

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

Pattern Recognition

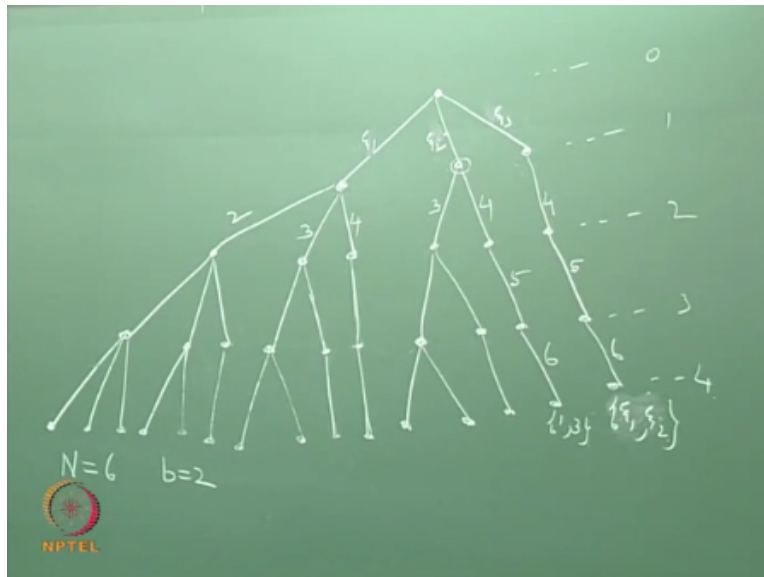
Module 04

Lecture 03

**Feature Selection:
Sequential Forward and
Backward Selection**

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

(Refer Slide Time: 00:20)

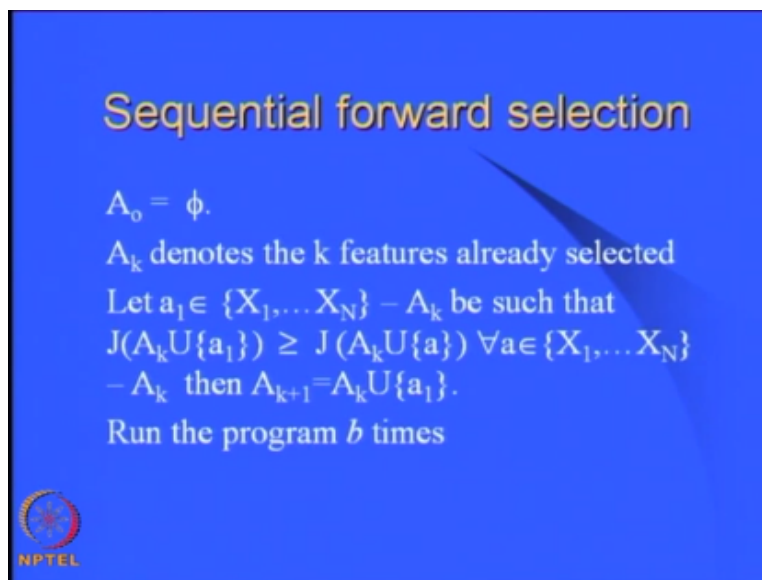


So you will find this one as a solution tree for branch-and-bound algorithm for $N = 6$ and $v = 2$ then one can give questions of the sort like draw the solution tree for branch and bound algorithm when $N = 7$, $v = 2$, $N = 7$, $v = 3$ are you can do some other numbers also so this can be given as class exercises okay now one can also give an exercise like write a C program for finding optimal set of features for branch and bound algorithm and you can give a criterion

function which satisfies the main assumption of the branch and bound algorithm and one can give some such criterion function and one can ask students to write a C program on implementing give a C program C code for implementing branch and bound algorithm.

So here branch and bound algorithm provides you optimal feature subsets for a special type of criterion function so but you have there are some other general techniques which do not use the properties of the criterion function and there are some famous algorithms famous methods they are sequential forward search sequential backward search generalized sequential forward search generalized sequential backward search you have something like what is called an LR algorithm.

(Refer Slide Time: 03:04)



Sequential forward selection

$A_0 = \phi.$


A_k denotes the k features already selected

Let $a_1 \in \{X_1, \dots, X_N\} - A_k$ be such that

$$J(A_k \cup \{a_1\}) \geq J(A_k \cup \{a\}) \quad \forall a \in \{X_1, \dots, X_N\} - A_k$$

then $A_{k+1} = A_k \cup \{a_1\}.$

Run the program b times



So these are some of the famous techniques for feature selection I will describe one or two of them here say this is the sequential forward selection method initially we start with a null set now A_k it denotes the key features already selected then what we will do is that we will select the $k + 1^{\text{th}}$ feature which I have denoted by A_1 this is selected from $X_1, X_3, X_n - A_k$ complement it is selected in such a way that if you include that one with a K_A the value of J is increased the maximum.

Then you write a $K + 1$ as a $KU a_1$ now run the loop not the program run the loop b number of times run the loop b times that means initially you are taking a null set now you have got capital number of features now which feature has the maximum value of J you will just find it and that you include it with a_0 a_0 is already null set so a_1 will have just one feature now the inclusion of

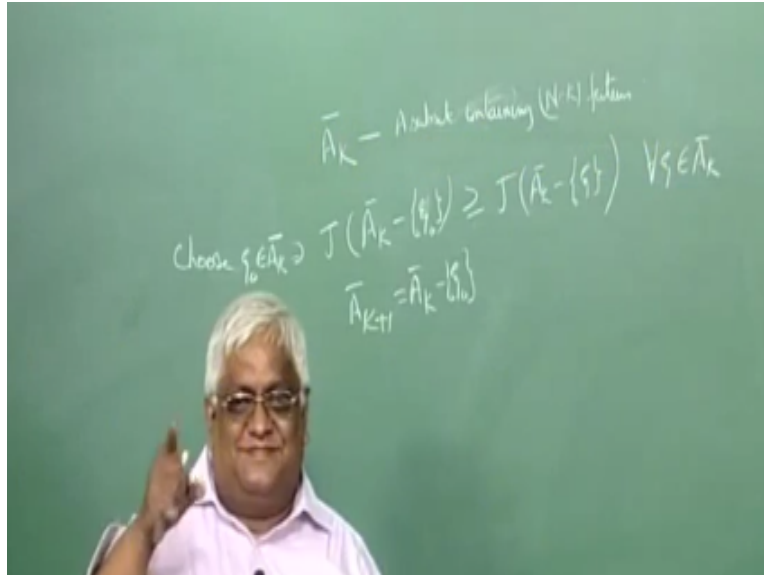
which feature to even will make that particular set has maximum J value that will include it then you will get a2 like that a3, a4, a5 up to a suffix b.

This is known as sequential forward selection algorithm probably the word the phrases sequential forward selection I suppose they are clear to you why the word for word is used each time we are adding features in that way the word for word is used sequential we are adding one feature at a time so that is why sequential forward selection okay and so since the word forward is used for addition of features we also have what is known as a sequential backward selection where you are going to subtract features.

No you are starting with a0 your question is the features that the feature said that we obtain is it really the optimal feature subset the answer is I cannot guarantee it this method does not guarantee that you will get optimal feature subset it will give you a feature subset but it does not guarantee to provide you the optimal feature subset have you understood because if first you selected a feature we selected a feature a1 some feature is selected that feature is going to be there throughout.

It may so happen that the optimal feature subset may not have that particular feature once a feature is included it will remain there whereas in the optimal one that may not be there if you apply the sequential forward selection method to the morning example we will select feature x1 and that feature x1 will be there throughout it will you will never get that X3 and X 4 in that way I mean if you apply this method yeah any other question it starts with the a0 yeah so we are using the word forward for addition of the features.

So there is also something called a sequential backward selection algorithm where you are always going to subtract features.
(Refer Slide Time: 08:34)



So how is it done you write A_k bar let us just say it denotes a subset of features subset of $N - K$ features a subset containing $N - K$ features let us say we already have an A_k bar then what we will do is that from A_k bar we will choose a feature X_i such that we will choose a features ξ_0 such that choose ξ_0 belonging to A_k bar such that J of A_k bar - $\xi_0 \geq J$ of A_k bar - ξ for all ξ belonging to A_k bar now you write A_{k+1} bar as A_k bar - ξ_0 .

Then you start with N features and you go up to b you start with N features and you go up to b that means initially you start with A_0 bar subset containing $N - 0$ features is N features which is A_0 bar is nothing but the whole space S then you start removing one by one features is this correct you will remove ξ_0 as such that J of A_k bar - $\xi_0 \geq J$ of A_k bar - ξ for all ξ belonging to A_k bar is this correct will it be \geq to or \leq .

We are we are doing the maximization it will be \geq to their also it is \geq to here also it is \geq to you need to convince yourself that it is \geq to you will remove that feature which has minimum information which has minimum information means removal of that feature does not reduce the value of J that much that means but officials have reduced the value of J a lot removal of ξ_0 has not reduced it that much so you remove ξ_0 have you understood it.

So here also it is \geq to so this is sequential forward search and sequential backward search between these two methods if I ask you which one you will choose and why what will be your answer let us just say $b < N/2$ this is what you are trying to ask me I am telling you if $b < N/2$

which one you will choose and why so you will use the backward one you will be use you will do more and more computations right.

In this case you will use forward number of computations maybe less but the accuracy of which one will be more from the point of view of accuracy which one you will find it to be better computationally you think forward one has less number of fun you need to do less number of computations than the backward one okay if $b < N/2$ right but then the accuracy wise which one you may find it to be better anyway think about it think about it probably.

There is no generalized answer for this regarding the accuracy but many times the accuracy of the backward one is slightly better than the accuracy of the forward one because in the backward one from the existing features which is generally a very big number from there you are reducing it so you are trying to take the interdependencies more into consideration than in the forward case so I am not claiming that the accuracy for backward selection will always be better than the forward selection I do not want to make this statement.

But there you are trying to take care of the interdependencies more that is why you are doing more computations so you are trying to take care of the interdependencies more so many times the accuracy of the backward selection that means accuracy from the point of view of J there that is many times it is found to be better than the one that you're getting by using the forward selection many times yes but not always.

So the problem is that the first feature that you are going to select the value of J is maximum for that feature but it may so happen that if you take the combination of few features you may get much better value of J and that much better value probably you may never get by using when this particular feature is present the first feature is present so I mean yes many times what you said that is going to happen but it still does not lead you always to the optimal feature subset it gives you some feature subset which is sometimes probably is optimal but not always optimal.

But your comment I think is valid many very many times it is valid very many times not much I can say so when our aim is to remove duplicate features and get unique subset of each our subset of features that contains the unique features then should we or should not be use this method first similar thing will happen even if you use backwards selection also in backward selection also the

features that are removed they do not contain much information after they are not expected to contain much information.

So similar thing will happen even in backward selection and if some such thing is present you can surely use forward selection I am not saying that you should not use it if I say that you should not use it I can I must be in a position to give you a better algorithm which I am not so I do not have any way of saying that you should not use it but what I am trying to say is that it will always give you the run the redundant features will always be removed if you want to make some such statement that I am not prepared to make.

But I can never tell you not to use an algorithm not to use FAC forward selection and back to backward selection that I cannot tell you are not exactly satisfied no so that is what my thing though I am just thinking like the way we do for data selection we generally cluster it take that cluster mean then we can say that these are the instances that can represent the whole thing can we do something for a case of features mastering of features yes it has been done I will come to it I think the time will permit at some point of time I will surely discuss clustering of features.

It has been done again it has been done the results are many times satisfactory but always they can be improved the main problem with feature selection methodology is that there is nothing called a best feature selection methodology method so that means whatever algorithm that are that are there they are they have some sort of negative points, so you can always improve upon them so when I say that you can always improve upon them as a researcher probably one may feel happy that there are some ways to improve there is so much of literature to be done but as a person working in the field if you everything is to be improved then what is the state of the art are you understanding.

So from another point of view it is not exactly good but that is unfortunately the state of affairs and feature selection there are too many methods are to be improved and there is a main drawback about the algorithms which I have already stated so there is always a scope for improvement and that is the state of affairs one more question please like I am again taking the analogy of clustering like in clustering the number of clusters it is so much important right so here less attention has been given on the optimal number of features to be selected like we assume that we will select some in a b number of features in the final subset but it may not be the optimal feature set in many cases like what it should vary with the data.

It should vary with the data why so less attention has been given in this are alike to find optimal number of features in the final feature set when do you call the value of b to be optimally so like in the IDs data set we I think the last two features are the best and it has got four features and I think the two features as far as I remember I do not know remember correctly I think two features are good and the two features are not that good something like that in IDs data set IDs data set.

So like in that I is data set if you have to select three features that will give poor result rather than to select two features is it the case with every data set yes that is that they know it is not the case with table data set okay every data set in IDs you are somehow able to say that to the number of figures to be selected is two and two is in some sense optimal can you say like that with a will data set probably because we do not know how many features are good and how many noisy features are there.

No probably yes that is true but on the other hand even if you are given the complete data set like in your you say archive there are too many data sets available okay and you can do your you can apply out feature selection methods there so there the information is given to you can you say for each data set or at least for some data sets can you say what is the optimal number of figures the word optimal number of features it is to be defined the phrase optimal number of features that is to be defined.

And that definition you would like it to be universal so that you can apply a will data set and then say that for each data set this is your optimal future optimal number of features to be selected so on one hand you would like to define the optimality of the number of features and second hand you would like to do it in a in a universal way and in my opinion it has not been done and probably in my opinion it cannot be done.

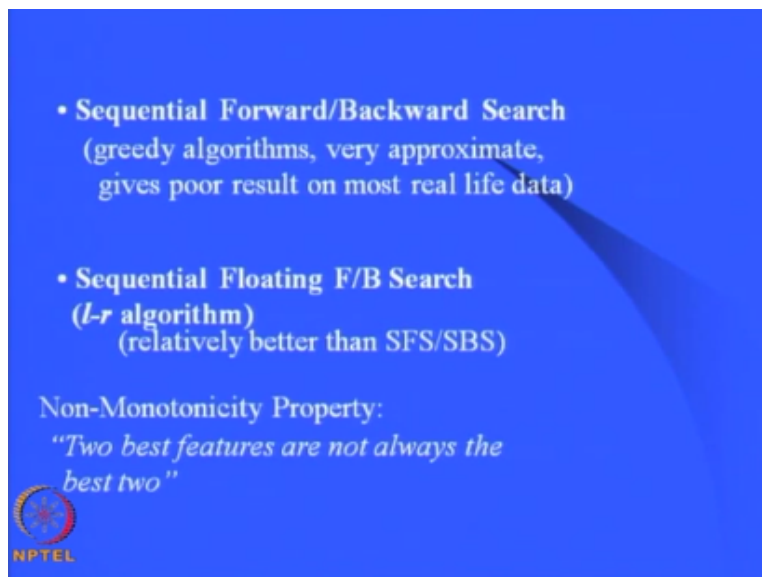
Because if you want to define the word optimal number of features in one way probably you can give it I can give a another definition about optimal number of features in some other way it is like saying that which clustering is better can you really say it and I give examples also where for the same data set different clustering have equal meaning maybe you I am sure you one must be able to get example even further number for the feature selection also where for the same data set

maybe you will get two or three different numbers of features which are sort of they had they might have the same importance value.

Maybe for let us just say for two number of features you may get x_1, x_2 has one set of features which will have the same value as let me just say X_3 and X_4 and it is better than all the other pairs so for if you have to select two features and x_1, x_2 is 1 and X_2, X_4 is 1 similarly for 3 figures also you might get some combinations I mean that uniqueness probably you may not have about the number of features as well as for the number when one you fix the number the set of features which particular set.


So I think the same way as we cannot say which clustering may be better in some examples so here also I am sure there are many examples where the which I mean the number of features optimality of the number of features one I do not think even your position to define it but this is my personal feeling and that can always be criticized that can always be criticized okay any other question you have any question.

(Refer Slide Time: 25:45)



- **Sequential Forward/Backward Search**
(greedy algorithms, very approximate,
gives poor result on most real life data)
- **Sequential Floating F/B Search**
(*l-r* algorithm)
(relatively better than SFS/SBS)

Non-Monotonicity Property:
"Two best features are not always the best two"

 NPTEL

The non-monotonic city property two best features are not obvious the best to this was what I was trying to tell you so we discussed a sequential forward selection and sequential backward selection there are some generalized algorithms also in this regard they are called as generalized sequential forward selection and generalized sequential backward selection the word generalized

is used from the point of view of the number of features that are getting added are getting subtracted.

Note that in the usual sequential forward selection we are adding at a time one feature and in the usual sequential backward selection we are deleting one feature at a time but in the generalized sequential forward selection we are going to add R number of features at a time and in the generalized sequential backward selection we are going to remove R number of features at a time.

So one can have algorithms also using the parameter value R and a slightly more complicated method is what is known as the LR algorithm what is known as LR algorithm since the analogy between clustering and feature selection has already been brought by one of the students in clustering we have algorithms where you split the clusters you are going on fitting the clusters and you are also going on merging the clusters.

Merging the clusters you have something like forward selection splitting the clusters you are something like backwards election now in clustering you also have split and merge techniques so similarly here there is an algorithm called LR algorithm wave in one iteration you will add L number of features and you will remove R number of features you will add L number of features and you will remove R number of features okay.

Now this L and R they have to be chosen in such a way that ultimately you will end up with the required number of features that is b you should choose L and R in such a way that ultimately you will end up with the required number of features b now there is a starting point for this LR algorithm the starting point is if you are starting with a null set then you should first add L features then remove R features that means R is less than L if you are starting with a null set then you add L features remove all R features again add L features remove R add L remove R .

So that when you are going on doing it ultimately at some point of time you will get b number of features so you have to choose the corresponding L and R now if the starting point is not the null set if it is the entire set then first you remove R features then you add L features remove R add L so then that means here R must $>$ than L because ultimately it should decrease here $R > L$ so again R and L I have to be selected in such a way that you end up with b number of features.

So that is basically LR algorithm LR algorithm is sort of more generalized than secret generalize the sequential forward search are generalized the sequential backward search and generalize the sequential forward search and backward search algorithms are more generalized than their original counter parts sequential forward sequential backward search algorithms you might be wondering why people are making it more and more complicated the reason is that somehow you should try to get the interdependencies taken into account as much as possible.

By adding at a time R number of features in your generalized sequential forward search somehow you are trying to take care of the dependencies in the features when you are adding odd number of features similarly when we are deleting R number of features in the generalized sequential backward step there also you are trying to take care of the interdependences and they allow algorithm which is the most complicated among all these five of them all these methods there at a time you are removing L features no you are removed you are adding L features and RR features you are adding L features and removing R features there also you are somehow trying to take care of interdependencies among the features as much as possible.

For all these methods there is a nice book divider and the kittler pattern recognition develop it lets developers look on pattern resolution there it contains that book contains six chapters on feature selection and extraction six chapters so all these methods are very nicely discussed there naturally you can see that the these methods even though they have been in existence for a long time they are people are still using many of these techniques the reason is that they have not been able to find really better methods than this.

Yes sometimes you have found some better methods in some applications that is quite true but these methods they are still relevant because they are people are still using them in their fields and they are still getting good results using these methods if you think that some method is really useless then somehow one must be able to say that some other method works always better than this and that you must be able to show it by using several examples so in that way people have not been able to show that much I mean they have not been able to show that much bad things I mean those many negative remarks about these generalized sequential and sequential forward sequential backward generalized sequential forward generalized sequential backward.

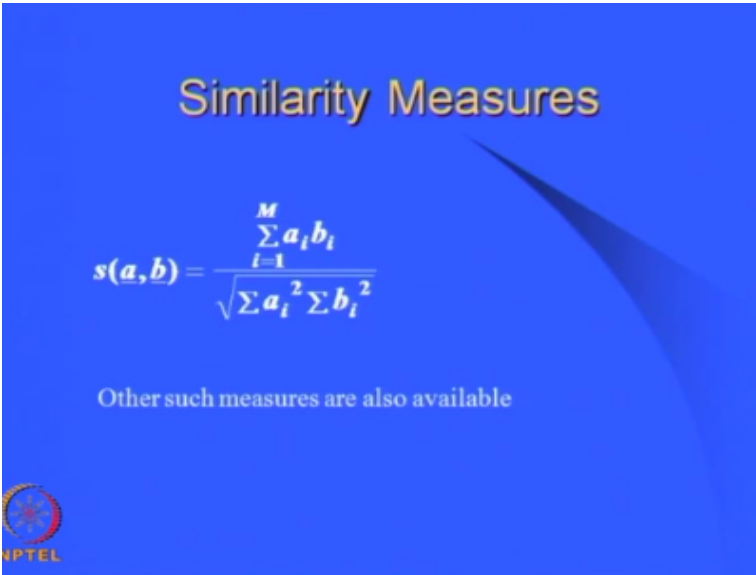
And LR methods they have been able to show some things regarding branch-and-bound algorithm because of that particular assumption about the criterion function and those many links

those many that tree may be really difficult to be implemented in I mean if the number of dimensions N is very large so people have been able to show some negative things about the branch-and-bound algorithm but not that much about the other five methods that I have stated in the last one hour one hour fifteen minute.

She was asking me about feature clustering about feature clustering some methods are there if you want to do something like feature clustering then let us go through whatever we have done about clustering where you have some similarities between points so when we are clustering points we have used a certain similarity or dissimilarity measures between points and we have taken those dissimilarity or similarities in our clustering algorithm so now for features also somehow you need to get the similarity between two features are that dissimilarity between two features.

How do you get it once you get this similarity or dissimilarity between features then you can sort of apply an algorithm similar to a clustering method and you can do the clustering of features right if you can get something like a similarity between two features are it dissimilarity between two features then you can do clustering of features then how do you get a dissimilarity or similarity between two features.

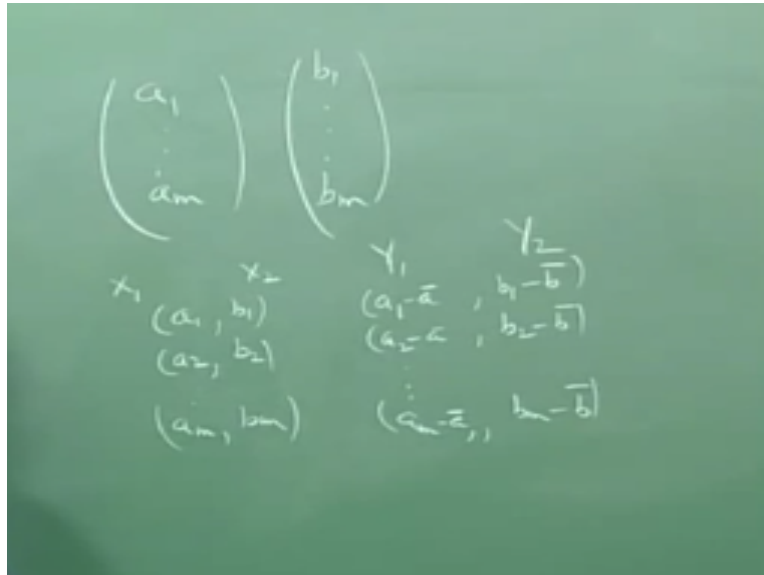
(Refer Slide Time: 36:28)



The slide has a blue background with the title "Similarity Measures" in yellow text at the top. Below the title is the cosine similarity formula:
$$s(a, b) = \frac{\sum_{i=1}^M a_i b_i}{\sqrt{\sum a_i^2 \sum b_i^2}}$$
 Underneath the formula, it says "Other such measures are also available" in white text. In the bottom left corner, there is a small circular logo with a star and the text "NPTEL" below it.

If you have two vectors a_1, a_2, \dots, a_n and b_1, b_2, \dots, b_m this is a similarity measure between two vectors I have shown this slide in one of my previous lectures.

(Refer Slide Time: 36:53)



So you have a vector a_1 to a_m . you have another vector b_1 to b_m then the similarity between these two vectors this is a way of defining similarity it is not the only way and it provides the cosine of the angle between those two vectors right now suppose feature x_1 and feature x_2 , x_1 takes when x_1 takes the value a_1 , x_2 takes the value b_1 when x_1 takes the value a_2 , x_2 takes the value b_2 when x_1 takes the value a_m , x_2 takes the value b_m then the similarity between these two features you can measure it by the same thing.

Is it clear to you or shall I tell you once again I will repeat it say there are two features x_1 and x_2 when x_1 is taking the value a_1 x_2 is taking the value b_1 when x_1 to a_2 to x_2 to b_2 when x_1 takes a_m then x_2 is taking b_m then you would like to find the similarity between x_1 and x_2 then you can use this please so recently one function from a_1 to a_m it is like for one person height is a_1 weight is b_1 for another person height is a_2 weight is b_2 for the a_m^{th} person height is a_m weight is b_m .

That is a relevant question there is another relevant question that is the other relevant question is varies correlation coefficient coming into the picture varies correlation coefficient coming into the picture if I write here I write vector y_1 as this is $a_1 - \bar{a}$, \bar{a} is the mean of a_i this is $a_2 - \bar{a}$ this is $a_m - \bar{a}$.

Then feature y_2 as this is $b_1 - \bar{b}$ this is $b_2 - \bar{b}$ and this is $b_m - \bar{b}$ then the similarity between y_1 and y_2 if you use that formula then you will actually get correlation coefficient please but again we are getting a linear correlation coefficient its we are just looking at the linear relationship between a_1 and b_1 like the correlation coefficient which you mentioned the last class.

I mentioned correlation coefficient here because I wanted you to understand the difference between this formulation and this formulation here in this formulation if you remove a bar then you will get this here in this one if you remove \bar{b} then you are going to get this and this is going to give you the correlation coefficient so just the removal of the means will give you the correlation coefficient and if you do not remove those means it is just simply cosine of the angle between these two vectors.

So I mean why I mentioned this one the reason is that you need to really understand that there is a relationship between this and this that is what I want you to understand I just removed the means right and here you are getting the correlation coefficient so if you do not remove the means it is just that cosine of the angle between those two vectors so this is a way of dealing with this but this is not the only way there are other ways in which one can define similarity or dissimilarity between two features.

Dr. Sukhendu das will teach you at some point of time principle components so after he teaches the principal components I will talk about similarity between two features similarity between two figures looking at somehow the covariance matrices I have to also tell you a few things since we have been talking about correlation coefficient you are talking about linearity between linear correlation coefficient that is some of the things that you wanted to say the reason why you use the word linear is that somehow whenever linear relationship is existing between the variables you are somehow able to get it nicely by using correlation coefficient whereas when the relationship is not linear you are not able to get the relationship value properly using the correlation coefficient.

That is why you would like to call it linear though I do not like the word linear used in that way though I understand why you are using it I would like to tell you some more things regarding correlation coefficient suppose you have a data set two dimensional data set two variables x and

y have found the correlation coefficient between them and what you do is that you rotate the whole data by some angle θ so you will get new $x_1, y_1, x_2, y_2, \dots, x_n, y_n$ and you are understanding.

Now you try to find the correlation coefficient between the new variables x and y do you think they will be same as the previous one or do you think there will be a difference the answer is they are not necessarily same they are not necessarily same if you rotate the whole data correlation coefficient values need not necessarily remain the same okay that is one problem there are also other issues regarding correlation coefficient but then how do you find one particular attribute or one particular way in which something that will remain invariant even if you rotate the data by any angle whatsoever do we have something of that sort my answer is yes.

There are some things there those things they would not change even if you rotate the data by any angle whatsoever 0 to 359° or if you look at it in radians you uncountable many such angles in which you can do it you will get there is something that is invariant with respect to this rotation those things you will know when we deal with the principal components where they are basically you will find that they are rotation invariant principal components they are basically rotation invariant I think will stop here.

**End of
Module 04 – 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 Krishnamurthy & Team

Animations
Dvijavanthi

NPTEL Web & Faculty Assistance Team

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

Administrative Assistant
K.S. Janakiraman

Principal Project Officer
Usha Nagarajan

Video Producers

K.R. Ravindranath
Kannan Krishnamurthy

IIT Madras Production

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

www.nptel.ac.in

Copyrights Reserved