Indian Institute of Technology Madras

NPTEL
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING

Pattern Recognition

Module 04

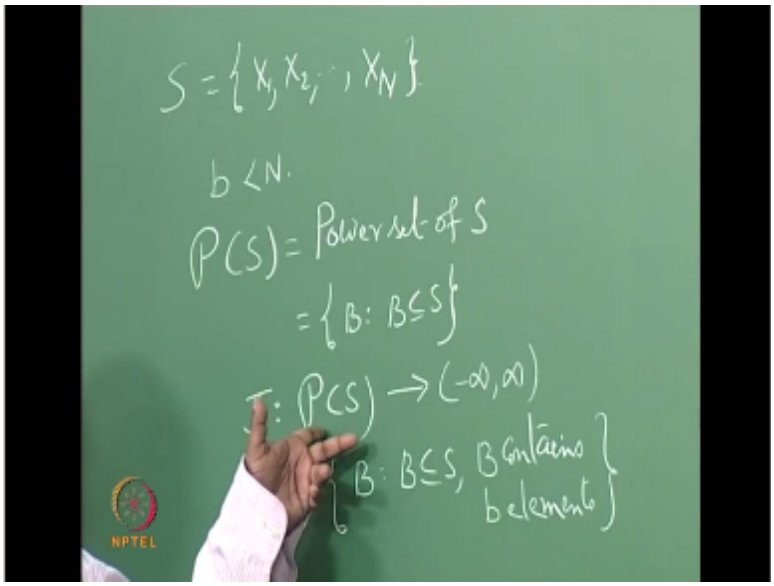
Lecture 02

Feature Selection:
Branch and Bound Algorithm

Prof. C.A.Murthy

Machine Intelligence Unit,
Indian Statiscal Insistute. Kolkata

(Refer Slide Time: 00:15)



Yeah so we have a set of features this is the set of figures B is the number of figures to be chose naturally B < 10 okay now we need to define a function J what is the meaning of defining a function J the meaning is the following let us look at what is known as power set do you all understand the meaning of power set power set means set of all subsets so I am looking at power set of s how many elements are there in power set of s to power capital n okay.

And what is the meaning of power set of X power set of s is any be set of all such beasts be the subset of s okay now the objective function J it should be defined from power set of s to see they used to be defined from power set of s -∞,∞ it may be 0 to ∞ also okay so I am writing it in a very general form -∞ to ∞ and J is to be optimized what is the meaning of J is to be optimized the meaning.

Let us look at this what is the meaning of this is the set of all B, B is a subset of s and B contains how many elements small B elements that means we are going to look at all possible subsets of s containing small B elements please for every subset of s we have to somehow attach a value which tells you the importance of the that section and the word importance from depending on the context to context situation to situation.

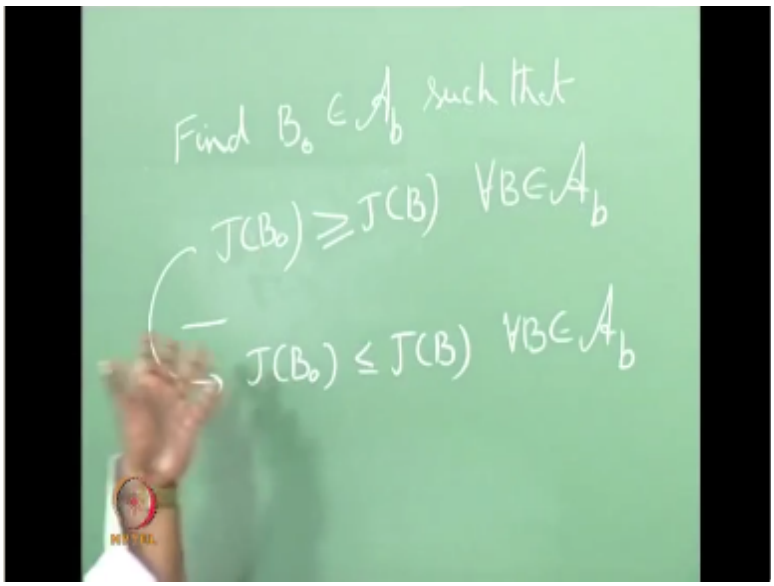
It may differ it may be a negative value also it need not always necessarily be a positive value depends on what sort of functions you are going to define it depends on that okay so in general I am just you may if you want you can always confine it to 0 to ∞ if you want you can always confine it to 0 to ∞ I am not saying that you should not confine it to this thing but what I am saying is that as a general formulation J which is supposed to give you some sort of importance of that particular subset.

Okay it may take values j may take values between -∞ to ∞ I am not saying that J will take all the values between -∞ to ∞ I am not saying that because this is an uncountable set infinitely many elements are there here you have finitely many elements so it can never take all the values in this interval are you understanding what I am trying to say it can never take all the values in this interval have answered your question.

I think your trouble is y -∞ J assigns a value to every subset what is this P of s P office contains all possible subsets when we say that F is a function from a to R we take an element here and then we calculate f of X right and what is an element here an element is the set so you are going to have J of that particular set a function.

That is defining the value of it is taking P of s as a input and giving some value it takes every element of P of s as an input and for every element it is it gives a value yes it gives a value and what is an element of P of s no it is a subset of B so J attaches a value to a very subset is it clear.

(Refer Slide Time: 04:57)



And we need to find this function J our objective is to find J what is J it is J is not unique number one J is not unique I will give you several examples of J in the later portion of my talk I will give you several, several examples of j, j is not unique I am writing all these things to give you how to tell you how to formulate the problem mathematically for mathematical formulation J is to be defined.

What is the meaning of defining J the meaning of defining J is a part of it is this J is to be defined for every subset of s what is the meaning after that means J is the domain of the function the, the domain of J is power set of s and this is the it is clear I am going to tell you how to define J I will surely tell you how to define J later after one or two lectures but this is the way that you are going to formulate the problem.

So once a function J is defined then we are supposed to find optimal value of J what is the meaning of finding optimal value of J first you are going to look at all possible subsets of s having small B number of elements we are going to look at all possible subsets of s having small B number of elements all possible subsets of X having small V number of elements I am denoting it by script a suffix be okay now J is to be optimized here.

That means if we are looking at a maximization problem then it is find B not belonging to a B such that J of B0<=J of B for all B belonging to this, this is for maximization problem for minimization problem this is going to be this is for minimization problem if J is to be maximized

then we should find B not satisfying this if J is to be minimized then we should find B not satisfying this and B0 should belong to script a B is this fine look I will tell you an example of J I think that will make it to more clear.

I hope all of you remember what B a yes decision rule is we know that it minimizes the probability of misclassification okay now we are supposed to find small B number of features for which if you use those small B features the probability of misclassification is minimized for those small B features I think it is not clear to you for every collection of small B features okay say let us just say capital n =10 you have ten features.

Let us just say small B is equal to two you are supposed to select two features so how many two element subsets you can have 10C to 45 subsets you can have okay you take one subset one subset is let us just say the subset containing 1 and 2 X1 and X 2 if you take X 1 X 2 as your only feature what is the minimum probability of misclassification that you can get there is one value similarly X 1 and X 3 you take what is the minimum probability of misclassification.

That you can get like that for this 45 subsets you are going to get for each one of them you are going to get minimum probability of misclassification now among these 45 subsets which one of them provides minimum of these minimums that is your optimal subset is it clear or it is not clear for the subject having elements 1 & 2 that is X1 and X 2 these two features what is the meaning of minimum probability of misclassification.

The meaning is if you take just these two features you can have several decision rules for each decision rule you have your probability of misclassification among all these decision rules which decision rule will provide you the minimum probability of misclassification are you understanding now if you take only X 1 and X 2 just these two features you can have several decision rules.

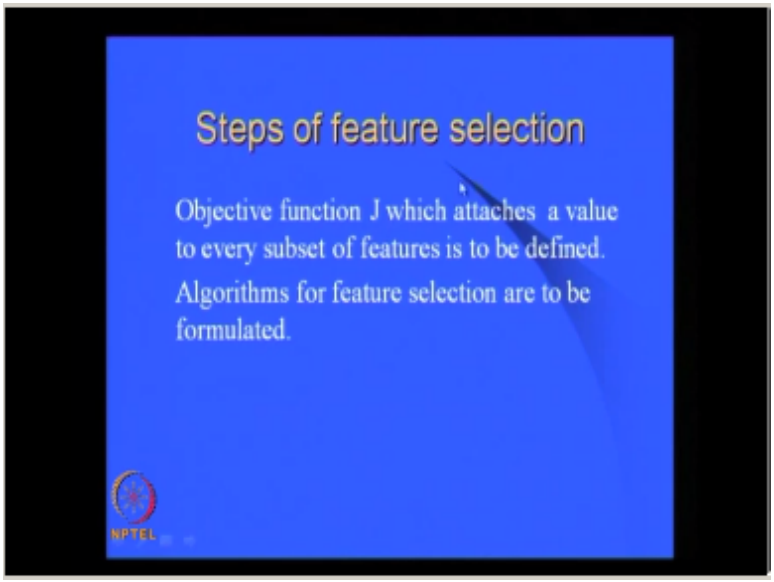
Let us say then the number of classes is 2 then I might have a decision rule like X 1+ X 2 <= 1 put it in class 1> 1 put it in class 2 this is class 1 and class 2 let us just say you only have two classes so this is one decision rule and I can have another decision rule in fact the number of decision rules is infinite and for each decision rule you have a probability of misclassification now among all these decision rules which decision will provide.

The minimum probability of misclassification you take that that is the value that I am looking at for just those two features what is the minimum probability of misclassification that you can get similarly if you take features X 1 and X 3what is the minimum probability of misclassification that you can get like that for every such double ton you are going to have probability minimum probability of misclassification.

Now among all these double turns which double turn will provide really the least that is the best set of features so this is a way of defining J a way this is not the only way have you understood okay I suppose now it is clear clearly this is a way of defining J okay.

So if J is to be maximized then B0 is the one that you need to find out satisfying this thing if J is to minimize then find B0 with this property okay now so there are two parts the first part as it is mentioned here.

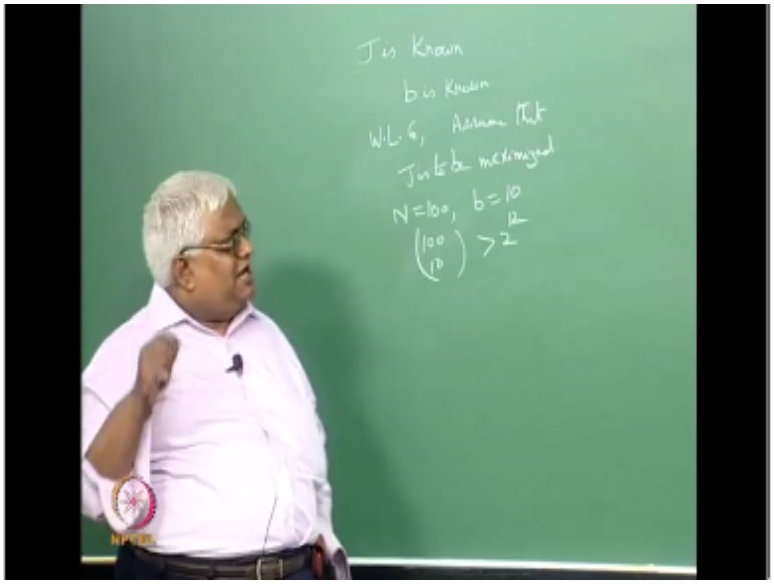
(Refer Slide Time: 14:45)



Look at this objective function J which attaches a value to every subset of frequencies to be defined now the second one is that algorithms for feature selection are to be formulated now in my next two or three talks I am going to assume that J is already given to you I will only talk about how to get the algorithms then in my later portion I will tell you a few ways of defining chain is this fine with you.

Initially I am going to assume that J is given and I will talk about the second part that is how to get the algorithms which provide you the maximum our minimum would provide you the optimal feature subset so now I am assuming that J is given to me.

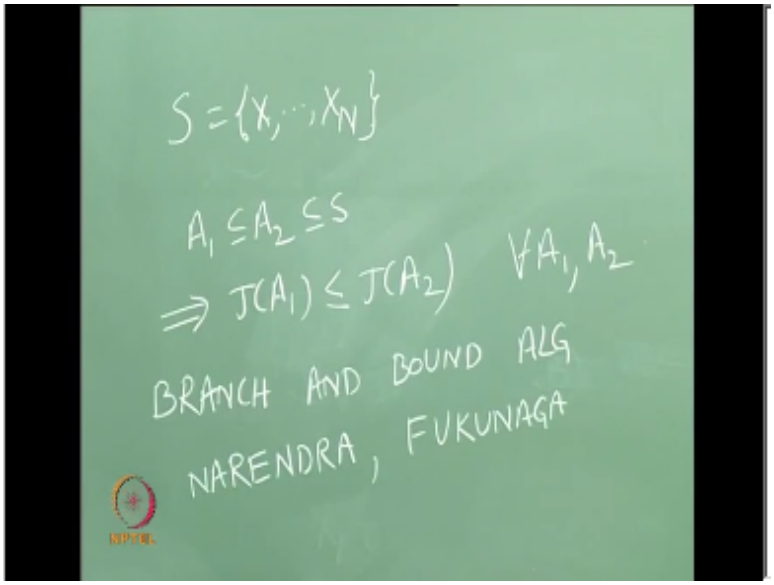
(Refer Slide Time: 16:00)



J is known and given J is given and B is also known the number of figures to be selected and let me also assume without loss of generality that J is to be maximized so without loss of generality assume that J is to be maximized now let us look at the algorithms I already told you some of the difficulties with the algorithms in the previous lecture in the sense.

That when capital n =100 and small V = 10 then 100 see 10 the value is more than 2^{12} right when capital n =100 and small B is 10 then 100 see 10 this is rather than 2^{12} so we are not in a position to do an exhaustive search here and I also mentioned that if you know some properties of the function J then many times then probably you can develop algorithms which provide you the optimal feature subsets without doing.

The exhaustive search without doing the exhaustive search you might get the optimal feature subset if the function J satisfies some properties let me give you an example of that that means what I am going to do is that I am going to assume that my function J satisfies some properties and I will give you an algorithm for finding the optimal feature subset with that with those properties and this optimal feature subsets I would like to find without doing the exhaustive search without doing the exhaustive search I would like to find the optimal feature subset.
(Refer Slide Time: 18:46)



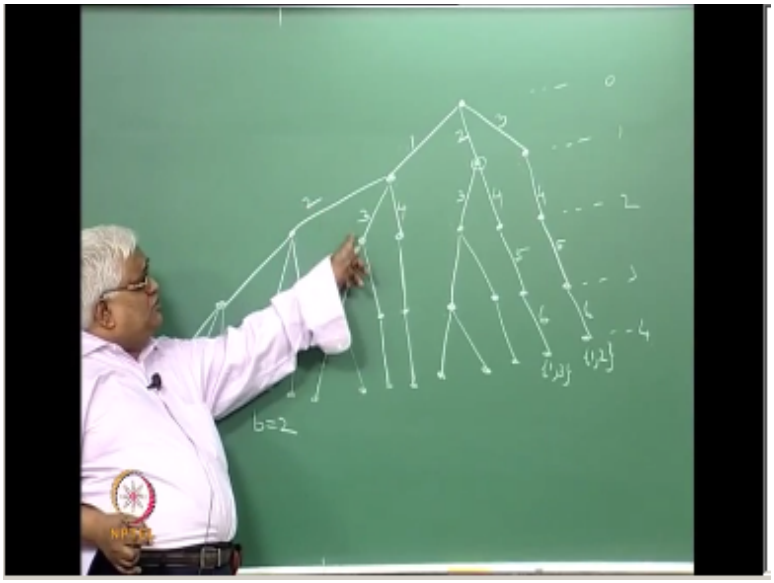
Let me first tell you the properties so this is your set of wickets now the property is the property is if you take any two subsets a1 and a2 such that the Evan is a subset of a2 then J of a 1 <= to J 2 is this property clear to you are including more features means you have more information that is what it is trying to tell you right and here we are trying to do the maximization if you add more features needs you are going to have more information this property need not always hold good.

This property need not hold good always if some features act as noise if you add more features then the importance of that feature subset may decrease in some problems it really happens it really happens so it is not necessarily true that this property holds good for all criterion functions this holds good for some criterion functions.

Now if this property holds good then some people have developed an algorithm for finding optimal feature subsets without really doing the exhaustive search see sometimes even if this property holds you may need to do the exhaustive search if you follow that particular algorithm

but most of the times you do not need to do the exhaustive search if you follow that particular algorithm the algorithm name is branch-and-bound algorithm.
It was developed by NARENDRA an Indian and I hope all of you know who FUKUNAGA is he wrote a book on statistical pattern recognition this algorithm was developed by NARENDRA and FUKUNAGA now I will give you the algorithm okay I will give you that algorithm this algorithm has some parts basically there is a tree for this algorithm now I will draw the tree I will draw the tree there here.

(Refer Slide Time: 22:50)



There are totally here six features and I would like to select two features I would like to select two features there are totally six features okay and so how many possible subsets are there six see - what is the value of 6c- 15how many nodes are there in this layer please count 15 the number of nodes in this layer is actually 15 each node de notes a subset of the original feature set this node denotes.

The whole set that is six features at this level each node denotes a set of five features at this level each node denotes a subset of four features at this level each node denotes a subset of three features and that this level each node de notes a subset of two features okay next this note you have five wickets this node you have six features this is a subset of this, this is a subset of this, this is the subset of this and this is a subset of this in fact if you go through any such branch.

This is a subset of this, this is a subset of this, this is a subset of this and this is a subset of this so what happens is that from here you remove one feature to get this you remove one another feature to get this one feature to get this okay let us say without loss of generality and just for explanation purposes from this one I will remove the feature one feature one is removed to get this say feature two is removed to get this let us just say there are six features x1 x2 x3 x4 x5 x6.

Let us just say feature one is removed to get this feature true is removed to get this now what I will do is that when I am drawing this whole thing this particular branch what I will do is that I always keep the features one the features one and two here that means let us just say feature 3 is removed to get this feature four is removed feature five years removed say feature 6 is removed so ultimately this will be now what I will do is that I get the value of J for this one I have some particular value of J for this.

Now I compare this value with this suppose this value is greater than this then this value would surely be greater than this, this and this are you understanding suppose this value is greater than this then this value will surely be greater than this, this and this and suppose this value is greater than this one then it will be surely bigger than all these things so if this value is greater than this and this then I do not need to search.

This whole thing I do not need to search the whole tree I will just stop if this value is greater than this and as well as greater than this then I will just stop it and I conclude that this is the best feature subset are you understanding I conclude that this is the best feature subset and supposing say this is not greater than this then what I will do let us just say I remove the feature 3here to get this then in this whole branch what I will do is that I will keep the feature 1.

And feature 3throughout here that means let us just say feature 4 is removed feature 5 is removed we can 6 is removed then this is going to be 1 3now what I will do is that I will search the value of J naturally when I am doing all these things when I come to this I am going to compare this with this also okay and let us just say this is greater than this now my value of J is this now I will compare this one with this and this if this one 3 is the value of J for this one if it is greater than this and as well as greater.

Than this then I do not need to search the whole tree NARENDRA and FUKUNAGA made the tree intentionally asymmetric the reason is that if the tree is symmetric then you do not get any

advantage here our aim is not to search the whole space so the place that we do not want to search we want it to be quite a lot.

And the place that we would like to search it will be a smaller area so the place that we do not want to search that is actually here that is the way they wanted to formulate and the place that they want to search most of it is here I think before I proceed further let me first let me tell you how to construct tree.

And then I will tell you the meaning of which place you do not need to search and which place you need to search let me first tell you how to construct the tree for the construction of the tree note that this node has three branches whereas this has only one this has two and somehow this has one and this has two and if you look at this, this has three branches the question is how does one decide the number of branches of, of a particular node how does one decide.

The number of branches of a particular node first let us see how many levels are there let us just say this is the zero th level this is the first level is the second level the third level this is the fourth level so the number of levels is small n -b one the number of levels is sorry capital n capital n is 6 - 2 4+1555 levels are there right now let me write down the formula number of branches from a node number of branches from a node plus number of features to be kept to be preserved or to be kept to be preserved at that node.

This must be equal to B +1 this must be equal to B +1 let me explain let us look at this node there the number of features to be preserved at that node there is nothing that is zero so for this you need to have three branches B is 2 so B+ 1 is 3 now you come to this so here this is one branch this is another this is another so you need to remove one feature to get this let us just say feature 1 is removed how to come to this figure 1 I will come to it later.

And here figure 2 is removed to get this now when you come to this one I am saying that you have to always keep these two features right now so if you look at this the number of features to be preserved at that node is 2 for this so the number of branches should be 1 to make it 3 again for this the number of branches is 1 to make it 3 again for this the number of branches is 1 to make it 3.

Now if you come to this node the number of features to be preserved is only 1 you have to always keep one in all these things here this one is to be always preserved here that means

feature X 1 should be present throughout so for this one feature is to be present so the number of branches is 2 the number of branches is 2 so let us just say the feature 3 is removed to get this now we threw out this portion 1 & 3 we have to be always be there so here one here one here one now if you come to this again feature one this is the one difficult to be preserved.

So for this you need to have two branches for this you need to have two branches the tree is constructed from right to left the tree is not constructed from top to bottom at every level like this it is constructed from right to left first you construct this then you construct these portions then you construct this if you come to this there are no figures to be preserved so you need to have three branches B +1 is three.

Now let us just say figure 2 is removed feature 3 is removed and say here figure 4 is removed now in this totality here you have to always keep 2 and 3 so the number of branches for this is 1 and 4 this is also 1again for this only feature too used to be kept that in one feature so you should have two branches so similarly fluffier here no feature is to be kept you need to have three branches so this is how that tree is generated now the question is we know how to generate the tree.

But then I said here feature one is removed to get this feature two is removed to get this feature three is removed to get this how does one decide those things right so how does one decide that here people have followed many ways I am telling you a wave but that is not the only way that you will find in books you will find this algorithm surely in FUKUNAGA book on pattern recognition you will find it in devourer.

And click list book on pattern recognition okay you will find it at many, many places branch and bound algorithm for feature selection it is a very standard algorithm people have been using it for a long time and you know when it was published actually NARENDRA and FUKUNAGA they wrote two papers both of them were published in IEEE tribute transactions and computers one was in 75another one was in 77 75 and 77 one of them was branch.

And bound algorithm for feature selection another one is branch and bound algorithm for Canaan root k nearest neighbor so one was in 75 one was in 77I do not know which is in which year I and I do not remember it right now you can search it in the internet to get it.

So this algorithm has been in existence for a long, long time now so people have made many modifications in this one now I will tell you a way of choosing this 12 3 say here in this example capital n = 6 okay choose three features randomly from these six features let those three features be.

Let us name the three features as say Z 1 Z 2 Z 3 let us name those three features as Z 1 Z 2 and Z 3 okay now let us look at let us look at the value of j4 s -xi-1 s - i2 s - x i3 s – I even me if you remove z1 you are going to have five wickets in s let us say the relationship is less than or equal to that is if you remove examine the value of J is reduced the maximum here, here it is second maximum here it is reduced the minimum.

Then in such a case here you write z1 here you write x i2 and here you write by three so that this is going to be z1 z2 now look at the way this is done J of s - xi-1>=s -X I 2 <= right 3 that means this value is greater than or equal to this value so if this is actually greater than this, this is anyway greater than this so if this is greater than this you do not need to check anything it is clear now why is it written like this YX I 1 is written here in situ is written like this here note that among those three features removal of Zion has reduce.

The value of s maximum that means z1 is expected to have more information than the second most important feature is excited and seemingly the least important feature is great 3 so between these three features which two features are expected to be more important z1 and z2 that is why we have kept z1 and z2 here they are here so when you decide the number of branches if it is 2 or if it is 3 or if it is 4 then from the set of features under consideration for example here the number of branches is 3.

And the number of features that you have is 5 here so from these five figures select three features randomly and find which one has more information that you write here I am that feature you remove it here and which one has slightly maybe it slightly less information than this but more information and the third one that you write here that means that feature is to be removed here and those two features.

They have to be kept in the third one so this is the way you are going to proceed I think it may be clear to you why the phrase branch and bound are used you will develop a branch only when the bound information is not satisfied you will develop a branch only when the bound information is

not satisfied are you understanding what I said first you will develop this and then you compare this with this and this anyway this is greater than or equal to this.

So if this is greater than this then you do not need to do all these things you have got the optimal feature subset otherwise you need to develop this and suppose this is greater than this then you do not need to do all these things you can just be satisfied with this so if this is greater than this note that quite a bit of portion in the tree you are not searching so my comment about vast portion of the search space we do not want to search.

I hope by now you have understood it is this clear to you I will stop here four minutes left I will stop here but I surely want your questions right I might select Zhang's idols I think he may select the joyful joy full sizes that can happen if at the starting they are selecting bad features that can happen yes I am okay I think it is not necessarily true that we will not search the whole tree always what I am trying to say is that many times we do not need to search.

The whole tree can we say what probability of searching whole tree given a random set of feature I think some such work probably they have done it some such work about what is the probability of searching the whole tree what is the average complexity I think some such works have already been done but the worst case is you have to search the whole tree here in fact searching.

The whole tree mean it is not only just the 6215 it is also these things plus these things plus this so it is much more than capital n see small beam it is much more than that but many times you do not need to do it please real scenario sir given a difficult to ensure that set of feature is holding the property agreed that was the first thing I said we cannot apply this thing for deal problem.

This algorithm of course whatever you are saying is perfectly correct if some real life problem are some criterion function does not satisfy the assumption that was me then this algorithm is not really useful because the main assumption is not satisfied there how will we ensure that our feature city is satisfying that particular criterion function yes sir that is itself a complex job to find it we cannot ensure it how can it be insured.

Because some functions they do not satisfy it some functions they do not satisfy it the example that I gave you minimum error probability minimum probability of misclassification that

example does not satisfy this I can assure you that there the moment the number of features is increased the probability of misclassification decreases that is not true I have a feature set and I want to apply this algorithm.

Then what should we look for can we apply this for the given set or not because you if you have a feature subset if you have a set of features and if you have a criterion function also and if you are satisfied that the criterion function satisfies the assumption that is made even then it may be difficult for you to apply this particular method because if capital n is equal to 100 say small B is equal to10 look at the size of the tree.

That you are going to draw look at the size of the tree that you are going to draw for really very high values of capital n this algorithm may be extremely difficult to implement even if the criterion function satisfies those properties and if it does not satisfy then you have a real problem please is there any way to choose this features like Z 1 Z 2 Z 3 at the first optimally so that the sort space is reduced ah.

So what some people do is they do something like sequential backward I mean the removal of which feature makes the J value decreased the maximum that particular ticket which decreases the value of J the maximum when it is removed then that one you can write it here instead of just Simon that means you have six features first you remove feature one then you have five meters find the value of J then from the six feature set to remove.

The feature two then again you have five figures find the value of J like that you do it for all the six subsets each subset is having five elements find out that subject where J is decreased the maximum from six to five that means removal of that particular feature made the value of J decreased so much have you understood that you write it here so then why cannot we just use a sequential backward search and we need to do we shall be discussing all these things I am going to sequential forward sequential backward selection in my next lecture any question sure.

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 Krishnamurty & Team

Animations
Dvijavanthi

NPTEL Web & Faculty Assistance Team
Allen Jacob Dinesh
Ashok Kumar
Banu. P
Deepa Venkatraman
Dinesh Babu. K.M
Karthick. B
Karthikeyan. A
Lavanya. K
Manikandan. A
Manikandasivam. G
Nandakumar. L
Prasanna Kumar. G
Pradeep Valan. G
Rekha. C
Salomi. J
Santosh Kumar Singh. P
Saravanakumar. P
Saravanakumar. R
Satishkumar. G
Senthilmurugan. K
Shobana. S
Sivakumar. S
Soundhar Raja Pandian. R

Suman Dominic. J
Udayakumar. C
Vijaya. K.R
Vijayalakshmi
Vinolin Antony Joans

Administrative Assistant
K.S. Janakiraman

Principal Project Officer
Usha Nagarajan

Video Producers
K.R. Ravindranath
Kannan Krishnamurty

IIT Madras Production

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

www.nptel.ac.in