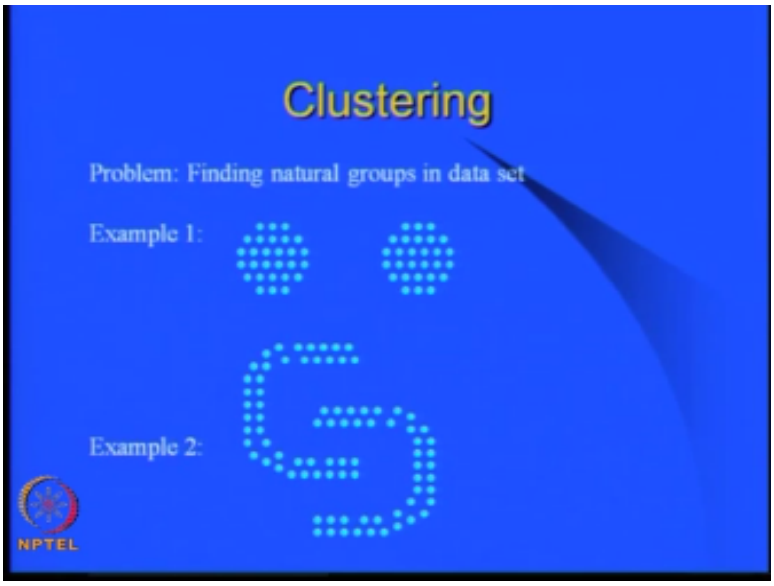
(Refer Slide Time: 43:41)



(Refer Slide Time: 43:42)



(Refer Slide Time: 43:43)

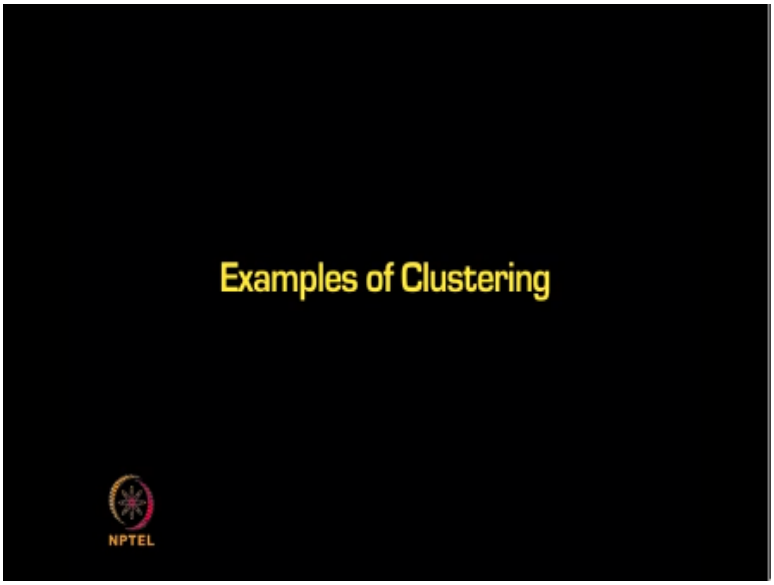


Look at the second one if you apply single linkage on this you will get those two clusters if you apply single linkage on this you will get those two clusters, so you can get non-convex clusters right but it is really expensive that is one problem there is another one came in sort of algorithms they basically look at the it is centroid based algorithm so you are going to look at something like a density that means you have got a certain mean and then there is some radius something like that you are going to have.

Basically you would like get cause you will get convex clusters there single linkage you will get non convex clusters can you is there any way of putting these two things together, so that to develop to have an algorithm which has the plus points of both this single linkage and as well as k-means, so that somehow you are able to get both the convex and non convex clusters that is one secondly I was mentioning in one of the earlier classes that there is also something called a came adults algorithm which is based on the median.

You would need something which is based on the median the reason is that medians are generally not affected by outliers many of you have image processing background you would have done at some point of time median filtering okay, if you do mean filtering and if the window contains really high values and low values then probably you would not like to do I mean filtering probably you would like to do median filtering to remove the outliers, so similarly here you have algorithms which are based on generalization of median which is known as made you have some monoid based algorithms also k many voids algorithm for clustering you are you have that.

(Refer Slide Time: 46:17)

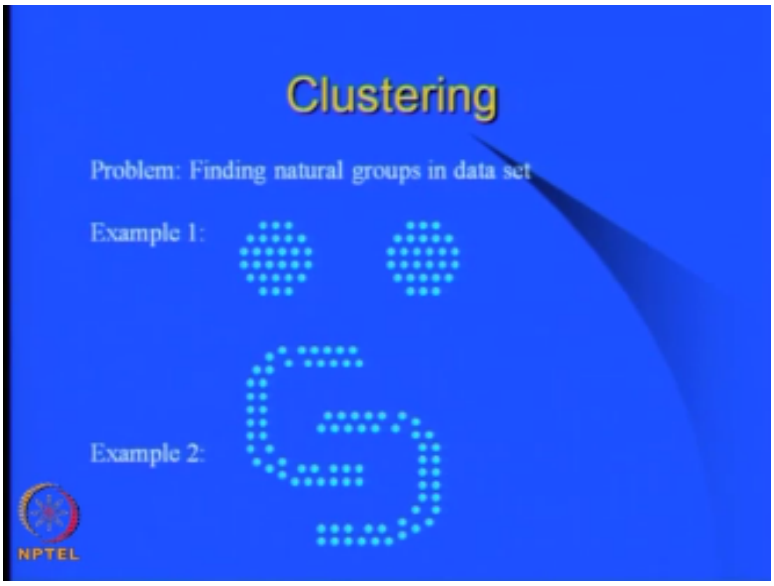


(Refer Slide Time: 46:22)

Yeah look at the example - in the slide I have drawn the points in this way for this cluster if you calculate the mean the mean will be somewhere here, it will be outside the cluster for this one if you calculate the mean the mean will be somewhere here okay, so this is the actual mean of this cluster this is the actual mean of this cluster, now if we had actually got these two clusters then what the k-means algorithm would have done if it would have calculated this as the mean and this as the mean then all the points which are falling on this side.

It would have made it into one cluster all the points which are falling on this side it would have made into another question, so we are not getting the clusters okay whereas in the first example.

(Refer Slide Time: 47:24)



k-means algorithm would generally provide the clusters that you would like to get in the first one in the second one the MST based clustering that is a single linkage would have given you the correct results if it had if we know if you feed the information that the number of clusters is 2 and if you measure the disability by Euclidian distance then you would get those two required clusters.

End of Module 03 – Lecture 02

Online Video Editing / Post Production
M. Karthikeyan
M. V. Ramachandran
P. Baskar

Camera
G. Ramesh
K. Athaullah
K. R. Mahendrababu
K. Vidhya
S. Pradeepa
D. Sabapathi
Soju Francis
S. Subash
Selvam
Sridharan

Studio Assistants
Linuselvan
Krishnakumar
A. Saravanan

Additional Post – Production
Kannan Krishnamurty & Team

Animations
Dvijavanthi

NPTEL Web & Faculty Assistance Team
Allen Jacob Dinesh
Ashok Kumar
Banu. P
Deepa Venkatraman
Dinesh Babu. K.M
Karthick. B
Karthikeyan. A
Lavanya. K
Manikandan. A
Manikandasivam. G
Nandakumar. L
Prasanna Kumar. G
Pradeep Valan. G
Rekha. C
Salomi. J
Santosh Kumar Singh. P
Saravanakumar. P
Saravanakumar. R
Satishkumar. G
Senthilmurugan. K
Shobana. S
Sivakumar. S
Soundhar Raja Pandian. R
Suman Dominic. J
Udayakumar. C
Vijaya. K.R
Vijayalakshmi
Vinolin Antony Joans

Administrative Assistant
K.S. Janakiraman

Principal Project Officer
Usha Nagarajan

Video Producers
K.R. Ravindranath
Kannan Krishnamurty