Indian Institute of Technology Madras

NPTEL
National Programme on Technology Enhanced Learning

Pattern Recognition

Module 06

Lecture 03

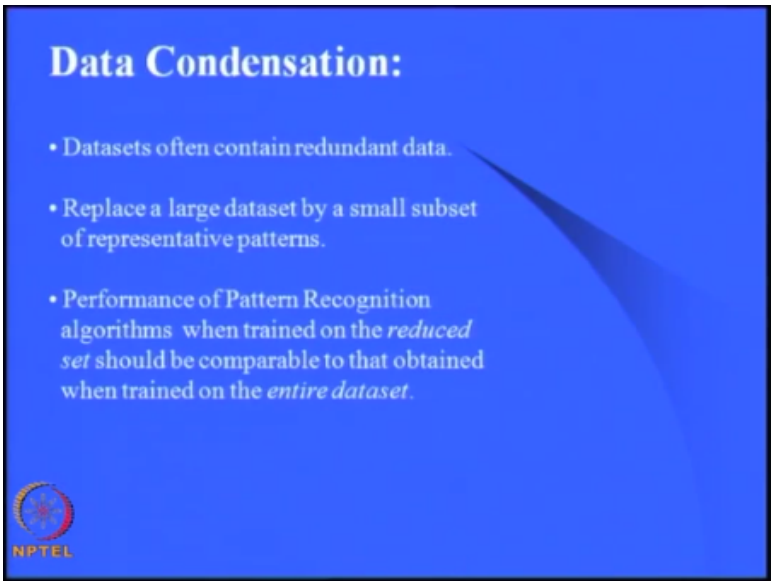
Data Condensation, Feature Clustering,
Data Visualization

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

So in this lecture I shall discuss a few topics which are very much related to pattern recognition techniques and which are in some senses they are becoming part of the pattern recognition techniques. The topics are data condensation this is one of the topic sand feature clustering which is again atopic in pattern recognition.

But it is now being vastly used these techniques have been are being vastly used in many other related subjects like machine learning and there is one another thing about which I would like to talk about that is data visualization which is very much in some way necessary to understand the behavior of the data. But there are some problems and some troubles in visualizing the data in higher dimensions so about which I would like to discuss little bit. So initially let me talk about data condensation, data condensation as you can see on the screen.

(Refer Slide Time: 01:33)



Suppose there are the data set size is huge now dealing with all the points in the data set that might be resulting in too many computations, so one may want to reduce the size of the data set and there are also other reasons the data sets often contain run and data. So what we would like to do is that we would like to replace a larger tacit by a small subset of representative patterns, so that the performance of pattern recognition algorithms when trained under reduced set should be comparable to that obtained when trained on the entire data set.

Now in this one performance of pattern generation algorithms when trained on the reduced set that is fine I am that is understandable in the sense that we can have some classification techniques or if we do not have any data that is labeled data if you have unlabeled data then what we might do is that we might have some clustering to be done on the entire data set. So instead of doing run the entire data set we would like to do it on a reduced data set, okay.

And so the third point as it is mentioned here this is fine but then replacing a larger tacit by a small subset of representative patterns that is the main a there are a few points which need to be discussed here, one is what is the size of the reduced set, what is the size of the reduced set if our total data set has let us just say 10,000 points then we can reduce it by half that is we can have 5000 points in the reduced set.

We can have 2,000 points that you set thousand points the reduced set 500 points 100 points 50 points maybe two points in the reduced set. So number one what is the size of the reduced set and number two when we are going on reducing the sizes naturally when we apply the usual

pattern recognition techniques on the reduced set and we would like to have the same performance as we can get when we use the original data set.

Now when the reduction is small when the performance would be expected to be very close to the actual performance, when the reduction becomes more and more that means from 10,000 when we come to let us just say 1000 or 500 then the performance may not be as good as what we expect because we have naturally reduced too many, too many points we have naturally reduced too many points so and when you want to make it from one 10,000 to just two points are so then it is going to be simply asked reduction.

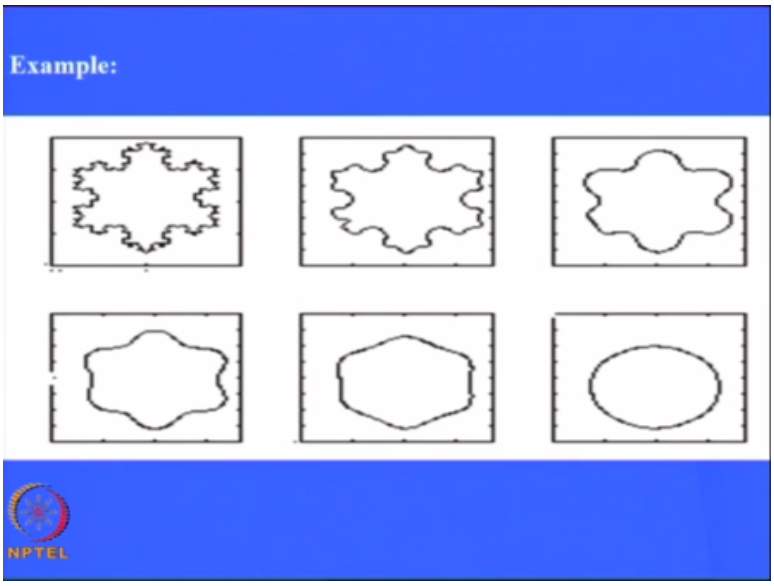
Now we have already done some such thing like this earlier 10,000 points reducing it to two points we have done it in clustering when we have done k-means clustering and if you take k=2 and you divide the data set by doing two means clustering and ultimately when the algorithm converges you will get two means these two means we can take them to be in some way representative of the whole data set.

So we have done this thing already earlier but we have not used this terminology as data condensation at that time we called it as clustering we did not use the word data condensation naturally those two means they are in some sense representing the whole data set. Now so actually on one hand clustering is one end of this data condensation clustering is one end of this data condensation and the another end is the entire data set whatever data set that you have.

And in between you have several different stages where we are reducing the size of the datasets in between you have several different shades where we are reducing the size of the data set. Now here there is another term that comes into the mind that is scale, if 1000 points are reduced to 1 over 10,000 points are reduced to two points then the scale is the ratio of 5,000 to 1, 1000 to 2 is same as 5,000 to 1.

Whereas from 10,000 if you come to 1000 it is 10:1 here it is 5000:1 and here it is 10:1 the same data set when we represent it by a different scales now you would actually the same data set when it is represented by different scales it looks in some sense different, let me just give you some examples.

(Refer Slide Time: 08:01)



Look at this, this one when the scale becomes coarser and coarser and the very last stage it might look like a like an ellipse because the finer details here number 1 they are missing okay, and number 2 they are also getting slowly and slowly joined you see this is your thing and it has some of these details they are missing here and some more are missing and here some more are missing and when it comes to this it has become like this and ultimately it has become this.

It is like there is another example we are standing here let us just say and you look at the sky during nighttimes we see too many stars and there are some patterns that you see there and let us just say we start going towards them we start going towards them and initially we have seen too many stars and when you start going towards them naturally some starts they start getting faded away and the one to which we are going nearer the details are going to become more and more clear, right.

And the closer you go there the finer representation you are going to get and it may so happen that whatever patterns that you are seeing from here quite a bit of maybe lost when you go nearer to them, I hope you are understanding what I am trying to say. So the same data when you are representing it in different scales probably different details you may be missing this is one point that one needs to take into consideration and there are in fact a few other issues which are related to this.

Let me just tell you one of them I do not know whether you people have used ink pens when you are in your childhood or writing on the paper have you done it and probably so now how did you

fill the tube probably you might have taken a filler and you would have put it in the ink taken the ink like this and put it here, right. Now you are supposed to fill up this whole tube but then probably most of the times you have you are never completely filled it up why, because you would see the last bubble that you are going to put you know how much volume atleast you guess how much volume it might occupy.

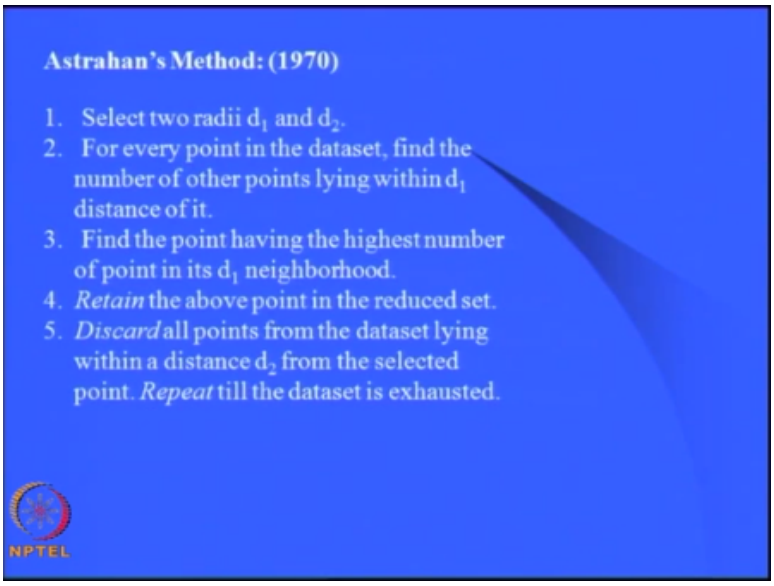
If you think that it is going to overflow then you would not put it there right, are you understanding what I am trying to say so it is confined to the size of I mean one I should say one drop of ink the amount of volume that may occupy. So everything is limited to that size you cannot reduce it further than that, similarly when you are doing digitization there also to certain degree you will do the distillation.

So everything is limited to that you are not in a position to make it go beyond that okay, everything is limited to that size you are not in a position to make it go beyond that, so with respect to that scale of digitization there with respect to the volume of the drop of the ink when you are filling up the tube. Here also there is a scale at which you are looking at the details so with respect to that scale there are some things which may be changing but you would like to keep many things as they are.

So totally here the problem is what is known as multi scale data condensation at different scales you are going to reduce the sizes of the data, what scale is needed that is the user who would say the scale that is needed and once that scale is given you are supposed to reduce the data set to that size, so this is the basic problem.

Now there are in fact a few methods already existing in the literature for this one of the methods is by Astrahan, what he did was in order to for k-means algorithm he wanted to get the initial seed points Astrahan for k-means algorithm he wanted to get the initial seed points.

(Refer Slide Time: 13:58)

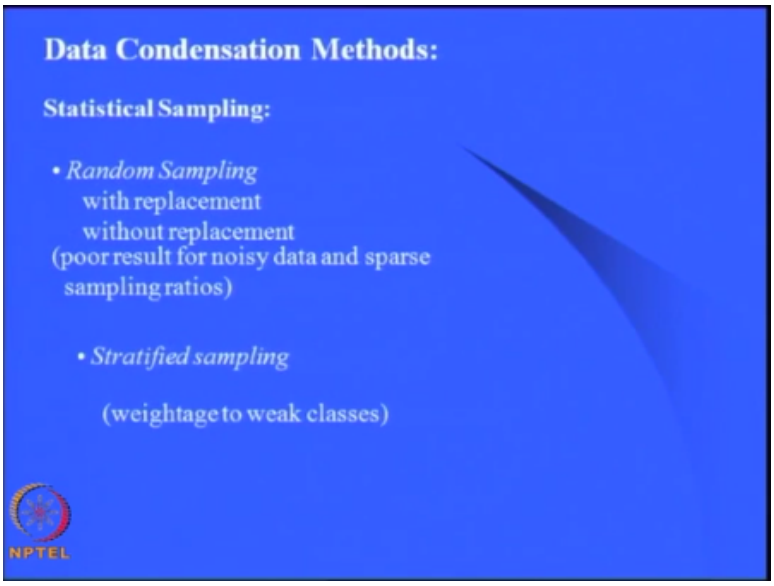


So what he did was now select to radii d1 and d2 and for every point in the data set find the number of other points lying within d1 distance of that, basically for each point you consider a radius of d1 and then you find how many points are lying in that data set within in that radius that means basically you are measuring the density of every point, basically you are measuring the density.

Now what you are going to do is that, you are going to find that point having the highest density and once you find that point you retain that point and you discard all points from the data set lying I mean within a distance d2 from the selected point and repeat till the set is you go on doing it if you are trying to get only K such points you just do it till such k such points are there and if you do not have any such restriction you go on doing it till the whole data set is exhausted.

What Astrahan was done it was he went on doing it till he has got k-means or k representative points, so he used it for extracting the initial seed points for k-means algorithm. So this method has been it is already existing Astrahan's method, but there is a small problem with this you have to select those to radiate d1 and d2 and the results are going to change if you change the values of d1 and d2 and there are some more methods.

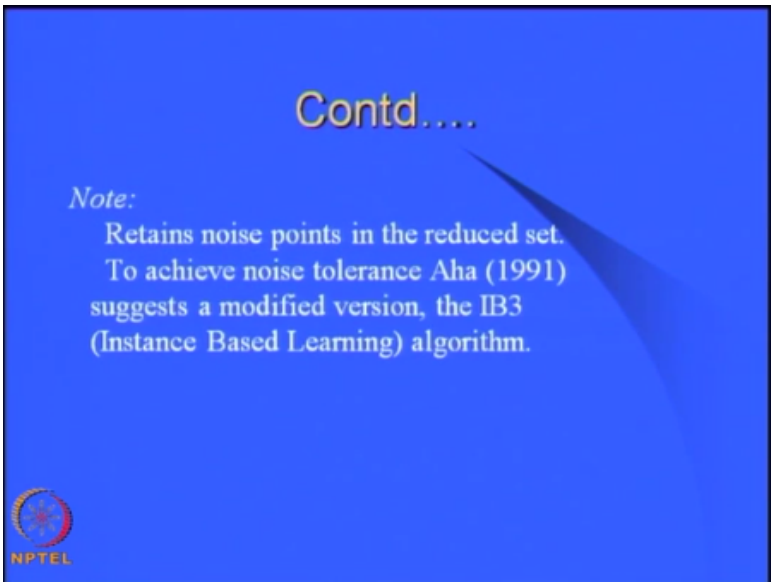
(Refer Slide Time: 16:16)



You have random sampling with replacement and without replacement, now what happens is that sampling methods are fine if you are really getting representatives I mean points from all over the data set. But if you are not getting fine if you are not getting points from all over the dataset then what happens is that the set is not exactly a representative of the whole thing. So you have these troubles with random sampling methods with replacement or without replacement.

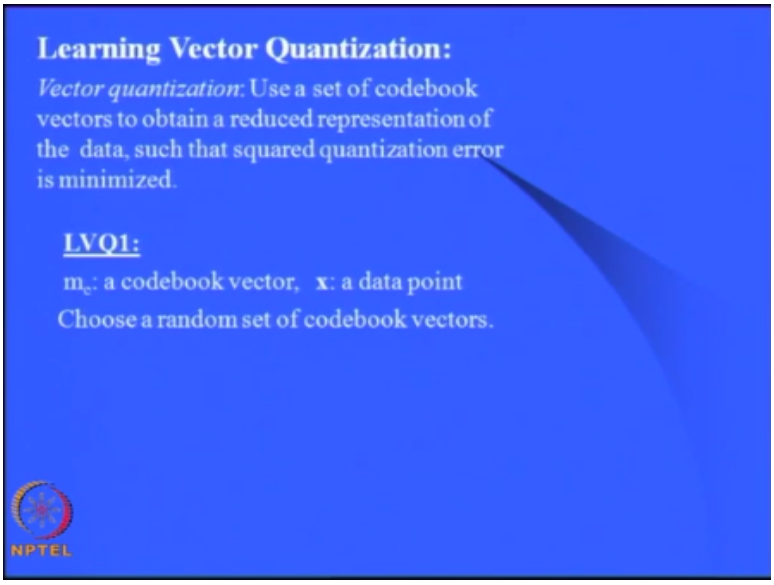
And you also have similar problems with stratified sampling you also have similar problems with stratified sampling.

(Refer Slide Time: 17:10)



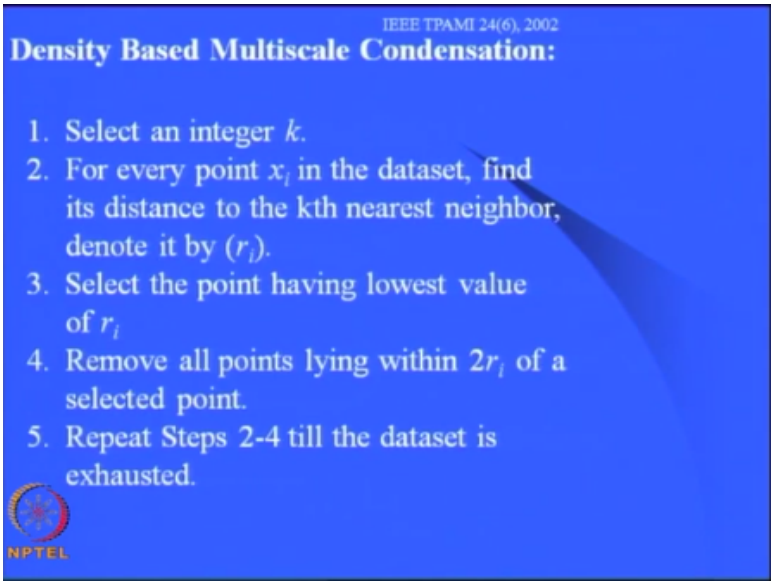
And there are some more methods I will not going to all these details about these methods I will just tell you a way of doing it one can surely use a Astrahan's method and I will just tell a way of doing and my point is that that is not the only way in which you can do it one can always improve up on these methods to get better and better results.

(Refer Slide Time: 17:46)



There are this learning vector quantization and there are many other such methods.

(Refer Slide Time: 18:05)

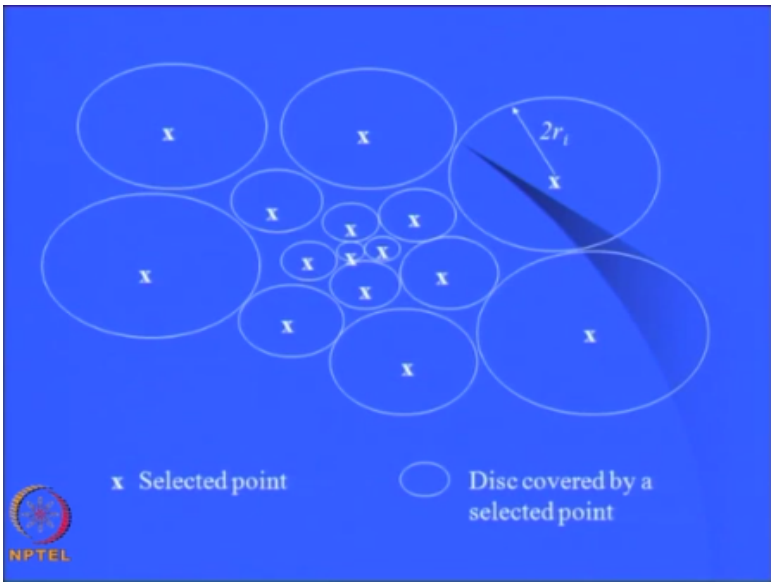


So this method that I am just going to tell you it is a slight variant of the what is known as the k-nearest neighbor decision rule which we have already discussed in one of the earlier classes in this lecture series. Basically what w eare going to do here is you select an integer k, k actually we will tell you this it is that multi scale data condensation it tells you the scale so once the scale the k is given for every point find its distance to the k^{th} nearest neighbor, okay.

You take a point in the data set and finally its distance to the k-nearest neighbor and if the point is x_i you call the distance as r_i now select the point having the lowest value of r_i select the point having the lowest value of r_i the lowest value of r_i means it has something like the density is quite high in a small radius you have many points lying so the density is quite high, then what we do is that remove all points lying within $2r_i$ of the selected point.

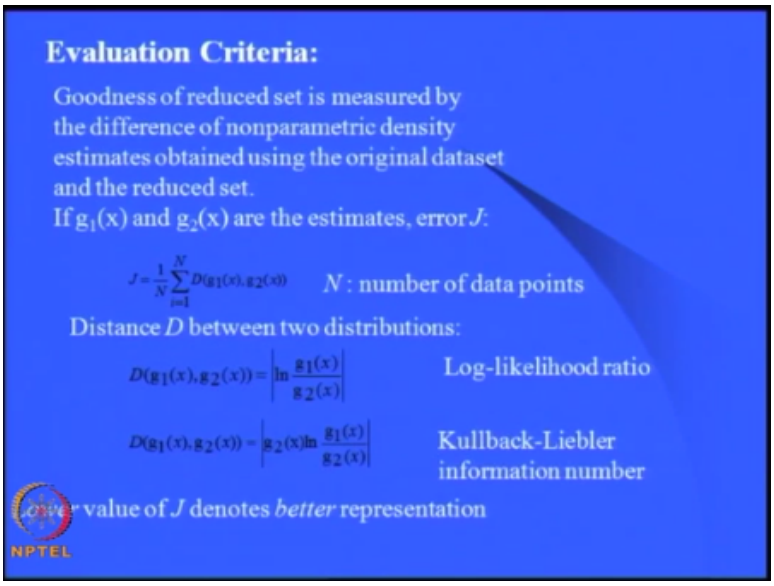
Two times the radius in two times the radius which ever point is there you just remove and repeat steps 2 and 4 till the data set is exhausted. So you go on doing it till the data set is exhausted.

(Refer Slide Time: 19:48)



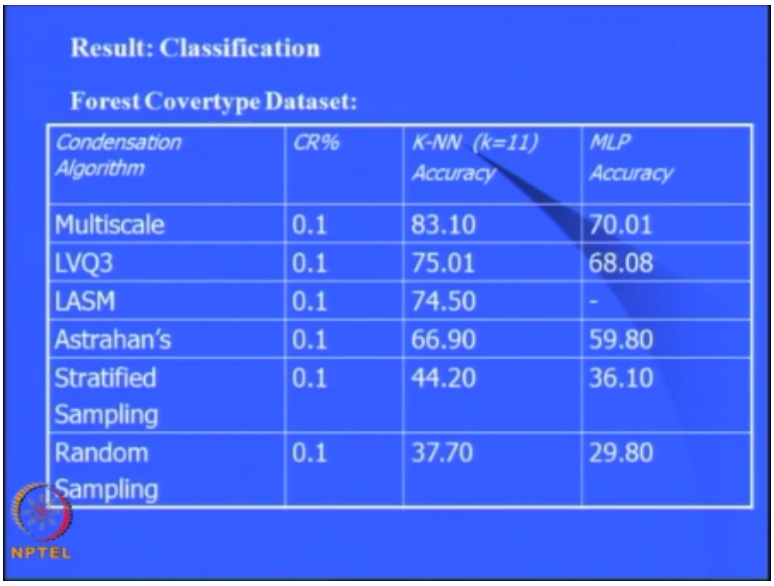
This is one of the results that is obtained when the data set I mean these are the representative points and these are discovered by this method.

(Refer Slide Time: 20:09)



And there are several evaluation criteria that are used you have something called Kullback Liebler information number okay, that is g2x lng1(x)/g2(x) and you have log likelihood ratio that in some way gives you the distance between two distributions and with these things you can try to evaluate the performance of the algorithms and the last method that has said it is found to provide good results on a few infect it is found to for provide good results on many data sets okay.

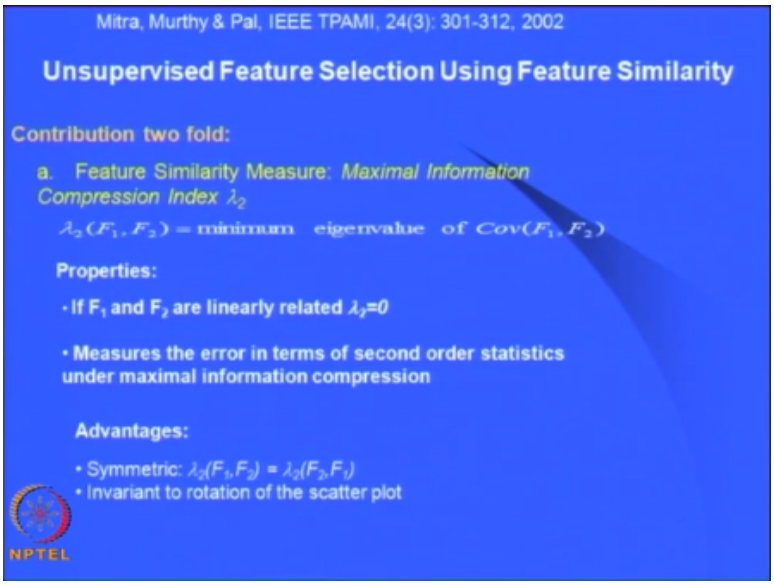
(Refer Slide Time: 21:01)



And compared to several other algorithms also compared to several other algorithms this is found to provide good results, but the point that I want to make is that always there is a scope for improvement and my aim here is to try to tell you what the problem is instead of trying to say what solutions are there which is better which is worse that is not actually my aim is to just tell you what the problem is and what are the main issues around it.

And so this is one problem that this is one problem very about which I wanted to tell you and there is another one that is clustering of features.

(Refer Slide Time: 21:51)

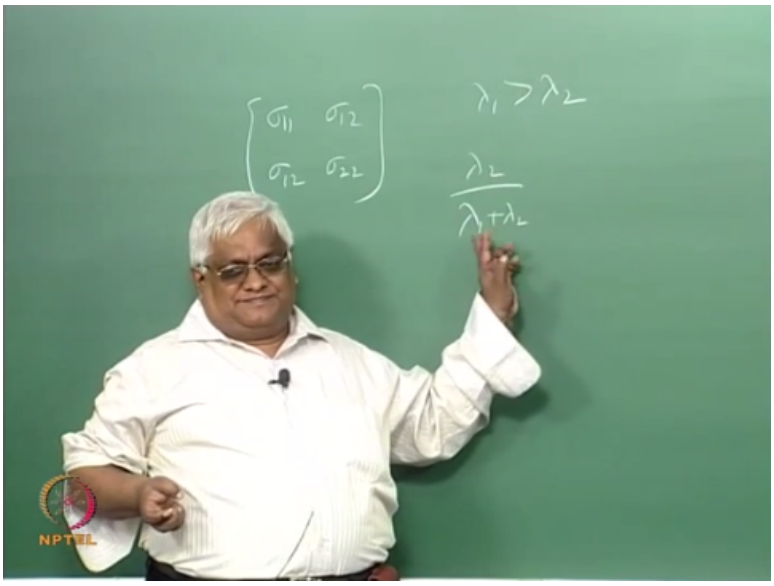


We have done already clustering of points now if you want to do clustering of features basically you need to have some sort of a distance measure are a dissimilarity measured you need to have between features a distance measure are a dissimilarity measure you need to have between features. Now there is already a similarity measure between features which is correlation coefficient there is already a similarity measure between features which is correlation coefficient but the main problem with that is that.

If you shift the data by some angle the whole data by some angle θ then you would like to have the correlation value to be the same but it is not going to remain the same at least one cannot guarantee that it will remain the same. So this is one of the troubles and there are a few I mean this is one of the main problems with the correlation coefficient. So basically what one would like to do is that you would like to have some sort of a measure which even when the data is shifted by some angle then it should not really change its value.

And there is one way in which one can suggest this one the way is one of the ways is for those two variables under consideration you please look at it the covariance matrix for these two variables the covariance matrix would look like.

(Refer Slide Time: 24:43)



For the two variables the covariance matrix would look like σ11, σ12, σ12 and σ22, now find eigen values and eigen vectors of this matrix let us call the eigen values as λ1 and λ2 and λ1 is greater than λ2. Now if the variables are related let us just say linearly then what is going to happen to the value of λ2, λ2 will be 0.

If the variables are related linearly that means there is a direct relationship between the values then this matrix will be of rank 1 since this matrix will be of rank 1 the determinant is 0 since the determinant is 0 the product of the eigen values should be 0, since this is a non-negative definite matrix both the eigen values have to be positive I mean greater than or equal to 0, so the second eigen value has to be equal to 0.

So if the variables are linearly related then the second eigen value is equal to 0, so somehow you can take the second eigen value to represent the distance between two variables which is what has been done here. So once you take the second eigen value whether you will take it in the absolute terms are you will write λ2/λ1+λ2 that is relatively that you can have many ways of doing this thing either you can take just λ2 or you can just take λ2/λ1+λ2 so that you can always make the distance between two variables lying between 0 and 1 lying between 0.

And in fact what is the maximum value of this is the maximum value of λ2/λ1+λ2, λ 2 is if for example this cannot be more than half can this be more than half the maximum value of λ2 is λ1right, when λ2=λ1so this is λ2/ λ+λ2 this is 1/2 okay, then the λ 1/2 run down it is have so the maximum value of this is it cannot be more than half. So then here we are trying to put everything in between 0 and 1/2 by taking λ2/λ1+λ2.

But again this is a way of defining dissimilarity between two variables now once you define this thing then you can always do a clustering of variable features you can always do a clustering of features which is I mean similar to the way we have reduced the size of the number of the way we have done the data condensation in the same way you can do freaker condensation, you take a point.

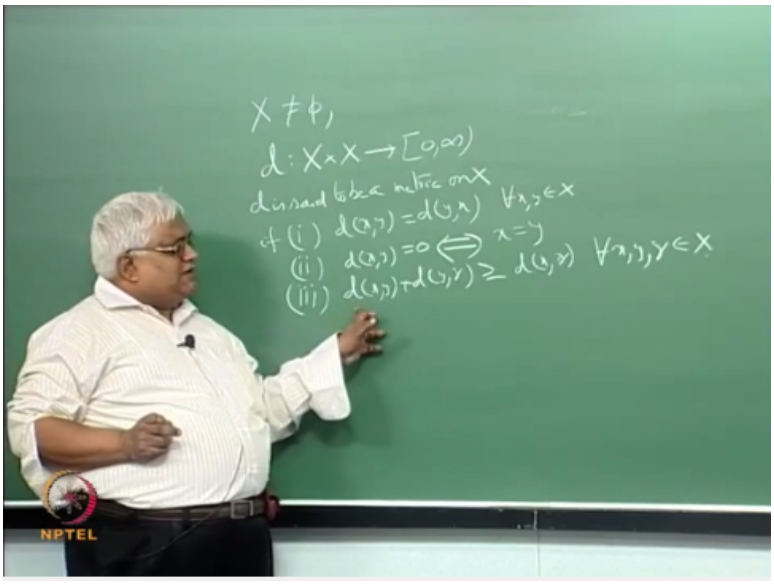
And find its k^{th} find a feature which lies find its k^{th} nearest neighbor find its k^{th} nearest neighbor means see how chosen the value of K let us just say k=10 then for this point you are going to find all features from this point and you are going to find the distances and you are going to arrange those distances in increasing order, you are going to arrange those distances in increasing order and you will take the 10^{th} such distance for the distance for which our variable which is going to provide that that is the 10^{th} nearest neighbor of this variable.

So you can do all these things and you can develop clustering algorithms using this sort of a similarity measure. Now this is only one such similarity measure one can always look at other similarity measures and in fact it is very much needed that and we need many other similarity measures which measure the similarity between two variables are two features in one of my earlier talks I was mentioning about dependency and independency.

We know the definition of independency of two random variables, so we say that two random variables are dependent if they are not independent but how much dependent they are we do not know. Similarly the distance between two random variables which is what I am trying to tell you here it is not exactly a metric whatever measure that was defined one can show that it is not a metric.

I do not know whether I have given the definition of metric corner did I give you okay, I think let me just define what a metric is.

(Refer Slide Time: 30:08)



Let us just say x is a non-empty set and d is a function, d is a function which is defined from XxX to 0 to ∞ and d is said to be metric on x, if this word metric is taken from mathematics this gives you the definition of distance in a mathematical way. The first property is that the distance it cannot be less than 0 the distance value it cannot be less than 0 it has to be greater than or equal to 0 second one is that distance between x and y it should be same as distance between y and x.

And this is if distance between two points that is equal to 0 then this those two points how to be same and distance between the same points has to be equal to 0 and this last one is the triangular inequality distance between x and y plus distance between y and z is greater than or equal distance between x and z for all xyz belonging to x. Now the distance measure that is proposed here it is not a metric that can be easily seen.

Now the question is can we define a metric between two features and if we can define one such metric does it really satisfy the inclusion of our inclusion of distance between two features number one how to define this a metric between two features okay, and the metric should be the definition of the metric that is given there it should be satisfactory to all the persons who are working in this thing in this field so there are two aspects.

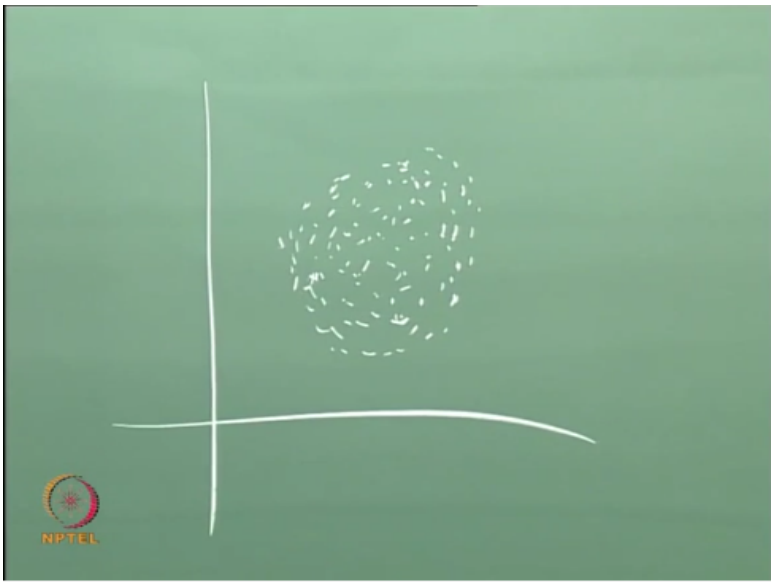
So these are all the problems on which people have been working, so here also by defining a metric we are trying to define some sort of a we are trying to measure the relationship between those two variables so by using this definition that is λ2 or λ2/λ1+λ2 some works have already

been done and there are in some cases some improvements are also made on this one by modifying a few concepts.

And it is very much necessary to devise some more feature clustering algorithms in different domains such as web mining domain and there is another domain that is bioinformatics where in these two domains the number of the sizes of the vectors the feature vectors that is huge and the number of points they become very small but the number of dimensions they become really large.

Now there is a third aspect on which I would like to discuss little bit the third aspect is data visualization, when we are given points in two dimensional plane.

(Refer Slide Time: 35:09)



You created in two dimensional Euclidean space so these are the points that are given to us we understand that this point is closer to this then the distance between these two points is less than

the distance between these two points similarly the distance between these two points is less than the distance between this and this. When we look at the data we understand these distances very well we see which points are closer which points are far apart we understand it, okay.

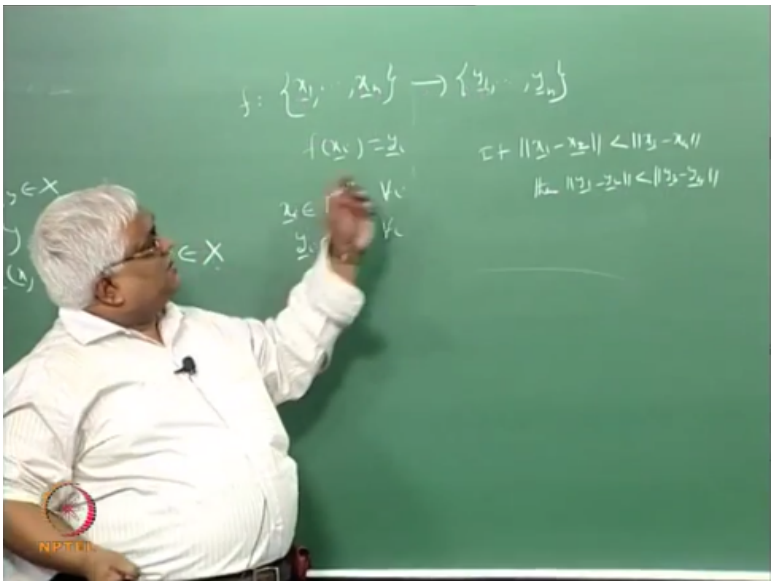
And we somehow try to get the shape of the data set for example or something like this we know that it is circular we somehow try to get the shape of the data set in two dimensions when we plot like this we understand the meaning of shape may not be always but many times we understand the meaning of distances between points which two points are closer which two points are further apart and which distances are smaller than which distances are larger than which distances all these things we understand them nicely.

But the moment the data is in three dimensions are more than 3,4,5,6 since we are not in a position to plot we are not in a position to visualize the data properly, so one of the problems that on which people have tried to work up on is how do you visualize the data of in multiple dimensions, multi-dimensional data in such a way that we understand those distances and there some distance is smaller than other distance some distance is larger than other distances so these properties we understand them well.

So in order we need to represent the data set in such a way that the property is regarding their distances are understood well by us that is the main issue of data visualization. This is in essence connected with cohesive self organizing map, I am not sure how many of you are ever of this cohesive self organizing lab are you aware of it they are also the issue is the same mapped into one dimension are two dimensions, right so that you would like to preserve the topology, topology preserving do you know anything about it, okay.

The issues are in some way related to cohesive self-organizing map in the sense that in that map from any dimension you would like to reduce the data set to two or three dimensions that is what Conan tried to do there and his aim was to preserve the topology as much as possible, what is the meaning of preserving the topology the meaning of preserving the topology is whatever I am mentioning regarding these distances that is what he wanted to do that is if.

(Refer Slide Time: 39:45)



That is if you have points say x1 to xn then you want to define a map from this to another y1 to yn okay, so that f(xi)=yi now let us just say this excise are in let us just say some10 dimensional space okay, and let us just say this is yiR in two-dimensional space, xiR in 10 dimensional space yiR in two-dimensional space.

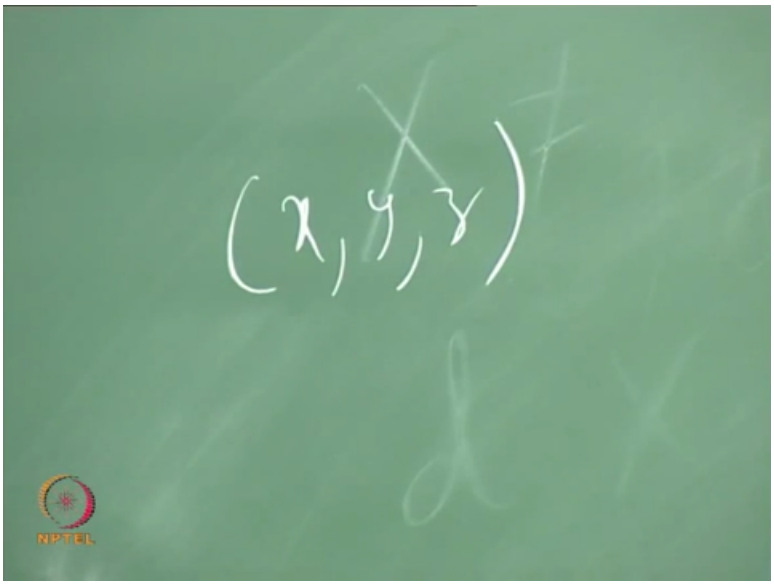
So from 10 dimensions you would like to have a map from 10 dimensions two dimension, so that you would like to preserve the topology, you would like to preserve the topology means if let us just say if distance between x1 and x2 is say less than distance between x3 and x4 then distance between y1 and y2 also should be less than distance between y3 and y4.

If this is your map and suppose between x1 and x2 and x3 and x4 this relationship is satisfied then correspondingly you want the similar relationship to be satisfied here. Now if there are n points there are totally NC2 pairs that means NC2 distances now if you want to compare one distance with another distance then two distances that is NC2 right, n into n-1/ 2, c2 so those many pairs you have those many inequalities or equalities you have.

And correspondingly you should have the same equalities or inequalities to be satisfied in among this yiR's also that is what you want. Now there is a small twist here the twist is you can show mathematically that it is always not possible to get it one can easily provide contradictions where the policy preserving maps are non-existent, you can easily provide contradictions in fact you can take your own data set I just want you to take you take unit cube in three dimensions.

How many points are there nine times 327 points I am not unit cube so +1 and -1 so 2,2x2 so you have totally 27 points are there from 000 on each side you go +1 or -1. So you are going to get 27 points correspondingly you try to get 27 points in two dimensions having the same equality or inequalities you cannot get it you can try to do it you cannot get it, are you satisfied with this 27 points you are not understanding right how the 27 has come.

(Refer Slide Time: 44:29)



You see each one you have three ways of doing it -1,0 and 1 each one you have three ways so 3x3x3, 27 right, so there are 27 points you try to get 27 points corresponding to this in two dimensions having I mean between any two vertices there are 27 vertices you can measure their distances and you can measure those inequalities and corresponding inequalities you should get in r2 and you must choose 27 such points sighs having the same inequalities you try to do it I can assure you that you cannot do it.

You simply cannot do it, what Conan tried to do was how much clothes you can go there, how much clothes you can go to that optimally that is what he tried to do and he gave an algorithm with that you have got the self-organizing feature map before Cohen came into the picture people tried it in too many ways some people try to give a tree structure for I mean looking at the some people try to give a tree structure and there are some people who just gave some faces.

There is a famous article and Chernoff ‘s faces, thesame Chernoff about which I was talking about Chernoff bounds there is a very famous article on Chernoff faces you just go through Google on Chernoff face they are also they try to represent data in multiple dimensions using phase. So like that you will find several different works, but yes, I said on one hand for some things it is impossible I told you very clearly but how close you can go to that which is what Conan try to do.

And probably what coherent right to do people can improve up on that but how to improve up on that and how to do that it is a matter of further research, basically if topology preserving maps are always not possible there are theorems and proofs corresponding to this statement if you go through some books on topology you are going to get the theorems and proofs for the statement that I made.

Topology is basically a subject in mathematics, it is a subject in pure mathematics from higher dimension to lower dimension topology preserving maps they are always not possible that this thing was proved by people long back. But data visualization is extremely important data visualization is extremely and extremely important and okay, something is not possible that is fine but to how much extent you can visualize are you in a position to develop techniques whereto the extent that one can visualize these techniques will make us visualize.

This is also another issue on which I mean so I mean what is going on and especially when you have simply very high-dimensional data present nowadays both in data mining literature and in bioinformatics literature as a web mining where you are dealing with web pages. So data visualization has become really and really important, I am stopping it here if you have any questions please ask me, please. So instead of data visualization like if we just find out a manifold of the data is that an alternative.

I do not want to say that it is not an alternative I surely do not want to say that okay. Manifolds are useful in many situations I am I should really agree that, I surely agree to that and if you can really have some idea if you can really get some idea about the data using these manifolds it is fine there are I mean I am always for it.

But how do you generate this manifold it depends on that depends on that very much okay, the generation of manifolds yes, I know that in face recognition literature people have developed

manifolds for I mean for changing the faces by using manifolds they are doing many things and face recognition that I have seen those papers. But it is not exactly data visualization you see they have constructed a manifold for some specific purposes that is absolutely fine and yes, if you can visualize or if you can get something more about the data than what you know already using manifolds it is fine for me I simply do not have any problems with that and I encourage that.

It is an alternative or not I do not want to use the word alternative you can use the word alternative when you are able to get good success with that then you can call it alternative, but I do not know whether you can achieve the success if you achieve the success value can call it an alternative I do not have any problem with that okay, any other question.

So can we do something like PCA and idea to visualize means reduce the dimension and then try to visualize the data. No, I mean you can always reduce the dimension and by reducing the dimension you are losing that information many times you are losing the information the information that you are losing it may be very small amount of information but you are surely losing the information yes, I mean you can I am not saying that you should not do it because now the question is what is the alternative.

I do not see any other way of visualization so when you are trying to do it in one way I cannot say that you should not do it if I do not give you an alternative and I am not in a position to give you an alternative, so you can always do that but that is not the end of it because you must understand that you are losing the information whenever you are doing any sort of PCA, because after all some Eigen values are greater than 0 and you are not taking them.

You can say that the amount of information lost is small that is a different matter but you are still losing the information whereas here what we want to say is that we want to preserve those distances are you in a position to preserve those distances when you are doing this PCA idea sort of thing are only the PCA sort of stuff are you in a position to preserve those distances for this you cannot make any guarantee you cannot give me any guarantee to the effect.

So when you cannot give me any guarantee so then you really do not know whether you are able to keep track of the distances or not, so I mean you can always do that but you need to understand the limitations this is what my point is okay, any other question, thank you.

End of
Moudule 06 – 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