

**Indian Institute of Technology Madras  
Presents  
NPTEL  
National Programme on Technology Enhanced Learning**

**Pattern Recognition**

**Module 03**

**Lecture 01**

**Basics of Clustering,  
Similarity/Dissimilarity Measures,  
Clustering Criteria.**

**Prof. C. A. Murthy  
Machine Intelligent Unit,  
India Statistical Institute, Kolkata**

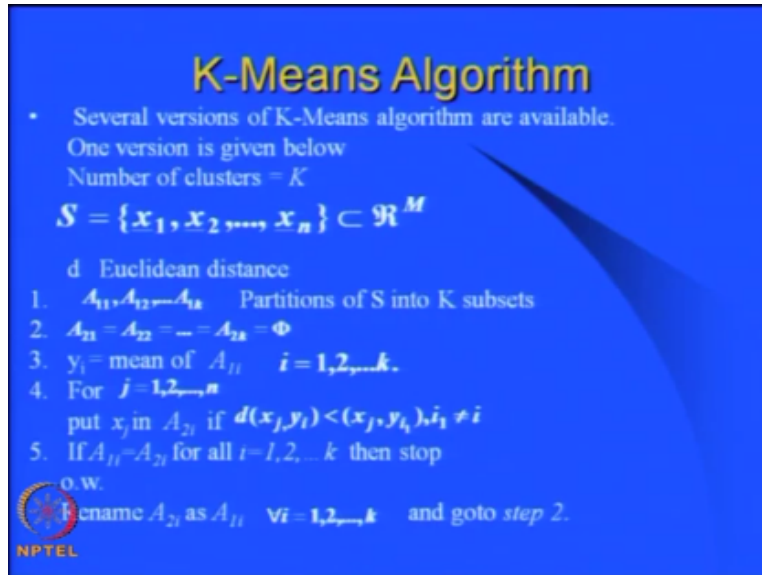
Basically in many of these clustering problems we would like to define an, optimization criterion we like to define an, optimization criterion and whichever clustering follows that optimization criterion that we call it as clustering of the data this optimization the definition of this optimization criterion that again that is that should be in principle problem dependent since we are starting the subject so we will start with basically if use I mean at least one such optimization function we start with basically one such optimization function and then let us see how it proceeds why do you need to define an, optimization function for doing clustering.

The reason is that the clustering problem as I said it is whatever someone wants you to find these properties you need to have so you find them you do the clustering according to those properties sometimes what happens is that you are just given a data set you are asked to do the clustering the person may not have anything on his or her mind you are just given a data set then what you may want to do is that find let us just see what sort of groupings are there if it is a simple example like playing cards one can easily find the groupings.

But usually the data sets that are given to us they do not follow such simple I should say properties they do not have those simple properties so what one may like to do is that somehow within a cluster you need to have I mean the points should be similar to each other and for between two different clusters you need to have some sort of a dissimilarity and you need to have


I should say more dissimilarity so what is the way of formulating this thing mathematically a way of formulating this mathematically is the following.

(Refer Slide Time: 02:36)

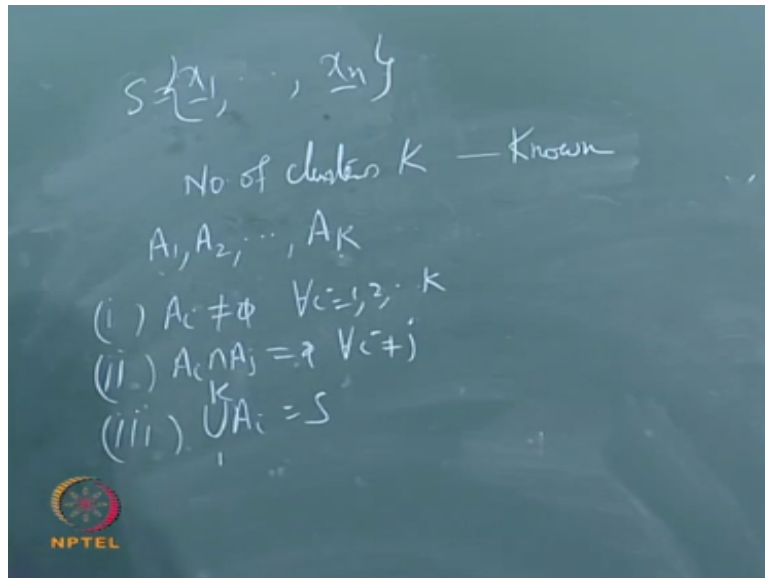
A blue slide with yellow text titled "K-Means Algorithm". It contains a list of steps and mathematical notations. The NPTEL logo is in the bottom left corner.

### K-Means Algorithm

- Several versions of K-Means algorithm are available.  
One version is given below  
Number of clusters =  $K$   
 $S = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^M$   
d Euclidean distance
  1.  $A_{11}, A_{12}, \dots, A_{1k}$  Partitions of  $S$  into  $K$  subsets
  2.  $A_{21} = A_{22} = \dots = A_{2k} = \Phi$
  3.  $y_i = \text{mean of } A_{1i} \quad i = 1, 2, \dots, k.$
  4. For  $j = 1, 2, \dots, n$   
put  $x_j$  in  $A_{2i}$  if  $d(x_j, y_i) < d(x_j, y_{i_1}), i_1 \neq i$
  5. If  $A_{1i} = A_{2i}$  for all  $i = 1, 2, \dots, k$  then stop  
o.w.  
Rename  $A_{2i}$  as  $A_{1i} \quad \forall i = 1, 2, \dots, k$  and goto step 2.

 NPTEL

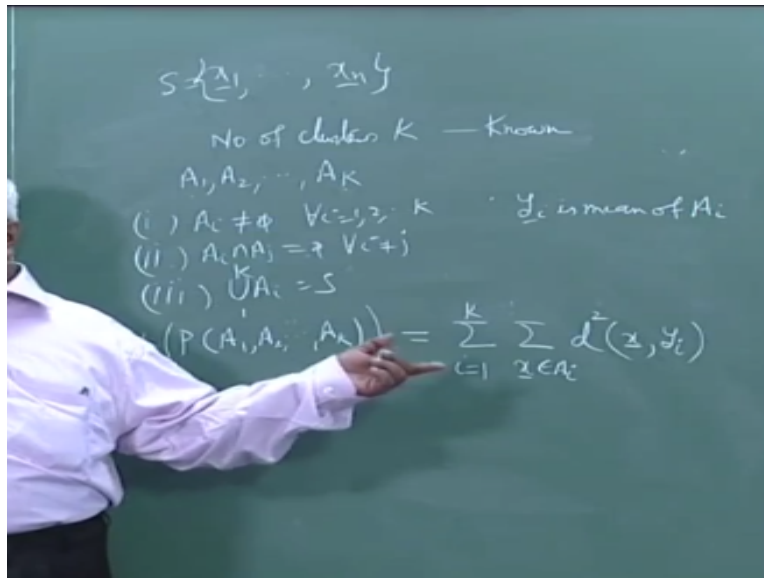
(Refer Slide Time: 02:46)



So you are given a data set okay and then the number of clusters okay let us say it is known the number of clusters  $K$  is known so basically what do you need to do you need to divide this data set into  $K$  clusters that means you need to make a partition am I right you need to make a partition so let us call one such partition to be say  $A_1, A_2, A_k$  this is a partition of this dataset so if I write this data set as  $S$  and these are  $A_1, A_2, A_k$ .

So what are the properties the first property is  $A_i \neq \emptyset$  for all  $i$  to  $A_i$  intersection  $A_j$  is  $\emptyset$  for all  $i \neq j$  and  $\bigcup_{i=1}^K A_i = S$ . So the first thing is you need to make a partition well this is a partition then what do you need to do somehow you need to have some sort of a similarity or dissimilarity let us assume that we are doing we are finding this dissimilarity by using Euclidean distance let us assume that we are finding this dissimilarity by using Euclidean distance.

(Refer Slide Time: 04:42)



Then for a partition  $A_1, A_2, \dots, A_K$  of the whole set  $S$  let us define this function  $L$ . What is this function? This function is let us take square here this  $y_i$  let us call  $y_i$  as  $y_i$  is mean of  $A_i$ .  $y_i$  is mean of  $A_i$  you find the mean of all the points in this set that is  $y_i$  so what you do is that you take a point  $X$  in  $A_i$  find the distance between  $X$  and the mean okay.

And this you take the square of the distance and this you sum it up over all  $X$  in  $A_i$  and you take  $i = 1$  to  $K$  this is the function that you are defining now you would like to get a partition  $A_1, A_2, \dots, A_K$  which minimizes  $L$  you would like to get a partition  $A_1, A_2, \dots, A_K$  which minimizes  $L$  right intuitively it looks nice okay you would like to get a partition  $A_1, A_2, \dots, A_K$  which minimizes  $L$  right.

Now okay let me start asking your questions let me not show this thing now let me just keep it like this you are given  $n$  points let us assume that the number of clusters is 2 that is not going for  $K$  how many such partitions can you have how many such partitions can you have how many such pairs you can have any one will answer  $A_1, A_2$  this  $A_1, A_2$  okay let us just say small  $n = 4$  you have 4 points you can put one point in one cluster 3 points in another cluster okay you can put 2 here 2 here for each one of the things you can calculate this right.

And you would like to find that partition which minimizes this now my question to you is how many such partition means possible for  $K = 2$  for  $K = 2$  how many such are possible that is right  $2^{n/2}$  there is a  $2^{n/2}$  because we are not considering all the points put in 1 cluster 0 in another cluster that we are not considering so  $2^{n/2}$  we need to have 1 and  $n - 1$ ,  $n - 1$  1 they are same so

you need to divide it by 2 so  $2^{n/2}$  that is for  $K = 2$  what about  $K = 3$   $K = 4$  and my next question is any value of  $K$ .

What is your answer you try to give me the answer tomorrow otherwise I will tell you the answer tomorrow okay for any  $K$  first you need to know how many such partitioning are possible then you know the complexity of the problem it is actually after order of  $K^n$  – something is and - something + something will coming but I want you to give me the exact expression I want you to give me the exact expression okay.

So that is about the computational complexity part of it the second part is fine you have got these clusters somehow you have done the whole exhaustive search or whatever it is one has done it and you have got the clusters what properties these clusters process what properties these clusters possess so this is the function that we want to optimize we would like to get hold of those  $A_1, A_2, A_K$  which provide minimum value for this function.

Now let us just see the properties of this function properties means if we really get those sets  $A_1, A_2, A_K$  that means if we really get those  $K$  clusters then what properties do those  $K$  clusters poses.

(Refer Slide Time: 11:17)

## Similarity Measures

$$s(a, b) = \frac{\sum_{i=1}^M a_i b_i}{\sqrt{\sum a_i^2 \sum b_i^2}}$$

Other such measures are also available



Before I do that.

(Refer Slide Time: 11:21)

$$A_1, A_2, \dots, A_K$$


(i)  $A_i \neq \emptyset \quad \forall i=1, 2, \dots, K$       $y_i$  is mean of  $A_i$

(ii)  $A_i \cap A_j = \emptyset \quad \forall i \neq j$

(iii)  $\bigcup_{i=1}^K A_i = S$

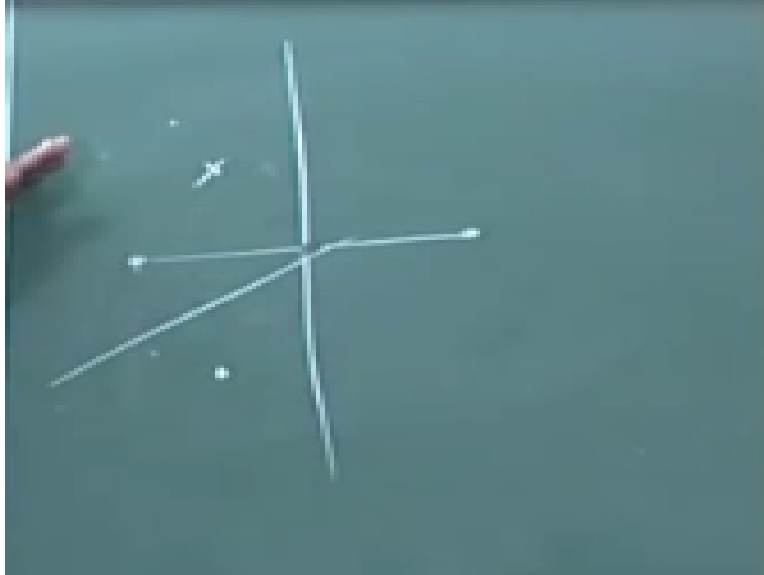
$$L(P(A_1, A_2, \dots, A_K)) = \sum_{i=1}^K \sum_{x \in A_i} d^2(x, y_i)$$

To find  $(A_1^0, A_2^0, \dots, A_K^0)$  such that

$$L(P(A_1^0, A_2^0, \dots, A_K^0)) \leq L(P(A_1, A_2, \dots, A_K)) \quad \forall P(A_1, A_2, \dots, A_K)$$


I will just write hereto find again  $A_1^0, A_2^0, \dots, A_K^0$  is for optimal such that  $L(P(A_1^0, A_2^0, \dots, A_K^0)) \leq L(P(A_1, A_2, \dots, A_K))$  for all  $P(A_1, A_2, \dots, A_K)$  we are supposed to find  $A_1^0, A_2^0, \dots, A_K^0$  is for optimal so  $A_1^0, A_2^0, \dots, A_K^0$  such that the partition  $A_1^0, A_2^0, \dots, A_K^0$  for last the last function corresponding to this partition is less than or equal to the last function corresponding to any other partition  $A_1, A_2, \dots, A_K$  for all partition  $P(A_1, A_2, \dots, A_K)$  of the set  $S$  so this is our a now if we do really find this partition  $A_1^0, A_2^0, \dots, A_K^0$  what properties does this partition possess let us see suppose note that we are using only Euclidean distance we are not using any other distance this is really important and you will actually probably understand it why I am saying it.

(Refer Slide Time: 13:03)



Suppose this is the mean of one cluster after we have got obtain the optimal partition and this is the mean of another cluster okay this is the mean of one cluster this is the mean of another question after the I mean of the optimal cluster one of the optimal clusters this is the mean of the other up this is the mean of one of the other optimal clusters now you have got a point here this is the point from the data set you are supposed to put this point to either this cluster or to this one now to which particular cluster you would like to put it.

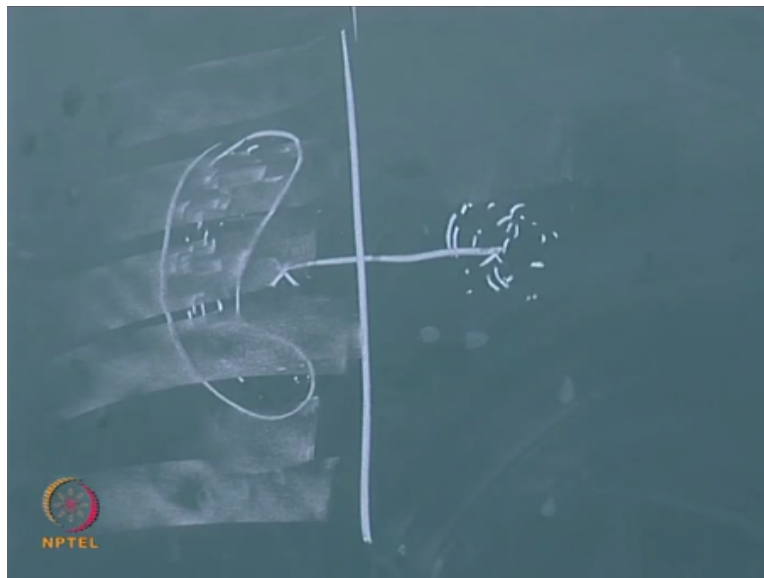
You will put it to this cluster because the distance of this point with this cluster is less than the distance of this point to this cluster mean and you are always taking the cluster mean in our expression for optimality we are always looking at the distance between  $X$  and the mean right the mean of the cluster right so that means what that means basically all the points which are on this side of the line they will go to this cluster and which are on this side of the line they will go to this cluster.

That means what basically say suppose this is another mean so that means you will have a line like this for this right and suppose there is no other than mean you will have one more line like this basically okay do you know the meaning of a convex set when do you call a set to be convex the line segment joining any two points should be completely contained in the set then you call that set to be convex now this is one half space this is one line this is one half space is a half space convex it is convex take any two points here the line segment is here only within that set only.



And then if there is one another point you are going to have on another line and this side there is another half space so it is the intersection of two half spaces intersection of two convex sets is it convex intersection of two convex sets is convex intersection of finitely many convex sets is convex so basically the cluster corresponding to that particular mean it is if you forget about the finite number of points basically it is the intersection of all these half spaces right and that is basically convex.

(Refer Slide Time: 16:55)



It may so happen that it may so happen that the shape of the cluster may be like this okay but take the mean here the mean will be somewhere here take the mean here this is convex basically this method provides convex shaped clusters have you understood if you use this optimization criterion the clusters that you would get they are basically convex shaped what is the meaning of convex shaped note that the shape of this is not actually convex this is actually a non convex shape but then once you do this one this is convex this side of this line this is a convex set this side of this line is a convex set convex set.

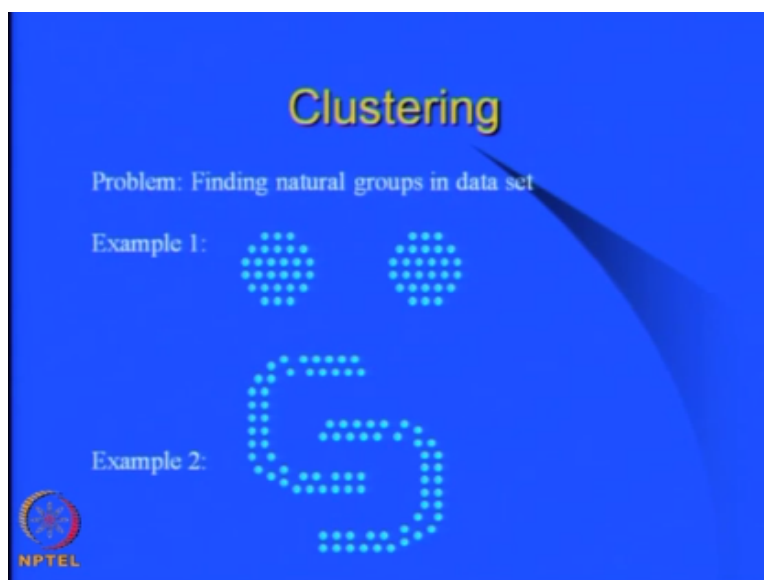
When I say that you get convex shaped clusters what I mean is that the finitely many points what shape does do these points have I am not bothered about that the shape of these points I am not bothered what I am bothered is that it is basically intersection of convex sets each convex set by the very nature of the definition a convex set is an uncountable set it is surely an infinite set and

since we are looking at points in  $\mathbb{R}^2$  and  $\mathbb{R}^3$ ,  $\mathbb{R}^4$ ,  $\mathbb{R}^M$  it is an uncountable set and intersection of convex sets if it is not a singleton it is an uncountable set.

If it has two points means then it is going to here all the points in between those two points then it is an uncountable set so intersection of convex sets is uncountable we are doing clustering of finitely many points we are not doing questioning of uncountable many points we are doing clustering of finitely many points okay so these finitely many points may form some sort of shapes but once you do the clustering the space corresponding to one cluster that is basically convex.

The space corresponding to are the set corresponding to one cluster that is basically convex because in this example this whole thing corresponds to this cluster and this is a convex set though the shape of this is not convex so any I mean this criterion essentially provides you convex clusters this criterion essentially provides you convex clusters and once I say this thing and you are going to get all the other complications the complications are that.

(Refer Slide Time: 20:34)

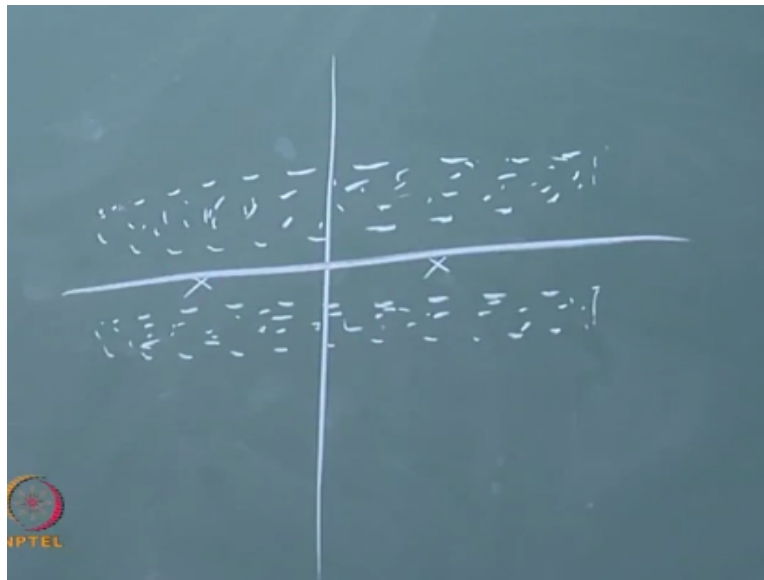


The slide has a blue background. At the top center, the word "Clustering" is written in a yellow, sans-serif font. Below it, the text "Problem: Finding natural groups in data set" is written in white. Underneath that, "Example 1:" is followed by two clusters of blue dots arranged in a circular pattern. Below that, "Example 2:" is followed by a cluster of blue dots arranged in the shape of the number "5". In the bottom left corner, there is a small circular logo with a red and blue design, and the text "NPTEL" below it.

Look at the example two here if you want to get those sort of clustering's then you will not be able to get using this optimization criteria see the power of mathematics is even without doing a single thing we are able to state that you will not be able to get it okay. You cannot be able to get this shape here in example two by optimizing this criterion okay so in fact I will make slides of all these things and I will give the slides if not in the coming one or two days the next system surely I will give the slides.

Yeah so when you have non convex shaped clusters it is not necessarily true that you will get those clusters by using this criteria then you have a next question the question is suppose the clusters are convex shaped do we get those clusters by using this criterion the answer is unfortunately know the answer is unfortunately no.

(Refer Slide Time: 22:16)



Suppose okay take two such elongated strips really long finitely many points only each one is this is convex shaped this is also convex shaped but if you optimize this criterion you may get it like this one may be giving you lesser value for this than this partition if you make that strip to be

very long if you make the strip to be very long then this one may give you a smaller value than this.

So it is not necessarily true that the convex clusters existing in the data set they are always obtained by this criterion it is giving will give you convex clusters but the convex cluster's that it provides they may not be the convex clusters that you are desiring they may not be the convex clusters that you are desiring you are desiring one type of convex clusters it is giving you other type of convex clusters okay.

So the whole thing that I want to state is that whenever anyone gives you any criterion for doing any work be it clustering or anything first you try to analyze its properties you try to analyze the properties you will get many clues from there and you see this is on every standard criterion many persons use this criterion a generalization of this criterion is used in fuzzy C means for this C means algorithm is based on a generalization of this criterion if you look at the whole of this thing optimization function there it is basically this one only slight generalization but then since this has all these drawbacks there is a slight generalization it will also possess at least some of these drawbacks if not all okay so this is a my general advice to you all anyone giving you any criterion you please look at the + points and - points just look at it I mean the mathematics are the things that I have done it is nothing I mean great I am just trying to explain what this is doing okay.

So it is not necessarily true that the Condor cluster is the existing in the data set you may be able to get by this criterion and the non convex clusters you may not get them so that is one part of it now the second part is I was asking you about how many such partitions are there it is a huge number it is a huge number this problem is similar to from an endpoint set to a K point search the number of onto mappings the number of onto mappings from an endpoint set to K point search.

How many onto mappings are possible it is exactly the same problem as this one it is a small difference if you are talking only about on two mappings then you will not get that divided by 2 factorial  $2^{n-2}/2$  factorial okay partition A1, A2 is taken to be same as partition A2 even that is how the 2 factorial is coming there so for K it will be K factorial so number of onto mappings from an array n element set to K element set divided by K factorial will give you the answer to this will give you the answer to this one.

And you will see that that number is huge I hope all of you know about travelling salesman problem and NP-hard problems right I hope all of you know about this thing know about it if you have K number of clusters and if it is of the order of  $K^n$  then you are going to be in serious trouble right you are going to be in serious trouble so whatever algorithms that are existing for optimizing this they are basically sub optimal algorithms that means you are not in a position to optimize the criterion there using those algorithms.

So if you can suggest an algorithm which really optimizes this without doing the exhaustive search then it will be a very nice contribute onto the literature and people will lap it up let me state this thing to you once again if you can develop an algorithm for optimizing this for getting this partition without making an exhaustive search for any data set and for any value of K then people will surely lap it up people will surely lap it up because we do not have algorithms for it.

So k-means algorithm it tries to do the optimization but it does not guarantee to provide the optimal solution k-means algorithm there are several versions available in the literature there is one version called 4G scheming there is one version called Macklin scheming there is one version called Jonze's k-means and there are modifications and generalizations of this in fuzzy c means there is instead of mean if you use median then you would have known something known as K merits algorithm okay.

And you have some things you have sometimes a data set where some for some points you know the classification there are some other points you do not know the classification and you would like to do some sort of a it is clustering that is semi-supervised which is nowadays many people are doing so you have a k-means version of this even in that case like that you have just several K means algorithms are there is one leader algorithm there is one leader algorithm professor M.N. Murthy more team IASC Bangalore.

He is one of the authors of the it is a very nice algorithm it tries to do things very fast well it does not guarantee optimality I mean it does not guarantee optimality I mean even the k-means sort of thing it does not guarantee but it is very fast and it gives reasonable results leader algorithm that is a very good algorithm M.N. Murthy IASC Bangalore so like that you see you will find several versions of this depending on the type of use depending on the constraints that you have you have several versions of this.

So I will do one of the versions here in the class the version that I am going to do it is a 4G scheming which was in 1965 McQueen's k-means algorithm it came in 1967 I am doing 4 G's k-means because many are many pattern recognition books you will find basically for just k-means algorithm but as a person I would prefer to use McQueen's k-means instead of 4G k-means if you want I will give you both the algorithms I would prefer to use McQueen's k-means instead of 4 G's K-means but I am giving you 4 G's k-means because many pattern recognition books have this algorithm shall we stop it.

**End of  
Module 03 – Lecture 01**

**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 Krishnamurthy & 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

**IIT Madras Production**

Funded By  
Department of Higher Education  
Ministry of Human Resource Development  
Government of India

[www.nptel.ac.in](http://www.nptel.ac.in)

Copyrights Reserved

