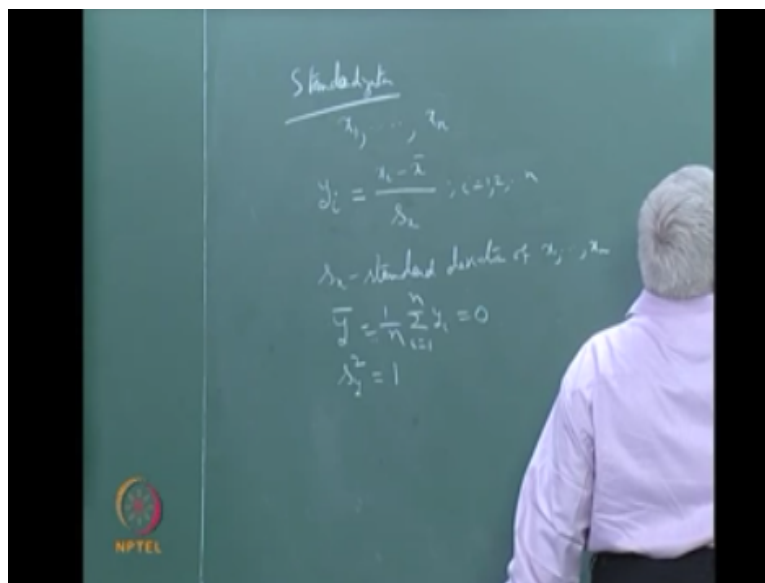**Pattern Recognition**

Module 02

Lecture 05

Standardization, Normalization,
Clustering and Metric Space

Prof. C.A. Murthy

Machine Intelligence Unit
Indian statistical institute Kolkata

Yeah I was talking about standardization and normalization if you have a variable.

(Refer Slide Time: 00:18)

 Which is taking values x1 to xn then a way of standardizing it is you just write new values as Y ice this is the right the new where new values as Y is why I you write it as X i- the mean x-bar the mean x-bar of these observations divided by the standard deviation SX is standard deviation of standard deviation of these observations x 1 pixel I hope all of you know the meaning of

standard deviation the square of this is actually going to give you the variance the square of this is going to give you the variance this is a way of doing standardization of the variables.

Now what is the specialty of this once you do this one then the mean of Y I that is y bar 1 over n summation I = 1 to n why I this is actually equal to 0 you are making the mean to be 0 and you are making the variance of wise to be equal to 1you are making the mean to be zero and you are making the variances to be one so the example that I was telling you zero to 100 since the variation is more you are going to get more distance values once you do this one you are making the variance to be equal to 1okay.

So if you have in have a were you have a variable where its variance is something like say point five then after you do this thing is going to become one and if you have another variable whose variance is the 100 after you do this one that also going to become one so both these variables where for which the variance is for one of the variables the variance is less than one for another man the variance is say greater than one for both of them the variances are going to be one and the means are going to be 0 so this is one of the methods this is one after this is generally accepted way of standardizing a variable but whether you will do standardization or not that depends on the problems that you have.

I am not saying that every time you should do it neither I am saying that every time you should not do it depends on your problem whether you will do it or not if you do standardization then your variance covariance matrix it is going to become correlation matrix that is all the diagonal elements will be one and all the off-diagonal elements will be the correlations between those two variables the correlation between those two variables that will be the they will after acknowledgements and the diagonal elements will be all one.

I hope all of you know the meaning of the word correlation okay I hope all of you know the meaning of the word correlation so if you do standardization then the variance covariance matrix is going to become correlation matrix and we general did we do pc on variance covariance matrix then you may try to do it over the correlation matrix normalization this is standardization.

(Refer Slide Time: 05:14)
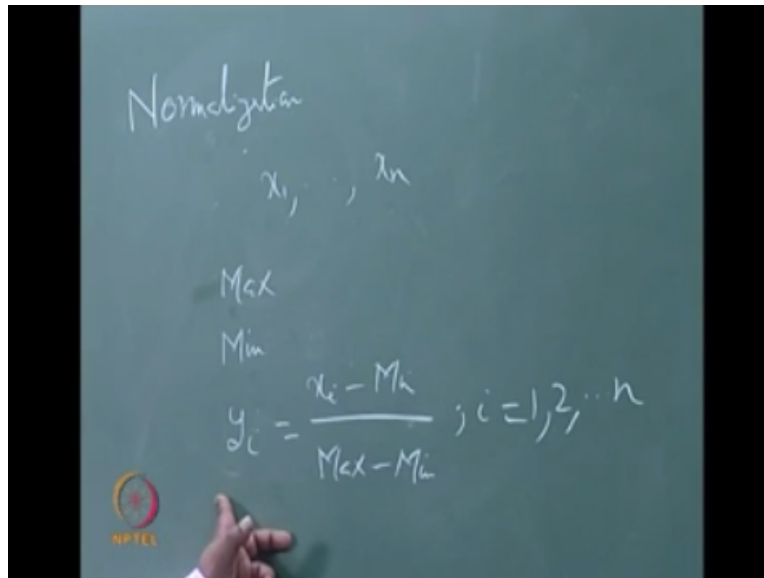
Normalization is again if you have x1 to xn then you write find max maximum of XI values find minimum of X I values then you write why I as X i-minimum by max minus min this is a way of doing it this max is maximum of the XI value min is minimum of the X I values so find the minimum of the X I values find the maximum of the exercise values then you write new variable Y I as X I -minimum / max - min so when this X I is actually the minimum you will get the value zero when the X I becomes maximum you will get the value 1 so then all the values are in the interval 0 to 1 all the values are in the interval 0 to 1this is slightly different from this here the mean is 0 variance is one here.

All the values are lying in the interval 0 to 1 this is a way which many persons follow for normalization again one can think of a few more ways of doing these transformations this is not the only such thing you can think of a few more ways of doing it whether you need to do it for your data set are not that again depends on you and if you do it then which particular transformation you will do and why that again depends on the data set that again depends on the dataset with this basically I would like to go to actually another with this one I will end this lecture 7 I should say classification class.

If with this something I will end the lectures on classification now I will actually go to clustering that is unsupervised classification this flow things yeah face the small modification the range can be changed to + 1 to -these windows root or you can do that I mean you can actually instead of making it from 0 to 1 you can actually make it from -1 to 1 so I mean probably a way of doing it

is this is from 0 to 1so multiply it by 2 then this will become 0 to 2 then you just subtract 1and this will become - 1 to 1 this should become -1 to 1 so you need not have to make it always between 0 to1 you can think of one particular interval and you can keep everything with confined to that interval please give an example they are doing.

The standardization and normalization will not be useful or may not be useful yeah basically there are some places where you need the information on variances to be preserved when you are doing the standardization then you are reducing are increasing the values of variances generally it is reduction but so those places where you want the information on variances to be preserved then that is the place where you should not use this thing and there are many examples where variance provides lot of information about the phenomenon under consideration about the phenomenon the consideration many times variance provides lot of information.

So they are the places where you should not you should not do the standardization you should not do the standardization and you the places where you are forced to-do the standardization is are those places where you have some variables with smaller variances but they have the capability of distinguishing between the classes and you do not want those variables to be submerged by the variable with larger variance but it does not have any capability of distinguishing between the classes then they are the places where you want the one with larger variance to have smaller one and the one.

That has smaller variance you want it is wise to be heard so that is the place where people try to-do and if they feel that they are the problem under consideration is like that then they do the standardization one of the places where you want the variances to be preserved is say the example about Punjabis and South Indians during the classification we are taking the height you see there the height information for I am assume that Punjab is are generally have larger heights but there are also many Punjabis we have smaller hives so probably their variances that variance has more.

 I mean variance is more for Punjabi class then probably for the South Indian class but there that variance should be preserved the variance information you should not change there and the other example about length of the notes that I gave that there you need to change their you need to change it because the variance for this thing this is very small and you need this variable to be heard so many of these things I will try to discuss in feature selection when I give lectures on

feature selection I will try to describe these things I will try to I will connect up this thing with that ultimately these are connected to feature selection.

When you are given a data you will do standardization or normalization that depends on the problem at hand and for the problem at hand which features are giving you the information about the problem at hand that you need to somehow do the make the decision and that actually becomes the problem of feature selection so problem of feature selection problem of standardization normalization they are very much connected they are very much connected other questions please so when we are doing standardization is the topology of the data prism what is the your meaning of the word topology either how its distributed the structure.

 I am not getting it is got me this evening please explain scatter of the data the distribution of the classes is the freezer I am not understanding but what I can tell you is this for two different variables the reduction and increase it is not in the same proportion if that is what you want me to say I think that is what is your doubt for two different variables the increase or decrease is not in the same proportion it is still not for all it is not in the same proportion different proportions so if you have different proportions information of the distribution information will be lost when you do standardization how.

The eye in the space feature space so that is not present suppose if you are scaling it down equally if you are scaling down so say you know I think you see there are many words the word for example use the word topology and there are some such words these words in my opinion they should be used very sparingly because they have they contain lot of lot of things probably in my tomorrow's lecture I will I will be actually talking about topology there I will tell you i mean the meaning of open set closed sets closures compactness connectivity all these things are therein topology so the word when you are using the word topology.

 I do not know which one you are trying to me I do not know which one you are trying to me basically the word topology is used by computer scientists in one way somehow if they preserve the distances then we say that it is topology preserving am a right between the most we could mean the is quant two lines and maybe same distance I do not think they are not going to preserve the distances that is what I said at different proportions they are going to change go through sheer yes for both the variables here you have two variables for both the variables the mean is going to become 0 and the variance is going to become one.

If you have two variables and I was giving you an example yesterday where one variable has variance to another variable has variance 1 if you take s if you look at the scatter plot how it looks like it looks like an ellipsoid if you have the covariance to be 0 now hear that two and one they are going to become one an done and the covariance is zero means there will be a circle so that means what is happening it has changed the distances have changed and then we also happen then they will come closer further that can I am not saying that in every situation you need to do the standardization neither.

I am saying that in every situation you should not do the standardization you need to make a decision whether you will do standardization or not and that decision is imperative in many problems in many problems the decision is imperative other questions ah nice yeah I had been talking about classification methods.

(Refer Slide Time: 17:25)
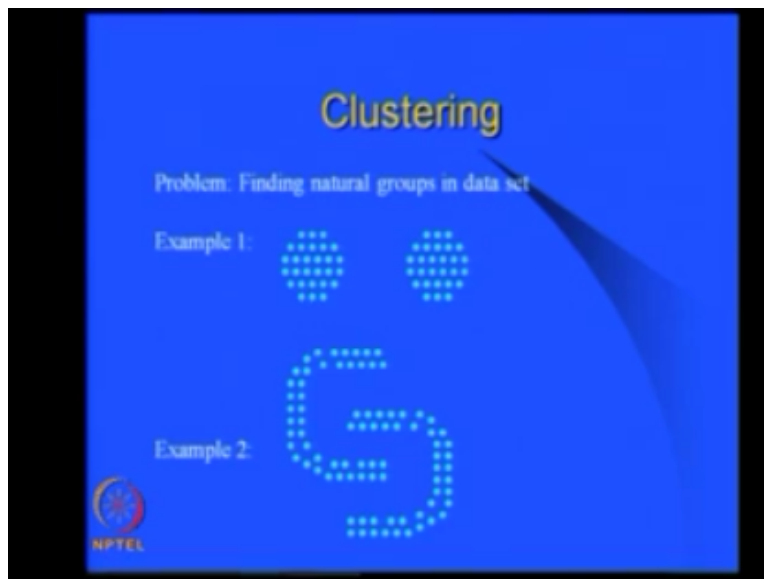
Now I will be discussing the clustering in clustering the basic problem is you are given a data set and you really do not know how many groups are there in the data set the word group I am using it loosely okay it is not the group that you would see in group theory rings and fields the word group is the way an ordinary man uses a word group uses the word group in the same way I

am also using it for example look at the two data sets that I have given here in this data set you would say that there are two clusters this is one cluster that is one cluster are there are two groups of points this is one group that is another group and in this data set again.

You would say that there are two clusters this is one cluster that is another cluster right now this is something when we are looking at the data set we are saying it and you want the same thing to be said are to be concluded by your algorithm your algorithm should automatically state that in these two examples the number of clusters is to number one and number two once it says that number of clusters is two in the second one it must be able to state that the first cluster is this is a cluster of points this is one cluster and that is another cluster on points similarly in this one this is one cluster of points and this is another cluster of points this is what you are algorithm automatically should state as the name of the subject is recognizing patterns pattern recognition.

So you are given a data set we are able to recognize the patterns and you want your algorithm to recognize them automatically now let us see what are the difficulties here.

(Refer Slide Time: 19:45)
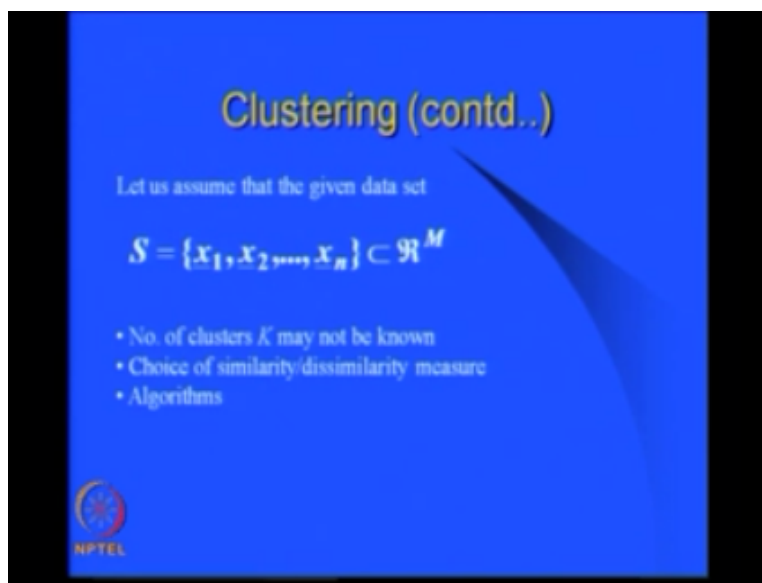
So if you want to state the problem the problem would look like this you have your given a data set maybe in M dimensional space here the number of clusters k may not be known and then there is something called a choice of similarity dissimilarity measure and algorithm what is this choice of similarity and the similarity measure you somehow feel that this whole thing since it is you are saying this one cluster somehow they are I mean similar to each other similarly these

points are in some way they are similar are they are connected to each other and here also these points and the same is the case with this you have some sort of a similarity relationship that relationship.

Somehow you need to mathematically state it the clustering problem which in many books it is stated as you would like to find groupings in the data set so that in each grouping the points are similar to each other and from one group to another group you will find dissimilarity let me repeat it in many books the problem of clustering is taken as you need to find groupings of the data so that in one group the points are similar to each other in every group the points are similar to each other and between one group to another group you will find lot of this similarity so this word is similar and dissimilar.

They need to be properly defined and that is this point choice of similarity and the similarity measure and then number of clusters has is it the case that someone gives you number of clusters and once you are given the number of clusters you will find them are in some problems you need to find the number of clusters on your own in some problem you need to find the number of clusters on your own it is not necessarily true that someone gives you the number of clusters always let me tell you an example where you may not know the number of clusters where you may not know the number of clusters the example that a stage it is again related to satellite images it is related to land cover types it is related to satellite images.

It is also related to land cover types my examples they are based on the Kolkata data because that is the place where I decide I used I always ask my students that how many land cover types are present in Calcutta then they would say water what are you should live on land cover type and then you have concrete structures sure that is building areas and you have vegetation where you have some paddy fields and there is open space that means barren land not in Calcutta but in the suburbs and there is one place called Madonna in Calcutta that is also just barren land so these are basically mean four types of land cover type mainly four types.

Then I would start asking them questions like well you have pond water and you have river water do you think there is any change in this then initially they would say and know all the water should be same then actually what happens is that when the satellite is taking photographs it is giving actually it gives two different signatures for pond water and river water and it is true

of Calcutta and I assume it to be true of even this place you have river kaveri right and i do not know because I have not seen the dataset.

But I presume that the kaveri river the pixel values would be different from the pixel values of the ponds I presume that I do not have any way of judging it I am not in the data but I presume it usually this is a mistake that people make all the water should be same the answer is not correct it is not true surely seawater is different from river water and in most of the cases river water is different from pond water seawater here in Madras you have see also there is water in the sea the sea water is different from river water we all know that sea water tastes salty right whereas pond water or river water at least.

The test is not salty this is something that we all of us know and you will see that they have different signatures so you cannot put the word the water add just one cluster you need to have that thing to be broken into two three components okay you need to have that thing to be broken into two three components they may not possess the same signature from thus the source which is giving you the values are these things it may not give the same value for all these types of waters okay so this is one thing that you need to keep on your mind you may think that you know the number of clusters.

But that may not be the case you may not you might have made some assumptions in saying that you know the number of clusters that may not be true you need to verify those things okay so this is an example where you may not know the number of clusters even though you know the place and you know the whole area yeah so number of clusters k may not be known in fact I can tell you very many examples where the number of clusters is normally just one example I can tell you one example relating to Telugu language my one mother tongue.

I can tell you one example into Telugu language telugu ovals we say ah ooh oh these are the vowels right but there is also one vowels that we always use but that is not mentioned in the text books there is a one word all these things I am Telling to him one word called kotuku to him pata Ko there the oval is not neither are not as it is something in between.

So this has come when is I has done projection till global recognition this particular thing has come as a separate vowel pat a coup that sound I that separate that has come as a separate verbal so there again the number of groupings is not known even though the mother tongue of most of

the persons who are conducting this experiment which is still go so I mean you assume that you know these things but you may not have an ultimately you may need to change so once you somehow get K and once you somehow get this similarity or dissimilarity measure then the next step is you need to have an algorithm.

To do the clustering before I proceed to all these things let me ask you a question E given the dataset do you think clustering is unique I will repeat given a data set do you think clustering is unique that means given a data set is it possible to get to three equally meaningful clustering the answer is unfortunately yes given a data set you can get many meaningful clustering I will give you an example probably all of you know what playing cards are Tosh in Hindi there is the same thing in Bangla you have 52 cards okay if you ask a child to do with the clustering of this 52 cards probably.

The child will divide the cards into two clusters all black cards all red cards and if you make a slightly grown up child to do a clustering probably he will put into four clusters clubs diamonds hearts and spades some other child may put all the face cards on one side and the other cars on the other side some other person may put all SS as one cluster kings queens jacks like that one may have cuddling clusters now so I have told you four different clustering's I am asking you which one is better and why can you remain an answer so given a data set you can have equally meaningful clustering's given a data set.

You have equally meaningful clustering's then when someone gives you a data set and ask you to do the clustering usually that person has some particular some specific property on the mind on his or her mind so you need to do clustering of the data based on whatever he or she tells you I hope you are getting the point that I am trying to state so it is not necessarily true that you have the same I mean unique clustering of the data so then when I say that the number of clusters is somehow you will find you will find some choice of similarity or dissimilarity measure the similarity or dissimilarity the example that I gave you I think the choice of similarity or dissimilarity is very clear.

If someone says all black cards and one cluster red cards there the similarity measure is very clear to you if someone says clubs diamonds hearts spades for different clusters they are also the similarity or dissimilarity is very clear to you right similarly aces kings queens just like that 13 clusters they're also similarity or dissimilarity is very clear to you note that if that similarity of

the similarity measure changes then your whole clustering will change I am right if the similarity dissimilarity measure.

If it changes then the whole clustering changes so depending on what you want to achieve you need to define the similarity or dissimilarity measure properly otherwise what you want to achieve you may not achieve okay so then the last part is algorithms yeah we all know we all know what Euclidean distance is before I talk about this dissimilarity measures let me tell you a little bit of what a metric space is there is something called metric space.

(Refer Slide Time: 33:07)
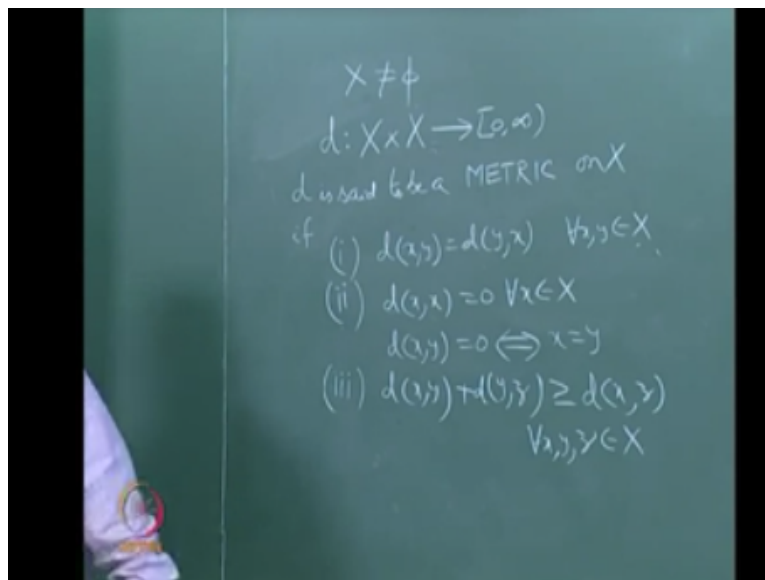
You have a set X this is not a null set okay, you have a function defined from cross product of XS 2 0 to ∞ now D is set to be a metric on X if d of XY is same as d of YX let me explain. To you these three properties the word d is used d is for distance the word d is used the symbol d I am using it for distance we use the word distance very loosely so in mathematics they have given a separate word for it and they gave a definition to that the word that they gave is metric ,metric and this is the definition of metric D is said to be a metric on X if these properties are satisfied now what are these properties.

If you go from X to Y whatever may be the distance that is the same as distance from Y to X which is totally true distance between X and Y and Y and X they are the same for all x and y and distance between X and itself is zero okay and if distance between X and Y is 0then X is equal to Y if distance between X and Y is 0 then X is equal to Y then what about this is distance between

x and y +distance between y and z it should be greater than or equal distance between X and Z for all X Y Z belonging to X this property is known as triangular inequality.

Take a triangle sum of two sides is greater than or equal to the third side I am right infect it is strictly greater than the third side then only you can have triangle okay so this is basically sum of two sides okay, if all the three points are on a straight line then you will get equality if all the three points x and y's XYZ that are on straight on you have already quality otherwise you will get inequality so this property should hold good for all X Y is that belonging 3x so a function D satisfying these properties is said to be a metric ton X is said to be a metric onyx and such a space is called metric space and there are the spaces.

That we are seeing are r square r cube R to the power of 4 matrix are defined on that and we are using those metrics probably without knowing that they are matrix unreal line it is basically if you have two values x and y the distance is measured as modulus x - y and you can show that is a metric it satisfies all these properties okay.

Dissimilarity Measures

- Metrics

$$\underline{a}^{'} = (a_1, a_2, \ldots, a_M)$$
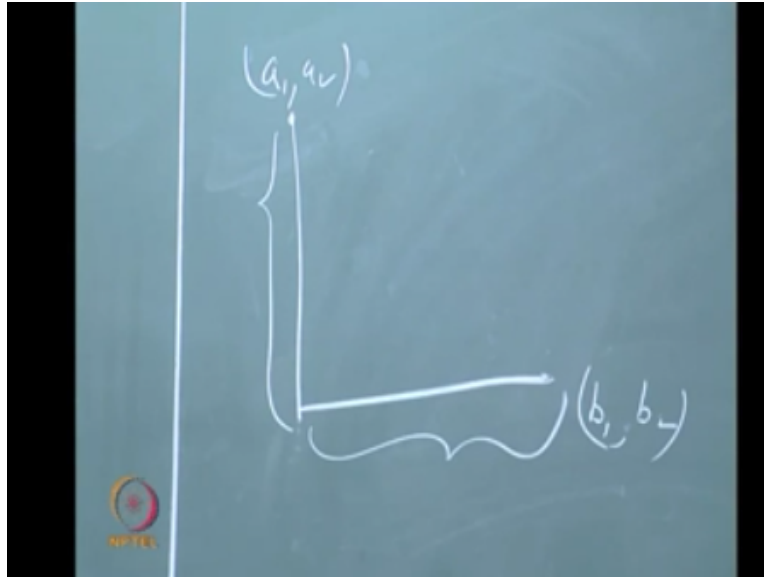
$$\underline{b}^{'} = (b_1, b_2, \ldots, b_M)$$

$$d_p(a,b) = \left(\sum_{j=1}^{M} |a_i - b_i|^p \right)^{1/p} ; p \geq 1.$$

$$p = 2 \longrightarrow \text{Euclidean distance}$$

And if you are in M dimensional space these are two vectors then for every p to take some value for P let us say P = 2 and you find this that you are going to get basically Euclidean distance if you take P is = 1 what are you going to get if you take small p is equal to 1what you will get is say this is a point.

(Refer Slide Time: 38:53)

A 1A two say this is a point B 1 B to say in two dimensions and if you take P = 1 what you are going to get is this Plus this right is a standard name for it what is it city block distance and these are all metrics if you take P = 1 you will get a metric if you take P = 2 you will get a metric if you take P to be a fractional value any fractional value greater than one then also you will get a metric any fractional value and any integer value greater than one you will still get a metric p is equal to 3 4 5 6 p is root 2 even geometric root 5 p = e metric p =P there is an irrational number even then you will get a metric so these are uncountable many metrics unaccountably many metrics even then you may define you are in a position to define.

Some more such metrics it is not the case that I have exhausted all matrix here so metric is a are this is a measure of dissimilarity between two points these are all measures of dissimilarity here this D this is a measure of dissimilarity distance is a measure of dissimilarity okay so there are very many metrics now the question is given a data set you need to choose which particular metric you would like to use given a data set you need to somehow find are you need to somehow calculate are in some way or other you need to get which particular metric you need to choose now.

I will give you some examples where probably for the same data set you may you may want to use two different matrix this example let us just say there are two students there are five subjects there are five subjects there are two students I will give you the marks of these two students.

(Refer Slide Time: 42:14)
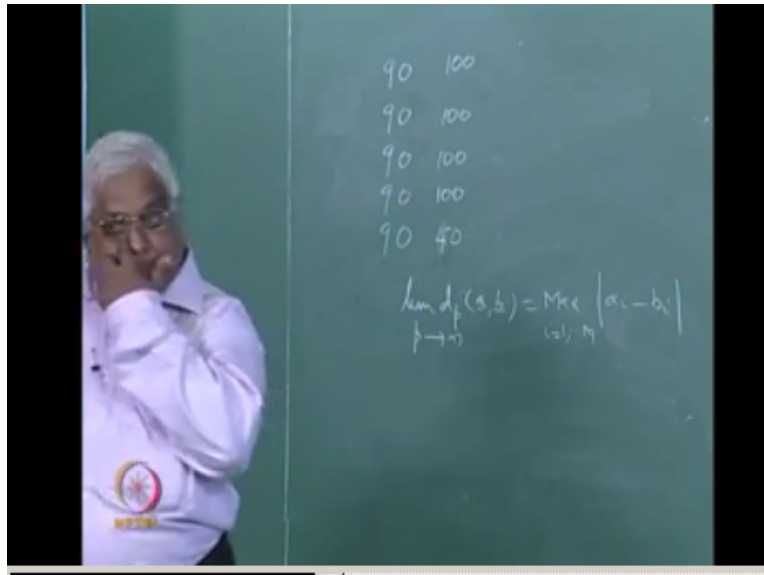
The first student he got 90 ,90, 90 ,90, 90 in five subjects okay the next student he got 100 ,100 100 ,50 and you would like to say who is better that is what is actually your problem 90,90,90,90 five subjects in the same five subjects other student has got 100 ,100 100 50 and you would like to say your problem is somehow you need to judge who is better and what you will do is that you would like to calculate the distance of this thing from the ideal one ideally is 100.100,100,100,100 whichever distance is minimum you would call that student to be better okay now my question to you is what distance measure.

You would like to use it in no start city block okay I mean let me just make it 60 because city block means then you are going to get to be same okay so let me just make it 60 then if you use city block probably he will be better okay because here it is 100 -90 10 10+10 +10 +10 50 you will get here you would get 40 so he is better but if you take P = 2 that is Euclidean distance then here it is 10 square 100 ,100, 100 ,100 ,100 ,500 and here this 40 square 1600 okay if you use Euclidean distance he is better now.
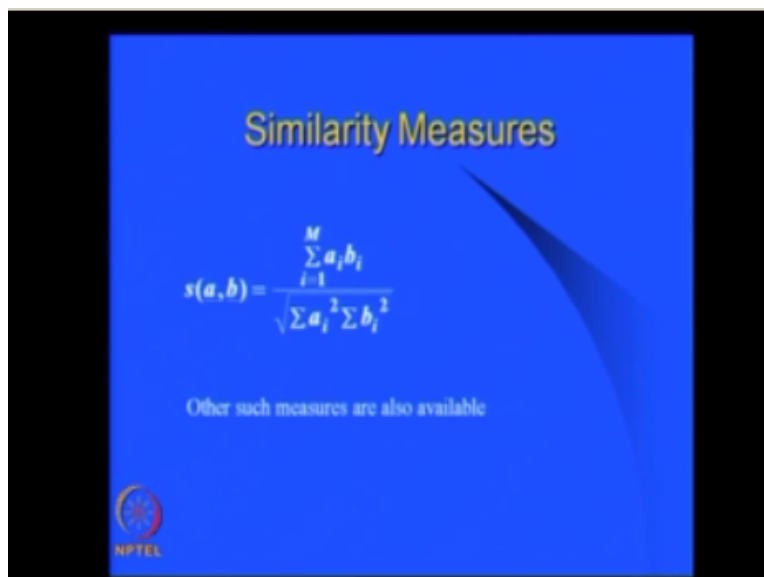
Which distance you would like to use it depends on whom you want to call better okay we would like to call this one to be better because it is 460 is 450 450is less than 4 50 60 but supposing this is 50 okay let me just take 40 then we can you think of some way of calling this one to be better suppose this is 40he just passed this subject is there any chance to call him better second one probably in our usual subjects mathematics physics chemistry algorithms data structures networking that one this one probably we do not we do not want to call the second student to be

better we would like to call probably the first student to be better but supposing you are five subjects are learning five different Carnatic ragas cannot music ragas you know that mastering a raga is extremely difficult and this person has mastered for ragas then probably.

You would call him to be a better candidate because he has a potential of mastering this one whereas this person he could not master any one of them maybe if it is a subject of about learning Carnatic music or Hindustani music learning those ragas probably you would want to put him to be a better student than this your choice of distance measure it depends on the problem at hand and how you would say someone is close to some or someone is far away on the other one it is very much problem dependent is it understandable to you shall.

We take a small break let us take us you yeah so the choice of distance measure it depends on the problem at hand and it depends on what the problem tells us common the problem actually forces us to calculate the distance between the quantities in that way and since you are trying to model the problem at hand you need to define your dissimilarity a similarity measures according to the situation are according to the constraints that you are facing so in the previous slide I was mentioning about dissimilarity measures and in this slide.

(Refer Slide Time: 47:52)

I am mentioning about similarity measure similarity between two vectors here is summation I Pm is equal to 1 to m these two are M dimensional vectors a Ba divided by square root summation Y square summation bi square can you tell me what the similarity measure is this is basically the angle between the two vectors a1 a 2 a.m. is one vector B 1B 2 BM is another vector how do you calculate cosine t θ this is the formula for it in fact your expression for correlation it came from this cosine θ has been existing in mathematics for a long time statisticians used it to get the expression for correlation coefficient.

That is it and with that they said that it lies between - one and one that is why this is cosine θ it lies between - one and one okay does it lie between -1 and 1 cosine values they lie between - 1 and 1 so this is similarity measure if the value θ is 0 cos 0 is 1 yeah maximum similar maximum similarities and when you are going to have θ to be π by2 then you will have get cos π by 2 to be 0 then they are perpendicular and one is on one side of one vector is like this another vector is completely the opposite that is 180 degrees then you are going to get - 1 then you will get - 1 so there you have the maximum dissimilarity that is the basic feeling.

When you are using this similarity measure this is also just an example there are many other examples of similarity measures in the literature and there are many other examples of dissimilarity measures also note that here in this slide I gave these metrics are all known as LP norms for P = 1 that is calledl1 not p = 2 that is l2 not and by the way can you tell me the value of this as P goes to ∞ what is the value of this as P goes to ∞ this one as P goes to ∞ what is the value of this is called L ∞ norm and this is also a metric this is called L ∞max modulus a - bi = 12 m now if you look at one of my previous lectures where I multiplied the differences by some meant by mattresses.

I called one of them to be positive definite matrix you attended those lectures positive definite matrix you may try to do I mean that is one way of looking at different I mean distances way you keep x 1 - y 1 whole square + x 2 - y 2 whole square root you keep it like this there you start adding first some weights to each of these terms then ultimately you will write a matrix there also you are generating many matrix here also you are generating matrix Mahalanobis distance which mall omnibus defined way back in nineteen thirties he wrote that paper to Calcutta mathematical society calculi mathematical society there that paper was published maulana miss distance paper that is I am not going to talk much because.

He will be decreasing all these things Mawlana this distance okay, there also you are going to get 1x I -y II mean X - y prime there is $\sum$ inverse x- y r mu 1 -mu 2prime $\sum$ inverse mu 1 -mu 2 which is again a quadratic form which is again a quadratic form and that $\sum$ inverse let us a positive definite matrix so that is why he called it as Δ square that is why he called it as Δ square so that we can talk about square root of that if that $\sum$ inverse is identity matrix you will get the usual distance the square of the usual distance that is why he divided by Δ square there is a square added to it so you can go on generating matrix indifferent ways depending on the problem at hand you.

End of Module 02-Lecture 05

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
Selvam
Sridharan

Studio Assistants
Linuselvan
Krishankumar
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
Santhosh 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