**Indian Institute of Technology Madras**

**NPTEL**
**National Programme on Technology Enhanced Learning**

**Pattern Recognition**

**Module 03**

**Lecture 03**

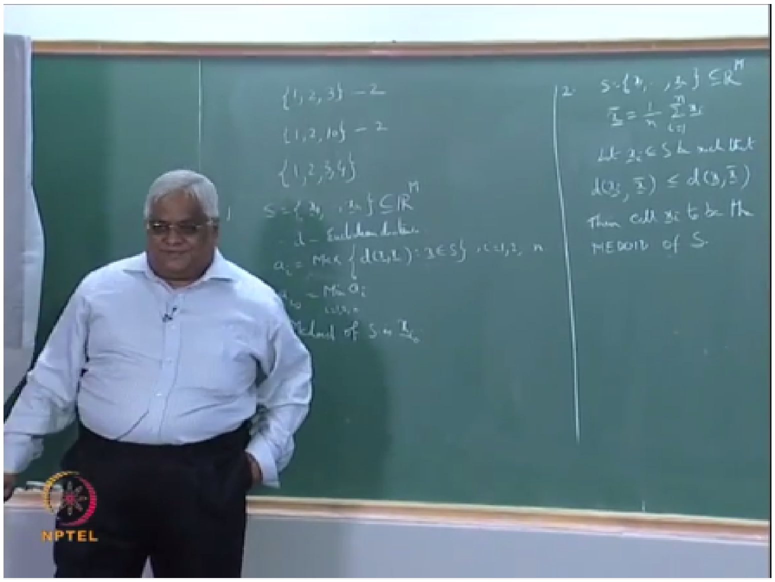
# K-Medoids and DBSCAN

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

We were discussing k-means algorithm single linkage algorithm in the previous few classes, so and today we will be discussing slight generalization of k-means algorithm this is a k-medoids algorithm. So instead of mean you are going to have medoid there, medoid the word it is it means it is a generalization of median for higher dimensions. I hope all of you know the meaning of the word median but let me just recapitulate for recapitulate to you the meaning of the word median.

So single dimensional data median is the middle most point ,what is the meaning of middle most point.

(Refer Slide Time: 01:09)



Suppose your data set is it has got just three points 1,2,3 then arrange them in increasing order in fact they are arranged in increasing order, the middle one is 2 so 2 is the median. Supposing your data set is so here it is 2, suppose your data set is 1 to 10 here also if you arrange them in the increasing order the middle one is 2 so 2 is the median here also. So basically arrange the points in the increasing order and the middle one.

But you can do it provided the number of points is odd, if the number of points is even then what is the meaning of middle one right, so let us see suppose your data set is 1,2,3,4 then arrange them in the increasing order there are four points here they are already arranged in the increasing order 2 you can call it to be the middle one 3 also you can call it to be the middle one, you can call 2 to be the middle one you can also call 3 to be the middle one.

So here the middle one is not unique then in this case there are a few conventions one convention is that you take the average of these two points 2 and 3 average is 2.5 and call it as the median that is so 2.5. But if you want to put a restriction that median has to belong to the data set, note that 2.5 does not belong to the data set okay, so if you want to put a restriction that median has to belong to the data set then that is fine then it can be either 2 or 3.

The median can be either 2 or 3 and it is not unique, so in single dimension this is what people do. But what will happen when you have multiple dimensions 2, 3, 4, 5 etcetera can you arrange the points in increasing order no, what is the meaning of arranging points in increasing order it is not clear in fact we do not have this ordering. I hope you know the meaning of ordering since I assume that you are all computer science students.

Probably you know the meaning of partial ordering, linear ordering etcetera okay, so for higher dimensions you do not have that ordering then what does one do there are the definition of median for higher dimension then the word is medoid that you will see different definitions at different places basically this concept is to be generalized to higher dimensions. So how does one generalize this let me tell you one definition in fact I will tell you two definitions let me tell you the first one of them.

Say your data set is x1 to xn points they belonging to the M dimensional space it is a subset of M dimensional Euclidean space and D is the Euclidean distance D is Euclidean distance, now what I do is that let me define what $a_i$, $a_i$ is Max yeah, I what you do is that from the point $x_i$ you calculate distance of every point in the set s and find the maximum, from the point $x_i$ you calculate the distance of every point and find the maximum that maximum I am denoting it by $a_i$ naturally i=1 to n okay.

Next I find $a_i$0, $a_i$0 is minimum of $a_i$'s and i0 is that subscript for which the minimum is attained then the median is our medoid of s is $x_{i0}$, let me explain. Let us take this one take the point 1 the distance of between 1 and 2 is 1, 1 and 3 is 1 right. So the maximum is 2 okay, when you take this point the value is 2 and when you take 3 this is 1, this is 2 and this is 1, for 3 also the corresponding value is 2.

What about 2 the value is 1 so you will get 212 minimum is 1 and that 1 is obtained for the value 2, so 2 is medium it is clear. Let us look at this example 1 to10 here for this the maximum is this is1 and this is 9 so maximum is 9, for10 also it is 9 but for 2 it is 8 so this is 9 8 9 it is clear, so minimum is at 8 and that is happening for 2 okay. Now let us look at this here for this the maximum is 3, for 2 the maximum is 2, for 3 also the maximum is 2, for 4 the maximum is 3.
So this is 3223 right, 3223 and the minimum is 2 and that is obtained for two values is this clear. So this is how this is A way of defining medoid, but why A way the reason is that this definition as it is it is fine but look at the number of computations you need to do. Look at the number of computations you need to do is not it very high right, so even though this definition is good okay, even though this definition is good but because these many computations are involved people do not like this definition.

So they have defined something else now that definition also let me just to do the definition is actually very simple, how does one define this one find the mean $\bar{x}$ okay, find the mean of the endpoints and let $x_i$ belonging to S be such that distance between $x_i$ and $\bar{x} \leq$ distance between x and $\bar{x}$ then all then call $x_i$ to be the medoid of S.

I will explain when we calculate the mean of any data set most of the times mean does not belong to the data set, most of the times mean does not belong to the data set it is something outside the data set then you find the point in the data set which is closest to the mean and call that to be the median or medoid okay, you find the point in the data set which is closest to the mean and call that to be the medoid.

So this $x_i$ is closest to the mean closest means this distance is the minimum among all distances right, see this is a mathematical way of representing it if you represent it in mathematical way there are absolutely no reasons for getting confused. Whereas if you tell something intuitively, intuitively it looks fine but the same intrusion you might be having different ways of putting it in mathematics.

So it is you how one has to have intuition but on the other hand one also should be able to write the corresponding mathematics, then everything will be crystal clear that is one of the reasons why in all my lectures I am trying to do the corresponding mathematics write everything in mathematical form so that there will not be any confusion whatsoever. So this is a way of writing this way of writing it in mathematical form distance between $x_i$ and $\bar{x} \leq$ all such distances that means $x_i$ is closest to $\bar{x}$.

Let me write this thing here because this is a vector so I am writing it like this, so $x_i$ is closest to $\bar{x}$ then call $x_i$ to be the medoid of this, please. Calculate and apply and going to a k-medoid given at the bottom then output will be similar to k-means only. Yes, the output would be similar to k-means no, no can you say that there will be absolutely no difference the difference would be small that I agree that I am not disagreeing to that the difference will be small that I agree.

I surely agree that the difference would be small but this is the way it is therein the literature I mean I can always say personally I do not like this okay, but you see one thing what is the mean of this one 1 + 2, 3, 3 + 10, 13 the mean is something like four point something what is the one that is closest to it 2 right, so 2 is the median and here also here 2 is the median and according to this definition also 2 is the median.

Basically when you are saying that it is very close to the which is very close to the k-means algorithm there is a general feeling in your mind that medoid is close to the mean that is true I agree, but the closeness may not be as much as you are thinking. In some cases it will be really close but in some cases this is the closest point but the distance may not be really as small as you think had it been 100 then also the median would have been 2 according to this one.

But then what is the mean 103/3 which is a very large value are you understanding what I am trying to say that is a very large value, so it is not really always as close to the mean as you are thinking I agree it is the closest because that is what the definition says but that distance may not be small it may be really large. From the mean right if determine very near and close then it will be very near to the mean in great.

See each one of the things I am understanding in what situation you are trying to say this thing but do you think that every situation would be like this that is one question I would like you to think just think about it and personally speaking I would prefer the definition 1 to definition 2 but then if you ask me that definition 1 needs too many computations I agree but maybe I am always more bothered about accuracy than the computations.

So I would prefer definition 1, but this is also one thing that you will see in many books you take the books on data mining where you will find too many clustering algorithms and you see the definition of medoid this is one definition that you will find in books, right. How much you like it that is a different thing but this is something that is existing okay, I mean you may not like it I may not like it that is a different thing but this is a definition that is existing other questions please.

Normally when we use just to take care of the outlet if you do a mean it might in that large value might influence my result so that is why we are normally goes for a median calculation but in this second definition like it is almost related to the previous question like actually we are not taking into consideration that we are actually taking that meaning to account so what probably might happen is that large value might influence the.

Here if you take 100 it is a very large value even then you are getting the mean median 1 to 100, 2 is the median and you follow this one you will get 2. This is see the difference in the second example is not that much like it is within that you did $10^{10}$ take $10^{10}$, $10^{100}$ even then you will get to as median according to this, are you getting and $10^{10}$ $10^{100}$ that is it is very much far away from 2 but even then you will get to as median if you follow this definition.

And it is not really as bad as you think that is one of the points that I want to mention it is not really as bad as you think and I am not saying that this is the ultimate you can always improve upon this, you can always improve upon this I am not saying that this is something like Bible or Quran and you should not change this thing that I am not saying you can always change it.

According to your convenience if you do not like this thing you do your own definition which has also some intuition and for which with these data sets you should get the same whatever medium that you are getting here you must get the same thing according to your definition after all people have tried to generalize it this is what they have given you may not always agree to this, okay you may not always agree to this but these definitions you find in books you can have your own definition.

Personally speaking I prefer this but I do not mind doing lot of computations okay, but that is again my personal choice and which I mean it is only my personal v let me just say that so I would prefer this but this is all one thing that you will gar you are going to find in books and it is not really that bad okay. So this is one definition this is one definition for medoid then you can have k-medoid algorithm which is you can make it exactly the same as the k-means algorithm for the k-means algorithm except that instead of mean you write their medoid.

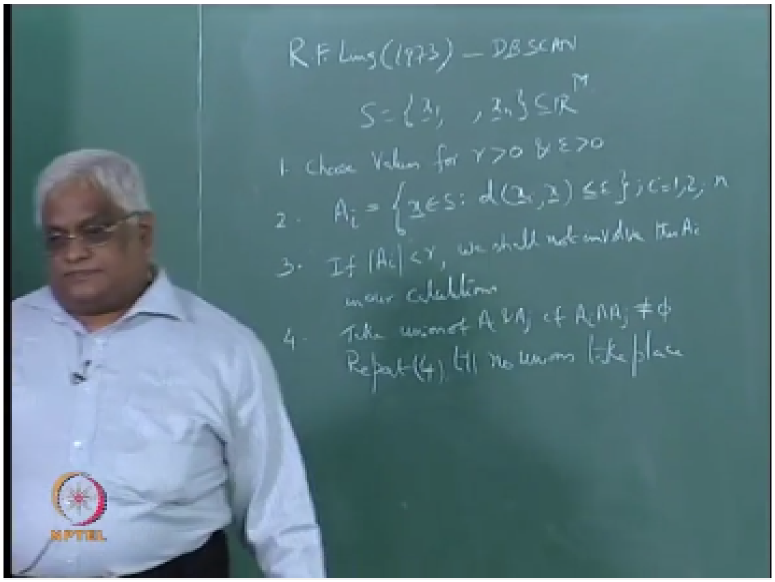
You can follow it in the same way and you can decide the number of iterations beforehand and for k-means algorithm people have tried to prove the convergence for a k-medoids I am not sure whether any proofs are existing for convergence, I am not sure of that but what people generally do is they assume the number of iterations beforehand maximal number of iterations is this much and then they run the algorithm and then they run the algorithm that is for k-medoids and which definition of medoid you will take that is again up to you that is again up to you okay.

Let us just take two minutes break just two minutes, yeah so I shall now teach you an algorithm which is popularly known as DB scan and but this DB scan it came into existence sometime in

95, 96 where the authors published the paper in a conference proceedings and it became really famous.

But unfortunately the exact algorithm it was published in 1973 in the Journal of American Statistical Association by an author called RF link Journal of American Statistical Association 1973 that Jogja 1973 RF link and he called it as generalization to single linkage method. So I am just going to call it as RF links algorithm and which is what I shall be doing it now.

(Refer Slide Time: 26:11)



R.F. Ling (1973) — DBSCAN

$S = \{\underline{x}_1, \ldots, \underline{x}_n\} \subseteq \mathbb{R}^M$

1. Choose values for $r > 0$ & $\varepsilon > 0$
2. $A_i = \{\underline{x} \in S : d(\underline{x}_i, \underline{x}) \leq \varepsilon\}; i = 1, 2, \ldots n$
3. If $|A_i| < r$, we shall not consider the $A_i$ in our calculations
4. Take union of $A_i$ & $A_j$ if $A_i \cap A_j \neq \phi$

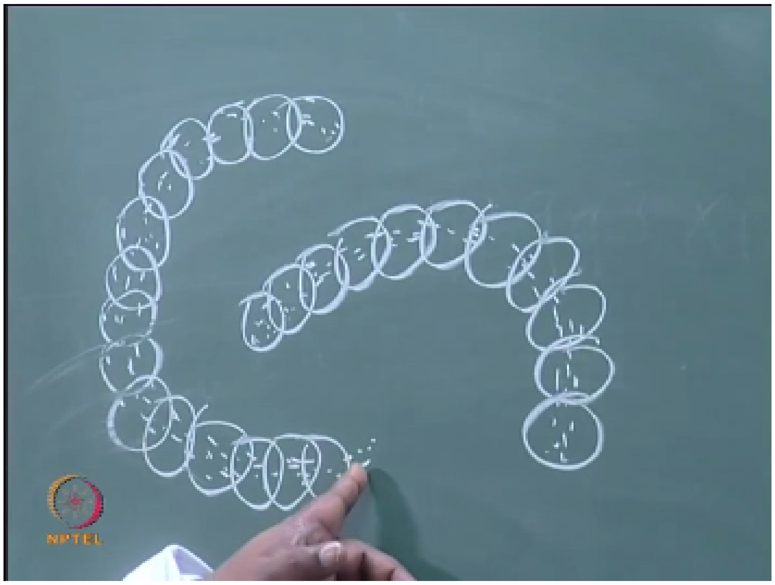
Repeat (4) till no unions take place

So RF link 1973 so you are given n points up set up $R^M$ you need to choose two values choose values for R>0 and epsilon choose values for R>0 and epsilon greater than 0 what are these $R^M$ Epsilon you will find it you will find out okay, 2 let us say capital Ai is the set of all x belonging to S such that distance between $x_i$ and $x \leq$ Epsilon that means for each one of the points $x_i$ we will consider a disc of radius epsilon and then find one all points from this set S they are existing in the set in this just in the disc.

So for each point x i's we construct a disc of radius epsilon and we find out all those points from S which are within the disc that we will do it for every $x_i$ so that particular disc that particular set we are calling it as $A_i$, i=1 to n. Now 3 if the cardinality of Ai<R then we shall not involve these $A_i$'s in our calculations that means we will only consider all those $A_i$'s whose cardinalities are greater than or equal to R.

We will only consider all those $A_i$'s whose cardinality is R≥R okay, now 4 what we do is that so from this step onwards all the $A_i$'s that we are considering there cardinalities are greater than or equal to R take union of Ai and Aj, if Ai intersection Aj is not is equal to ϕ, and this you go on and on doing it this you go on doing it, this step 4 you should go on and on doing it I will tell you the meaning of this.

(Refer Slide Time: 31:37)



So this is your data set say this is your data set and you will have to choose the values for R and Epsilon okay, you need to choose the values for R and epsilon. So basically you need to have certain radius and once you consider the disk how many points must be there in that disk that you need to have the value for R okay, what values you are going to choose take them that we can consider confer the present moment we can forget about it.

Let us just look at this step suppose the $A_i$'s that we got they are say something like this the $A_i$ is that we got say there are something like this these are the $A_i$'s that we have got and if you do this thing take union of Ai and Aj if Ai intersection Aj is not is equal to ϕ then in this one you are going to get all this as one cluster and you will get all this as another cluster, when I say that you should repeat this step you should go on and on doing it till no unions take place.

Repeat 4 till is no unions take place note that this method is going to give you the number of clusters automatically is this clear, this method is going to give you the number of clusters automatically provided you choose the values of R and epsilon appropriately you will get really good clustering if you choose the values of R and epsilon appropriately.

And if there are some points that are remaining for example here there are some points that are remaining which you are not able to put them in any one of the clusters now you can have your own convention about putting them into one of the clusters maybe you can have something like nearest neighbor to whichever cluster it is closest to put it in that cluster I mean you can have your own conventions.

The number of points that is less left out it will be very small percentage of points compared to the whole set that is the size of the whole set given that you have chosen the values of R and Epsilon properly. So this is basically DB scan what is that if then we shall not involve this Ai in our calculations that means all the Ai is that we are considering the sizes of those sets is at least equal to R, the number of points in those sets is at least equal to R.

It may be our R+1, R+2, R+3 any such number greater than that the number of points in a sense. If what is this the cardinality, cardinality is less than R okay, then we shall not involve this Ai in our calculations okay, we shall not involve that Ai in our calculation. So it means so happen that some points may have been left out I mean one cannot guarantee that I mean once you write this step you cannot guarantee that every point will be put into one of the clusters that cannot be guaranteed.

So if there are some points which are not going to any one of the clusters then you can have your own conventions like the nearest neighbor or any such one, so this is basically RF links method and as you can see if the value of R if it is equal to 1 what is it that you are going to get if the value of R okay. my question to you is can you get single linkage method if you choose the values of R and epsilon appropriately, can you get single linkage method if you choose the values of R and epsilon appropriately. If R is equal to 2.

Then you will get single linkage method think about it that is why RF link called it generalization of single linkage method that is what he called it way back in1973. Now there is the word density used here density is coming into the picture because in a disk of radius epsilon we want at least R number of points to be there.

So it is something like density of the point $x_i$ is at least equal to R density of the point $x_i$ is at least equal to R okay, so this is an algorithm with which you find the number of clusters automatically in clustering one of the first questions is that how do you decide a number of clusters.

If someone gives you it is fine if someone does not give you then is there any way in which you can decide then this is one of the algorithms where you can decide the number of clusters automatically given that you have chosen the value of R and Epsilon, given that you have chosen the value of R and Epsilon okay. I think we will stop here.

**End of**
**Moudule 03 – Lecture 03**

**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

**IIT Madras Production**

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

www.nptel.ac.in