

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

Pattern Recognition

Module 04

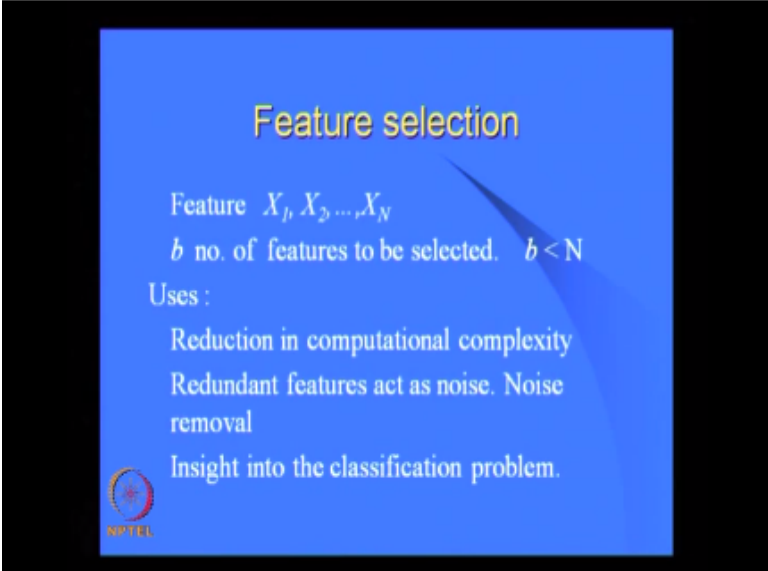
Lecture 01

**Feature Selection:
Problem statement and Uses**

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

Good morning I shall be talking about feature selection today, let me first basically tell you the problem of feature selection.

(Refer Slide Time: 00:18)




Feature selection

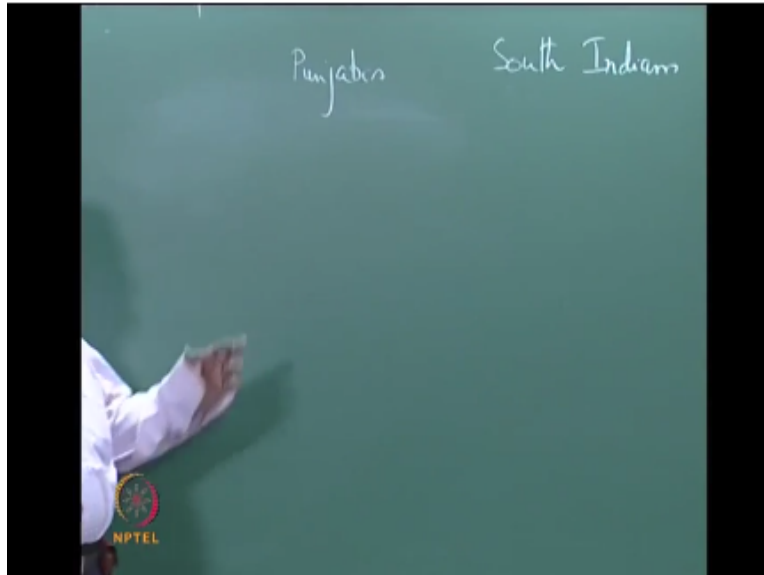
Feature X_1, X_2, \dots, X_N
 b no. of features to be selected. $b < N$

Uses :

- Reduction in computational complexity
- Redundant features act as noise. Noise removal
- Insight into the classification problem.

 NPTEL

In order to understand the problem of feature selection let me give you one or two examples in, my first example I am considering two communities of people.
(Refer Slide Time: 00:31)



One community is Punjabis another community is another community is South Indians, we generally know that Punjabis are taller than South Indians it is something that we have seen it many times with just too many persons. So one of the distinguishing features which separates these two communities is the height, it is something that we all of us know. Now the question is if you are given a data set consisting of Punjabis and South Indians and you have measured the features like height, weight and certain are some other features.

Height is measured say in centimeters weight is measured say in kgs etc maybe you have around 40, 50 such features values which are taken on everyone of these individuals that means let us just say you have a 40 dimensional feature vector. Now using this information how does one say that the is one of the distinguishing features between these two communities, which we know that it is one of the distinguishing features that data should automatically tell you, that height is a distinguishing feature.

How does the data will tell you naturally unit right an algorithm which produces the result that height is one of the distinguishing features, it distinguishes between South Indians and Punjabis there may be many other features I am not saying that height is the only feature there may be some other features also which may be distinguishing, these two communities but height is surely one of them. The feature selection method should result in automatically the feature height as the distinguishing features as a distinguishing feature between these two communities.

How does the feature selection method give height as a distinguishing feature that is the problem let me tell you another example, probably even this I think I assume that you might have noticed it if you look at people with mongoloid features, basically people living in the hilly regions of the northeastern states if you look at them, their noses they have shorter lengths compared to the noses of our people we people who are living in plates. Now the length of the nose is a distinguishing feature with that it distinguishes between the people who are living in the hilly regions of the north eastern states.

And the others like that if you look at here I give you examples of different communities they may not be different communities between I mean if there is a classification problem there might be some features which are able to distinguish the classes much better than the other features. So this is one of the problems that is there in feature selection now what are the other issues. Usually let us just say your number of features is capital and features X_1, X_2 and this is the mathematic formulation.

You are supposed to select B number of features where B is less than capital N many times beam the value of B is known to you, sometimes the value of B may not be known to you what is the meaning of value of B ? being known in the experiments under consideration the users generally have some constraints the constraints may be regarding the amount of computations that he may need to do, if the number of features is quite high and if you are looking at something like variance covariance matrices and their Eigen values are Eigen vectors are etc.

The number of computations maybe too large so in some situations that readers the user may say that I cannot have my covariance matrix ease of my covariance matter it cannot be more than the size, some such constraint the user may put in that may naturally give a constraint on the number of features. So I mean sometimes the value of B the number of switches to be selected is known from the point of view of the computational complexity. Generally it varies from problem domain to problem domain the value of B .

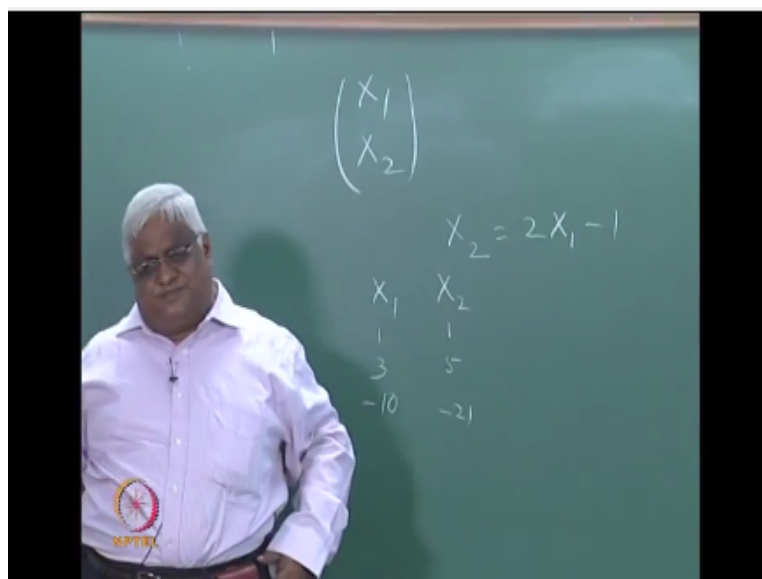
And many times it may not be known the user what may be the exact value of B that he would like to have he might sometimes say that the number of features that I would like to get is between these two limits anything between 10 to 20 is fine, some such thing also a reader may say and sometimes the reader the user may say that I do not I really do not know the number of

features that I would like to have. So B may not be known we may B partially known and B is known.

Here I have not considered in these slides that I am going to say I generally assumed that number of speakers the number of reduced features is known to the user, I generally assumed that the number of reduced features is known to the user any practice that may not be the case, usually real-life problems are more complicated than what is written in text books, it is almost always the case that real-life problems are much more complex than what we teach in the class.

So now the number of features to be selected is small b naturally be is $< N$ now why do we need to do feature selection, reduction in computational complexity this point I have already explained to you, now there is this one redundant features act as noise so we are doing something like noise removal. Now I use the word redundant what is the meaning of redundant? Let me explain.

(Refer Slide Time: 08:35)



Say I have just two features X_1 and X_2 but then let us just say my X_2 is $2X_1 - 1$ that is when X_1 takes the value 1 X_2 is taking the value 1 when X_1 takes the value 3 X_2 is taking the value 5 when X_1 is taking the value say -10 and this will be -20, -21, so there is a direct relationship between X_1 and X_2 in such a case do you really think that we should keep both the features, when there is a direct relationship existing like this the answer is no. So here since there is a naturally there is a direct relationship one of the features is redundant that is it is useless.

Whatever you can get from x_1 you can also get from x_2 because of this linear relationship, so we can remove one of the features please, I was expecting this question, when the relationship is nonlinear there are usually some issues involved. Let me tell you the issues let us let me first assume that there is a cubic relationship there is something like this that is this I just taken some polynomial of degree 3 I could have taken a polynomial of degree 2 because his question is nonlinear, nonlinear means it can be degree 2 degree 3 it may be a polynomial it may not even be a polynomial.

It can be some other relationship also, so I am breaking down his question into a few parts, my first part is I am taking a polynomial of degree three, the reason for taking, polynomial of degree three is, so if you take the value of if you take some values for x_1 you are going to get corresponding values for x_2 okay but note that here from x_1 you can go to x_2 from X_2 you can go to x_1 , x_2 is equal $2x_1 - 1$, whereas x_1 is $=x_2 + 1/2$ here x_2 is $=x_1^3 - 7$ then $x_1^2 + 5x_1 - 7$ can we express x_1 also in terms of x_2 like this?

Have you understood the point here you have both the relationship may not always have a both way relationship right, unique relationship and the problem exists with squares also instead of x_1 cube I could have written from square. So if the relationship is not there for example something like this I can say that x_2 is redundant but probably I cannot say the x_1 is redundant, probably I cannot say that x_1 is redundant, I can say that x_2 is redundant, whereas in this situation I can say one of them it does not matter.

X_2 is redundant or x_1 is redundant I can take any one of them but probably in this case I cannot say that. Now suppose you do not have a polynomial sort of relationship you might have something like a exponential relationship and you might have some other thing some other relationship, in such a case number one are we in a position to establish the relationship, even if

such a relationship exists. Suppose one feature x_2 is some 10 times 2^{-x_1} - say 3 times E^{x_1} + say some 4 times something like $\log X$, log to the base 2.

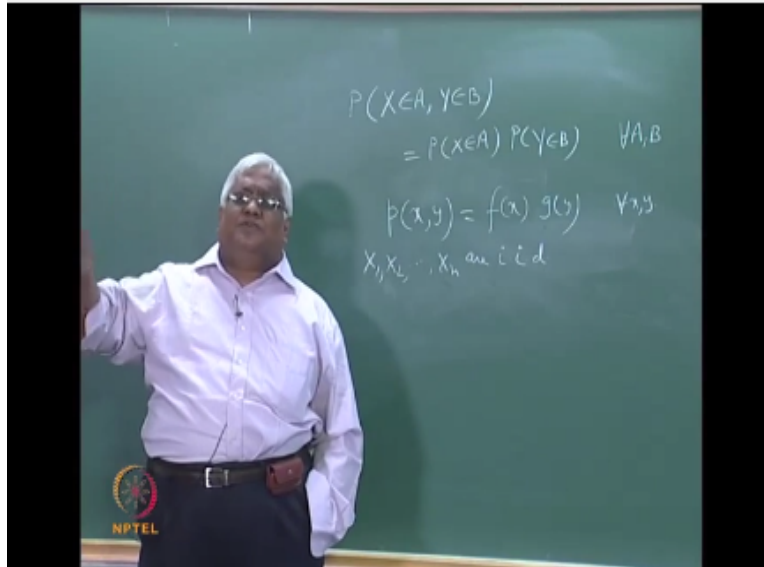
Let us say I have taken there is some such relationship actually it is existing now are we in a position to get this relationship, this is one of the problems one problem is that there may not if there is no relationship it is fine but if there is a relationship but we are not in a position to get it, now the second part let me just expand a little bit I hope many of you have had courses on numerical analysis okay. So you are given the values of say $x_1, y_1, x_2, y_2, \dots, x_n, y_n$ you can always fit a polynomial of size of degree $n - 1$ right.

If you are given $x_1, y_1, x_2, y_2, \dots, x_n, y_n$ we know that you can always fit a polynomial of degree $n - 1$ and usually in pattern recognition problems you always have finitely many values. So here we have come to a barrier on one hand if there are n observations you can always fit a polynomial of degree $n - 1$ but on the other hand, so probably there is no relationship between those two features.

The feature the relationship between features it has been a difficult problem and this problem has been faced by different research communities, at different times and in different situations. Let me talk a little bit about statistics, there is one concept called correlation coefficient, which tries to measure the relationship between two variables x and y , two random variables are there then correlation coefficient it tries to measure the relationship between two variables.

And then there is a theorem, the theorem has as the statement is like this x and y are two random variables, if x and y are independent random variables, then the correlation coefficient value is 0 there is a definition of the word independence, there is a definition for the word independent, so x and y are independent random variables means it is the definition says that.

(Refer Slide Time: 17:31)



Probability of X belonging to a Y belonging to B is = probability of X belonging to X times probability of Y belonging to B for all A B, then you call X and y to be independent random variables you call X and Y to be independent random variables, if probability of X belonging to a and y belonging to B this intersection of two events is = probability of X belonging to a times probability of Y belonging to B.

Probably we all have we all know the meaning of independence of events, the independence of events you say that C and D are independent, you say that C and D are independent if probability of C intersection D is = probability of C times probability of D, this was something that we have read long back. So this definition is generalized here for random variables, yes this is this is the independence of events from here this is generalized to independence of random variables.

Now he is talking about joint probabilities that is also true from here you can go to the whatever you can go to the point that got hit but he is trying to say, that is you say that two random variables x and y are independent if the joint probability density function, the joint probability density function if I write it as this is same as the product of the marginal density functions. The density function for capital X is say small F the density function for capital y is a small g and the joint probability density function for these random variables is a small p.

Then P of XY is =FX into Gy for all X Y this is also true but this is taken as the definition and consequence of this is this because if you have random variables you may not always have probability density functions, you might have discrete probability mass functions also okay. So

this is the basic definition and from here you will get this. Independent and identically distributed so IID independent and identically distributed we say that X_1, X_2, \dots, X_n are IID that is independent and identically distributed.

Independent means the definition of independence that I gave you that should be used for all these n random variables okay, identically distributed means the probability density function of X_1 is same as the probability density function of X_2 is same as the probability density function of X_n and it may not be density functions, if it is a discrete probability then probability mass function of X_1 is same as probability mass function of X_2 same as probability mass function of X_n okay.

So a simple example of IID independent identically distributed random variables is you take a coin and you toss it say n times, the result of the first trial you call it as X_1 secondary trial you call it as X_2 , N trial you call it as X_n okay. Then those X_1, X_2, \dots, X_n they are IID they have this first they are independent because the result of the first trial is does not have does not make any impact on the result of the second trial, so they are independent identically distributed since the coin is the same probability of head is same throughout in the first trial second trial third trial fourth trial etcetera.

And you have only two outcomes probability of a head or tail so if the probability of heads are saying the probability of tails are also same, so they have the same distribution identical because the random variable can take only two values head or tail head with probability small p if I say then tail with probability $1 - p$ and that is true for X_1 that is same for X_2 same for X_3 and same for X_n , so they have the same distribution and but they are independent, that is the meaning of independent and identically distributed random variables.

We are talking about little relationship between random variables, correlation coefficient, so if the random variables are independent then correlation coefficient value is 0 but it is not the converse is not true, that is the correlation coefficient can be 0 but the random variables are not dependent and the example that people give is this.

(Refer Slide Time: 24:20)

$Y \in \{0, 1\}$
 $Y = X^2$

X \ Y	-1	0	1	
0	0	$\frac{1}{2}$	0	$\frac{1}{2}$
1	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$
	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

NPTEL

I have X is taking values - 0 and 1 okay and Y takes values 0 and 1 in fact $Y = X^2$ okay $Y = X^2$ now let me just so X takes values -1 0 and 1, Y X values 0 and 1 these ones they are going to give you the probabilities, so this is $\frac{1}{2}$ this is $\frac{1}{4}$ is $\frac{1}{4}$ so probability of Y is =0 this is $\frac{1}{2}$ probability of Y is =1 is $\frac{1}{2}$, probability of X is =-1 is $\frac{1}{4}$ this is half this is $\frac{1}{4}$ what is 1, you can very easily see that the value of the correlation coefficient is 0 but the relationship is this.

So here the problem is that correlation coefficient is not the right one to measure the relationship okay, it is not the right one to measure the relationship, you may have so well what is the other one I mean which one is going to measure the relationship better this is a long-standing problem in statistics. We know the meaning of dependence of random variables and independence of random variables we know when the random variables are independent. The definition of dependency is the following x and y are said to be dependent random variables if they are not independent random variables.

X and y are said to be dependent if they are not independent but how much dependent they are that there is no measure, how much dependent they are there is no measure and this problem it has been there for along long time and the same problem persists even now and the same problem is there even in future selection, how do we know that these two features are related the word related I am using it loosely not from the point of view of correlation, are not from other points of it.

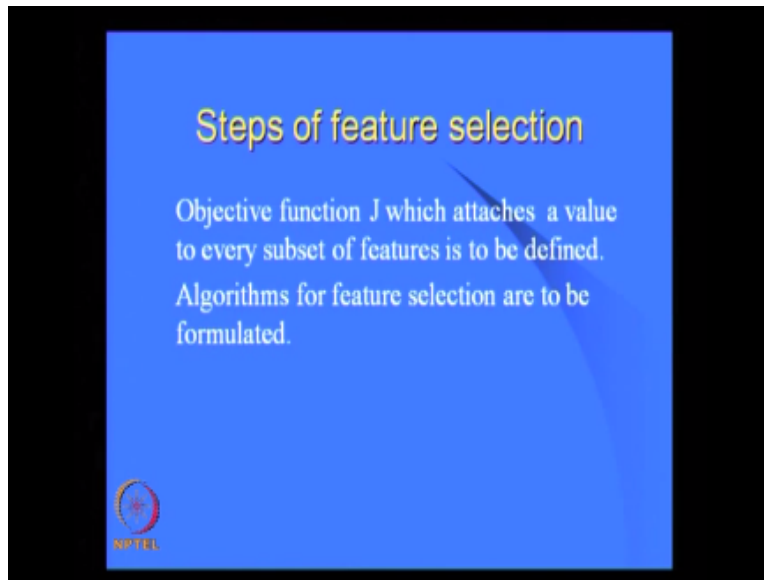
What is the our general meaning of the word relation there is some sort of relish some function or something which is from one variable if you use that particular function we can go to the other variable then how to get hold of that function, number one and number two is it should not be a function like when you have $x_1 x_2 \dots x_n$ and the corresponding $Y_1 Y_2 \dots Y_n$ you fit a polynomial of degree $n - 1$ say that then the error value will be 0 it is not a function of that sort.

So here the problem is not well-defined and I really do not know how to define the problem properly because this is a research issue, how to measure the relation between two features this is a research issue, there are already very many papers on this already quite a bit of literature in fact on pattern recognition is actually devoted to measuring the relationship between two features. The problem is even though relevant because it is still not solved to the satisfaction of all of us and there is still a lot of work that is needed to be done.

I think with this I will go to the next part of use of the feature selection, insight into the classification problem it will provide you insight into the classification problem, insight in the sense of if I want to classify Punjabis and South Indians, then my feature selection criterion are the algorithm and the algorithm should automatically provide height as one of the main features, that means from the data I am getting a conclusion that this feature is highly relevant for classification.

This sort of conclusions I would like to get from the data by using these methodologies, then all these methodologies then they are giving us an insight an in-depth idea about the classification problem which features are more important, which features are less important. So these are the users there are in fact one can think of a few more users there are a few more users also but these are the main three uses.

(Refer Slide Time: 30:45)



So what are the basic steps of feature selection the basic steps of feature selection are first initially you need to define an objective function which measures the importance of a collection of features, which somehow measures the importance of a collection of features that objective function is to be defined first? Then once you define the objective function the second part is we need to optimize it minimize it or maximization, depending on the type of optimization function it is for some functions you need to do minimization, for some other functions you need to do maximization.

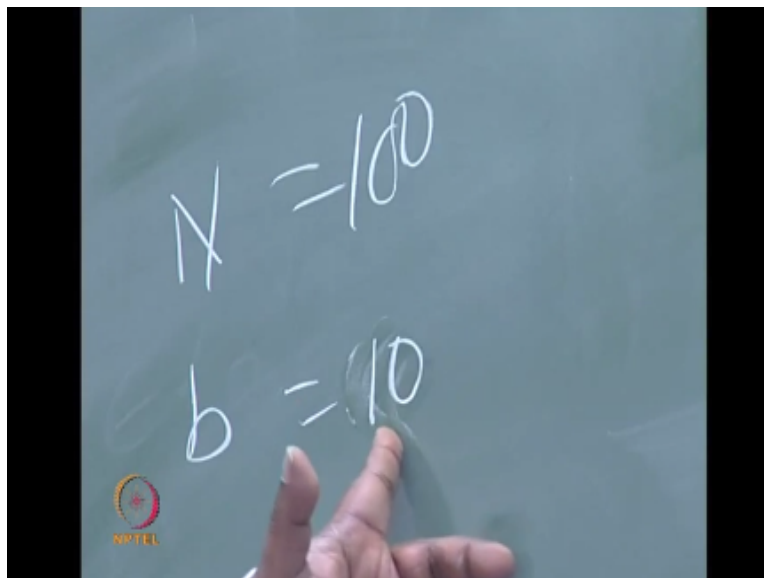
So basically there are two steps in featured selection, the first step is you need to define a function J an objective function on every collection of every subset of features this objective function provides you somehow the importance of that particular collection of figures. Now the second part is that once you define the objective function then you need to do the optimization maximization or minimization depending on the type of function. Now how does one do how does one define the optimization function or the criterion function J ?

As I said there are quite many papers on this there are very many papers on this, so for the present moment what I will do is that I will assume that I will discuss about how to define this criterion functions later but for the present moment what I will assume is that the criterion function is already given to us, then how does one do the maximization or minimization optimization of the criterion function and I shall assume that the number of features is number of

features that unique you want the thing to be reduced that is known that is the value of small B, I am assuming it to be known.

Actually you will understand why I am giving importance to this particular thing algorithm development.

(Refer Slide Time: 33:34)



Suppose your capital n value is 100 that means you have 100 features from this 100 features you would like to select 10 features, then how many possible such sets containing 10 elements can be found from capital N it is $100 C 10$ these many subsets of size 10 we can have from 100 features. Now what is the value of $100 C 10$ and you tell me approximately $> 10^{12}$ now my question to you is are we in a position to go through all these subsets to find the optimal subset what is your answer, the answer is no 10^{12} is a huge number after one you have to put 12 zeroes, that is a huge number.

So generally we do not like to search all these many subsets to get the optimal one, now you have the other side suppose I do not search the whole space can I guarantee that tile all I will get the optimal, have you understood the problem or shall I repeat it. On one hand we do not want to

search the whole space to get optimal so maybe out of 10^{12} let us just say I would like to search 10^9 this much then can I claim that within this 10^9 optimal will always be there?

We cannot say that, that means we have a very big problem here, so let me state the problem the problem is like this, there exists no feature selection algorithm which provides you optimal subset of features for any criterion function without doing exhaustive search. I am repeating it to this day there exists no feature selection algorithm which provides optimal feature subsets for any criterion function without doing exhaustive search, first I will expand on this thing then I will come to you okay.

Let me tell you the meaning of any criterion function that any criterion function means there are some criterion functions, they satisfy some properties there are some criterion functions they satisfy some properties. Now you can use those properties to obtain a feature selection algorithm which is not which will give you a optimal feature subset without doing the exhaustive search because the criterion function has some properties but when I said for any criterion function it is not necessarily true that every criterion function has some properties.

Some criterion function may not have any properties, are you understanding what I am trying to say the meaning of any criterion function that means you have a criterion function you do not know what criterion function is but given the set of features it will give you the value. Now you want to get your I mean optimal set of features, now you want to write an algorithm so you want to write this algorithm without using the properties of the criterion function, then without doing the exhaustive search you cannot guarantee optimality that is what my statement is.

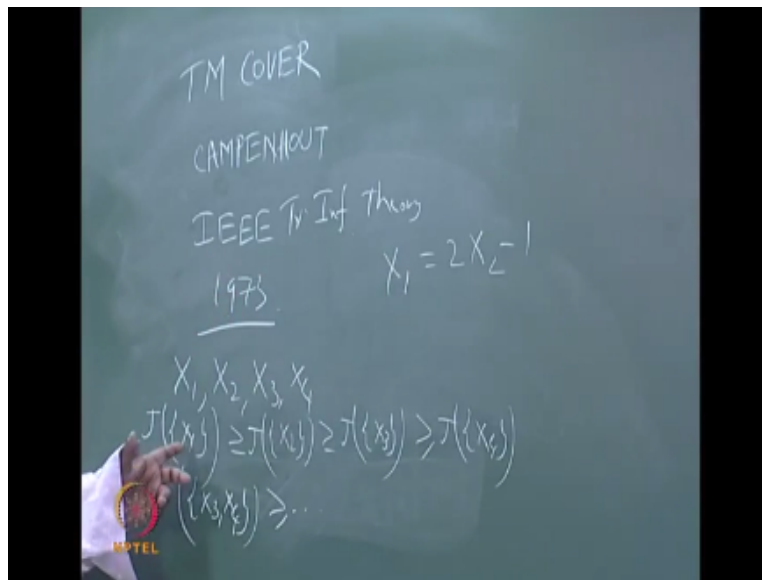
Whatever optimal algorithms that you are seeing optimal that means you are able to get optimal set of features without doing exhaustive such they are using the properties of the criterion function sure it is please, sure it is possible but what any means it is you should not use all those properties. So if you people are interested in you may try to develop algorithms which without doing exhaustive search you would like to get optimal features.

Anyway this I sort of mentioned it jokingly I do not know whether it is possible or not okay but to this day there is no such algorithm, so that means whatever algorithms that are existing they do not guarantee to provide optimal feature subsets for any criterion function. So yes you are not

guaranteed to get the optimal solutions though naturally you are going to get some sub optimal solutions.

Before I go into the algorithms I would like to make one more comment, this comment was actually made by TM Cover Campenhout.

(Refer Slide Time: 40:44)



This is an ice Ripley transactions on information theory, they wrote a paper the title of the paper is the two best features are not necessarily the best, to the title of the paper is that two best features are not necessarily the best two, there they just provided an example in that example they have taken one particular criterion function J and in that example they have taken four variables let us represent the four variables as X_1, X_2, X_3, X_4 okay and the criterion function let us just say we have to maximize it.

Then the criterion function is X_1 it has the maximal value, than X_2 it has the second maximal value than X_3 let us just say it has the third one and the least impressive is X_4 okay, now they are supposed to select two features then what do we expect? We expect that X_1 will be there in those two features but then the example is constructed in such a way that this is better than the rest $x_3 x_4$ as a pair it is better than any other pair 4 is six pairs, so this pair is better than the other five pairs.

That is why the title of the paper is that two best features are not necessarily the best two jointly yes this is a I think one and off page or two page paper if you go through, the general you will find it, usually in I triplet journals there are something called a short paper, something called a what is that regular people and for regular paper they publish the photographs of the authors. Now it is naturally it is a short paper but they publish the photographs of the authors because I think they found the paper to be really important because it just tells something that we all of us probably are feeling but then probably we are not able to put it quantitatively which they have done it.

And the function J is entropy based function the function J is entropy based function, you know you can go through the reference it is a very standard reference, that probably is not available that probably is not available to say I mean it is very old paper by the way do you people know anything about TM cover have you heard the name? In 1967 he wrote a paper with HART this HART is same as the on HART on nearest-neighbor decision rule cover and HART it is pronounced as COVER but the spelling is COVER see where we are as you can see.

He is an electrical and electronics engineer originally this is 1967 and 73 papers I am talking about we are in 2011 okay, he is even now active he is associated with I mean in Stanford University he is from Stanford University Stanford University as you know it is a I mean yeah it is a great university one of the greatest universities in the world Stanford MIT okay and he is associated both with the Statistics Department as well as the electrical engineering department.

You will find I mean he received several awards he received several awards he is one of the father figures of pattern recognition he worked quite a lot on feature selection, nowadays it is called as portfolio management it is one of the terminologies it is used portfolio management anyway it is one of the consequences of the feature selection problem that has been stated here and he was quite a lot on that portfolio management okay.

So now probably you have understood why feature selection is a difficult topic because of that particular example by TM cover and his student, so and this will also tell you why probably you need to do the exhaustive search this will also tell you why probably you need to do the exhaustive search I was expecting a question from you, anyway I will ask the question myself and I will give you the answer, how is this happening?

The reason is like this note that I was giving you some examples X 1 and X 2 is directly related to X1 linear relationship then if you put X1 suppose X 1 is directly related to X2 relationship then if you put them together there is no extra information whatever is there in X 1 the same thing is there in X 2, so if you put X 1 and X2 together you will get whatever is there in X 1 only that one only you are going to get you are not going to get anything extra, have you understood what I am trying to say.

Whereas if you put X3 and X 4 together that may be you will be getting more than what is there in X1, how you understood what I wanted to say have you understood suppose X 1 is a linear function of X 2 then if you put X1 and X 2 together it is same as just writing x 1 to x. So there is nothing no extra thing that you are going to get the J X 1 X 2 is going to be same as J of X 1 okay, whereas X 3 and X 4 individually they are probably they do not have much importance but when you put them together collectively they may give lot of importance.

So that may become the value may become more, I would like to put it like the next one is = something like 2x2 -one yes so the linear relationship existing if you put x1 and x2 together it is the same as just using X 1/2 times, so you are not going to get anything extra there x1 x2 will be same as J of x1 whereas if you put x3 and x4 together individually they are unimportant but then if you put them together probably the importance may be more, maybe in the next class already the time is over.

**End of
Module 04 – Lecture 02**

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