Indian Institute of Technology Madras

NPTEL
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING

Pattern Recognition

Module 04

Lecture 05

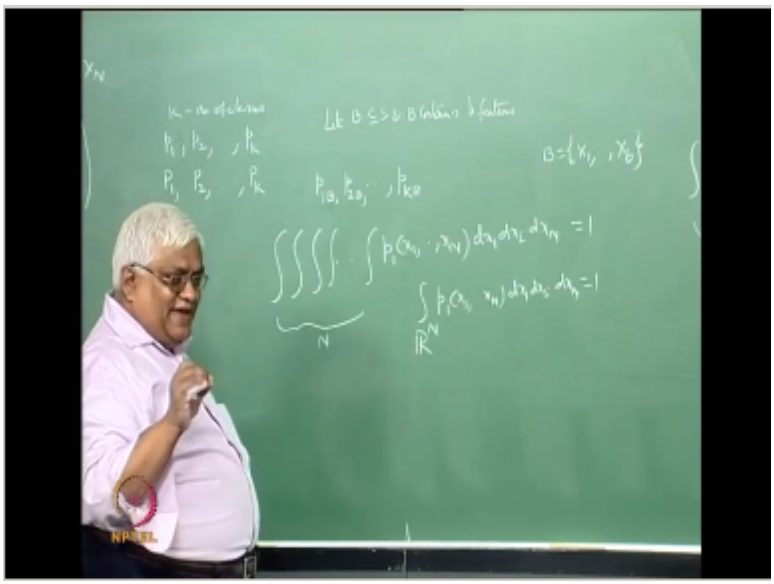
Feature Selection Criteria Function: Probabilistic Separability Based

Prof. C.A.Murthy

Machine Intelligence Unit,
Indian Statistical Institute. Kolkata

Continuing the feature selection we have been discussing so far the algorithms like sequential forward sequential backward generalized sequential forward generalize the sequential backward LR algorithm and branch and bound algorithm there are many other algorithms available so now I will start discussing the feature selection criterion functions how to choose the, the choice of the feature selection criterion functions.

(Refer Slide Time: 00:54)



So again so there are x1 to xn you have totally n variables are capital n number of features and so your feature vector is going to be off if I write the feature vector as for the I earth observation if I write the feature vector then this is for the I earth observation the value of the first feature is x1 I for the earth observation the value, value of the second feature is x2 I for the earth observation the value of the NF feature is xn.

I so it is your feature vector will be in capital n dimensions okay now let us just say that there are how many classes let us just say there are K number of classes the number of classes is K the number of classes is K and let us assume that we know the class conditional density functions that is for the first class it will be small p 1for the second class it will be smallp24 diet class it will be small PK so there are K class conditional density functions probability density functions and.

Let us just say there are the corresponding prior probabilities p 1 p2 up to PK naturally each of them is greater than 0 and then the summation is actually equal to 1 let us just say we know this we know these things and as well as we know this now we are supposed to select small V number of features small B number of features they have to be selected so let us just consider one such subset let us just say capital B is the subset office let us let B is a subset of s and B contains small be features okay.

Let us say B is a capital B the subset of s such that capital B contains small V number of features now with respect to those small V number of features we can have the density function p1b p 2

b. p k be using those features in this subset capital be the corresponding conditional probability density function for the class one is small p 1 b do you know how to get this small p1b from the small p 1do you know it if you do not know it.

Please tell me how you will get is you see the small p 1 is going to be here these integrations are how many such integrations capital n number of such integrations small p one of this you know because since small p 1 is a density function and it is number of arguments is capital n so you have to write those capital n variables x1 to xn and dx1 dx2 DX n and these number of integrations is we are going to integrate over capital and dimensional space.

Okay maybe each of them is from each of them is our okay our this is equal to one okay integration over capital R to the power and right now you look at those be features here let us say without loss of generality this let us say the subset b is equal to let us just say X 1 to X B these are the B features that we have taken let us just say then what we will do is that look at this function P 1 of x 1 x 2 x b XB + 1XB + 2 x n.

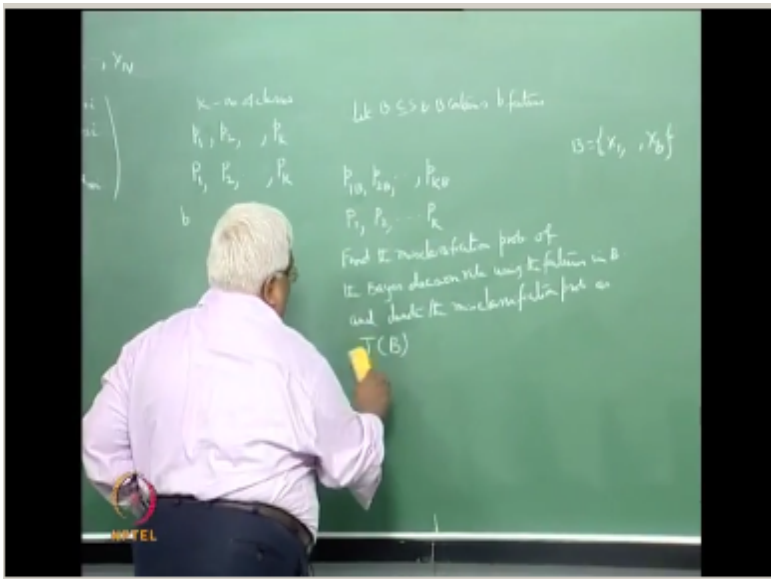
Okay this I will do the integration over D X this I will do the integration so this will be how many n -b over n-B dimensional space are basically this is same as integral R to the power of n - B p1 so if you do the integration what are you going to get let us just say the function that you are going to get let me just write it as G and for the class 1 let me just write it as g1actually here I wrote p1 be.

So let me just write it like this only p 1 for the SEC be what are these arguments because X 1 to X B they are there you are doing the integration over B + 1B plus 2 etc are you understanding you are doing the integration over B +1 B + 2 etcetera be up to capital n so there are some variables which are their remaining there what are the variables that are remaining X 1 to X B they are this in capital B you have these features so that is your P 1 B have you understood what I wanted to say similarly you can have P to B p3b pkb.

And that you can have it for any such subset of s it may have small B number of elements it may have I mean this capital B can be any subset of s having small B number of element which are those small B figures that you can vary by varying capital B you can get the corresponding this set.

These density functions these are actually called marginal density functions you are doing integration with respect to the rest of the variable so that the variables that you want they will just remain as they are okay so once you get this p1 BP to be pkb then what you can do is what you can do is.

(Refer Slide Time: 11:07)



You can use these prior probabilities and you can get the base decision boundary and you can get the base miss classification probability so find the Miss classification probability of the base decision rule using these be features using the features in Band denote the Miss classification probability as and denote the Miss classification probability as JB is this clear denote the Miss classification probability as JB now you are supposed to do the feature selection.

So now what is your criterion function your criterion function is I mean this j and you are supposed to find that particular subset b not a containing small B number of elements for which j of B0 is minimum that is find, find be not subset of s be not containing be elements such that such that symbol is this j of b0 <= j of be for all be subset of s containing small be umbra filaments is this clear to you.

This is this is this should be the way in which feature selection should have been done but usually people do not go in for this what is the reason the reason is very simple for most of the problems you really do not know the probability density functions for most of the problems you really do not know the probability density functions that is one part of it but the second part even

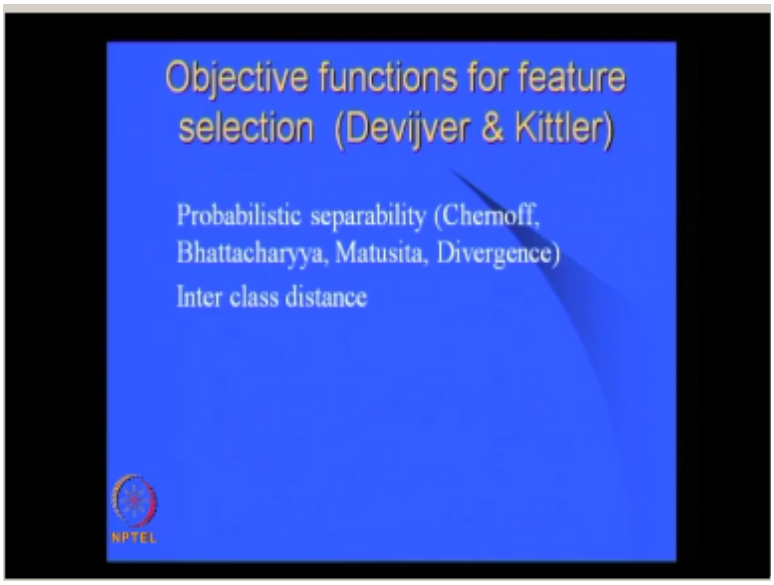
if you know it is difficult to calculate jfb it is difficult to calculate the Miss classification probabilities if you look at my earlier lectures on base decision rule which was supposed to be the best rule.

Then why at all we went in for other decision rules the reason was that even if you know the probability density functions it is extremely difficult for you to find them is classification probabilities to find which one is actually the minimum it is difficult to get those I mean decision rules and the corresponding miss classification probabilities it is difficult to get the Miss classification probabilities.

So this should have been the rule that people should have followed it but actually it is difficult to follow it because you really do not know how to I mean though the expression is fine but it is difficult to evaluate the expression for Miss classification probabilities then people started wondering is there any way in which one can actually choose features using probability density functions.

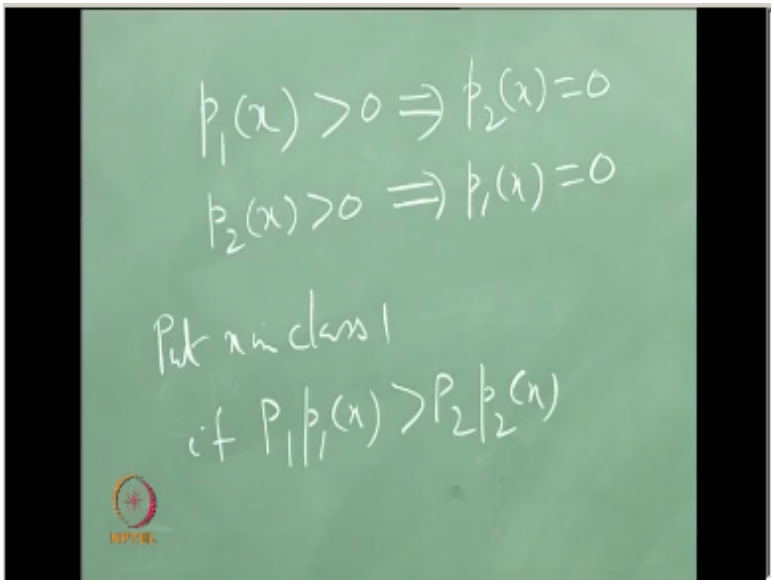
But without actually going in for calculating based service classification probabilities how to look at density from without really trying to calculate them is classification probabilities this was the question people ask themselves and then they came out with the probabilistic supper ability based criterion functions now what is the meaning of probabilistic separability based criterion functions.

(Refer Slide Time: 17:10)



In order to explain probabilistic separability.

(Refer Slide Time: 17:35)



Based criterion functions I will explain things in this following way say these are two density functions for the classes say this is the density function for the class one say this is the density function for the class to write and then note that there is a gap here okay note that there is a gap here so these two functions they satisfy a property what is the proper property p 1x greater than 0 implies P 2 x = 0p 1 of this thing is greater than 0 what about p2 for these points that is equal to 0.

Similarly okay now suppose we have chosen the required number of features in such a way that those features they separate out the density functions like this then what is going to happen to the base decision rule in this case if your density functions p1 and p2 are like this what will happen

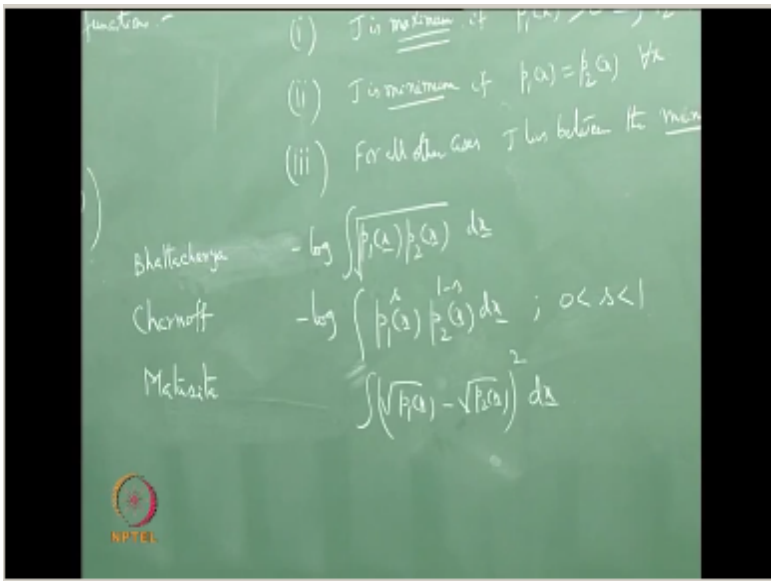
to your base decision rule does it give you any miss classification yes or no my claim is that there is no miss classification why what is the base decision root base decision rule.

Is put X in class 1 if p1p1 x > p 2 p 2 X this is in the case of two classes now here whenp1 x > x is 0 so for all the points which are here they will go to class 1 and for all the points which are in this region they will go to class 2 because p 2 p 2 x will be > 0 but p 1 p 1 x =0 so then there is no miss classification right then there is no miss classifications so what people thought was that they would like to choose features in such a way that the density function the separation.

Between the density functions is as much as possible the separation between the density functions is as much as possible that is the basic idea that is why it is called supper ability that means separation between the density functions since we are looking at density functions people called it as probabilistic supper ability have you understood it now so this is called a probabilistic separubility based feature selection criterion functions.

Now the question is how do you define the separation between two density functions if you define that then you can talk about which features are going to give you the maximum separation are you understanding what I am trying to say if you somehow define the separation between the density functions then we can say which particular features gives you the maximum separation so how does one define the separation between the density functions let us see.

(Refer Slide Time: 22:57)



So you have say two classes so you have density functions p1 and p2 you have prior probabilities p1 and p2 okay now you need to define a function which you need to define some g some sort of integration may be done okay and then you need to have let us just see re define some g1 and then again you need to define AG.

So this is a general way in which you can define separation between density functions you need to have some g one here and then some g two but then it should have some properties what are the properties the properties are one jay is maximum if p1 x >0 implies P2 x = 0 that maximum value for some functions it may be of the order of 1 for some functions it may be ∞ by varying this g1 and g2 you may get several.

Such maximum values I mean people defined it in just too many ways okay to j is minimum if p1 x = p 2 X for all X then there is no separation between the density functions am I right then you should have a minimum value are you understanding first I am trying to give you the intuition now I will give you the formulas it should be maximum whenp1 x < 0 implies P DX = 0 okay minimum when p1 x = p 2 X for all X I am expecting a question from you look let me ask you that question here I wrote only one p 1x < 0 implies P DX = 0 do I need to write the other one p2 x < 0 implies P 1 X = 0what is the other one do I need to write it.

I claim that I do not need to write it I will tell you why I do not need to ride it suppose this is not true that means there is one x for which PT x <0 as well as p1 x <0 but that does not happen okay if there is an X for which both p1 and p2 I have < 0 then it contradicts the previous statement are you understanding have you understood the logic so I do not need to write this just this one statement is sufficient is it clear to you.

And the third one is that so I wrote maximum and minimum and all the other cases they should be in between this maximum and minimum for all other cases J lies between the maximum the minimum and maximum this minimum is same as this minimum and this maximum is same as this maximum okay so using these principles several criterion functions have been defined several criterion functions means several separable ax t measures have been defined.

And using those separatebility measures you can have the corresponding criterion functions for feature selection okay now what are those supper ability measures which tells you which tell you how to look at the I mean how to quantify the separation between the density function okay the

first one I am going to give this will be by Bhattacharyya okay this is integral this is web Bhattacharyya this log is to the base.

Okay it looks to be quite complicated but actually it is not let us look at the case too that is the second one suppose p 1 x = p 2x then what is going to happen to this one this is ∫P 1 X and you are going to do the integration over about the whole space so this value is going to be one log 1 = 0 and - 0 is 0 so the minimum value is 0 now let us look at the other one what will happen to this product the product is 0√0 ∫ of 0 is 0 and log of this thing it is -∞ actually log 0 is not defined when something that is going towards 0.

Then this goes towards -∞ and - log is plus ∞ have you understood now have you understood why such an expression is taken is it clear and you know who this paddock area is he was a professor in Calcutta university as I was mentioning in one of my discussions with you is I was created by maulana beez in the Calcutta university way back in 1931 okay he started is I in 1931 in Calcutta university there has always been a close interaction.

Between the scientists of Indian Statistical Institute and Calcutta university okay and Fatah josh was a professor in Calcutta university I think he died around eight to ten years ago he never went abroad he never went abroad in a very famous scientist and this distance is known as pat archery distance probably you might have heard the term padded area distance so this is the one package area distance is this okay it was a very famous statistician.

And he actually developed this one buttock area distance and this was generalized by Cherenkov so how did he do it as you can see here if you put s is equal to half you will get better carrier distance right so he had judged this was generalized he took any s lying between zero and one so then you will get a generalization and Matthew criticism a too sitar their arm how did he do it this is integral square root P 1 X -√ p 2 X ^{2}DX okay.

And there are in fact many, many more there is something called divergence and there are many, many more such measures each of them uses the density functions okay using these separable ax t measures between the density functions one can always define the corresponding criterion functions for feature selection.

For example how do you define a criterion function using this Bhattacharya distance how will you define it you say that you take those small v features for which the vertical distance value is

maximum okay you take those small B features for which the Fatah carrier distance is maximum so for each for each such set you will get the corresponding what I wrote there p 1 BP to be I wrote when I explained it to you, you will get the corresponding p 1 B and P to B.

And then you calculate the perfect area distance so for every such set B containing small B number of elements you have the corresponding vertical distance value and you find out that b0 for which the distance is maximum it is the separation is maximum okay similarly using Cherenkov are using meth you Site are I mean looking at the there are many other measures you will get the definitions of these things from divider.

And Keith Earth's book you can get there are in fact too many of them he goes on giving the corresponding formulas for those measures right now note that in these cases I assume that you have two classes but supposing you have K number of classes instead of two classes suppose you have K classes that means p 1 p 2 PK and here you have capital n capital B to capital PK now how are you going to get the separation.

Between K such density functions how do you get the separation between the K such density functions we will look at what is known as the mixture density function px let us just define it as this is mixture density function that is one terminology for this and you also have another terminology for this, this is actually the mean of these I density these K density functions is not it the mean of this K density functions.

These are the probabilities multiply the function by the probability take the summation you will get the mean and we are supposed to look at separation between any two such density functions and whatever I did earlier that you do not need to look at the separation between any two such things you always look at the separation of that function with respect to the mean and take the summation that will give you the separation between any two of them.

And take the summation have you understood what I wanted to say what I did about the variance you are looking at the difference between the mean and each individual value taking the square and then doing the summation whereas we are supposed to look at the separation between any two of them and take the square and look at the summation here the problem is we are supposed to look at the separation.

Between any two of them and then take the summation to get the overall value so instead of looking at that we can always look at the mean and take the difference separation between each of them with respect to the mean and you sum it up have you understood what I wanted to say so you can have the corresponding generalization of these things not with respect to density functions but the generalization with respect to M density with k density functions.

So what is the generalization the generalization is for Bhattacharya it is going to be - log √ p I X and this is PX DX this you multiply it by π and you take the summation I = 1to k you take the summation I= 12KPI bracket and you write it with respect to π and p have you understood this is just the generalization of Bhattacharyya distance but not for two classes but for K classes taking it from the for the two classes whatever may be the formula that formula is generalized4k classes here.

And you can have the corresponding generalizations for Cherenkov and also for Matthew zeta and for other such measures other such separubility measures and there are simply many of them in the literature they are simply many often in the literature since the earlier part of the literature it was for feature selection people wanted to directly use the statistical principles so there are several, several such papers using these distance measures which you would find in the early part of the literature on feature selection.

This is denoted as probabilistic dependence measures this is the terminology that was used by divider and Hitler in their book they called it as probabilistic dependence measures for feature selection which are nothing but generalization of the probabilistic supper ability measures to K classes simply the generalization of the probabilistic separability measures to take asses we have any questions in my next lecture I will start dealing with when that density functions are not available you have a training sample set.

Then how do you do that how do you get the criterion functions for feature selection here in all these cases I assume that the density functions are available so we try to develop the criterion functions for feature selection now in my next lecture I shall deal with the case where the density functions are not available but a training sample set is available for feature selection and so then how do you get the criterion functions in that case it is you.

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

Copyrights Reserved