Indian Institute of Technology Madras
Presents

NPTEL
NATIONAL PROGRAMMEON TECHNOLOGY ENHANCED LEARNING

Pattern Recognition
Module 02
Lecture08

Bayes Theorem

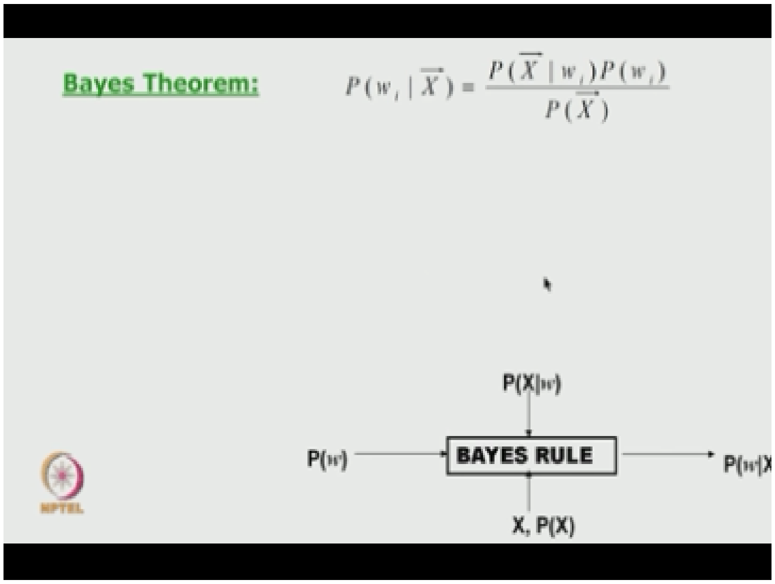
Prof. Sukhendu Das
Department of CS&E, IIT Madras

Welcome back to the lecture series on pattern recognition in the last class we have been discussing about decision boundaries decision regions discriminate functions following the discussion on normal distribution and for the task of pattern classification we took a very simple example of a minimum distance classifier or needle nearest neighbor classifier.

In which you had simply taking the distance of the sample X from the mean without worrying about the class priors and the class distribution functions and we also discussed that this is not a correct method to estimate or to perform the task of classification. Because you will get good amount of errors.

In this process and there are lots of examples to also show that which you will worry about it later okay so to incorporate those we will bring in the classic Bayes theorem okay so let us look at the slide for the classic Bayes theorem.

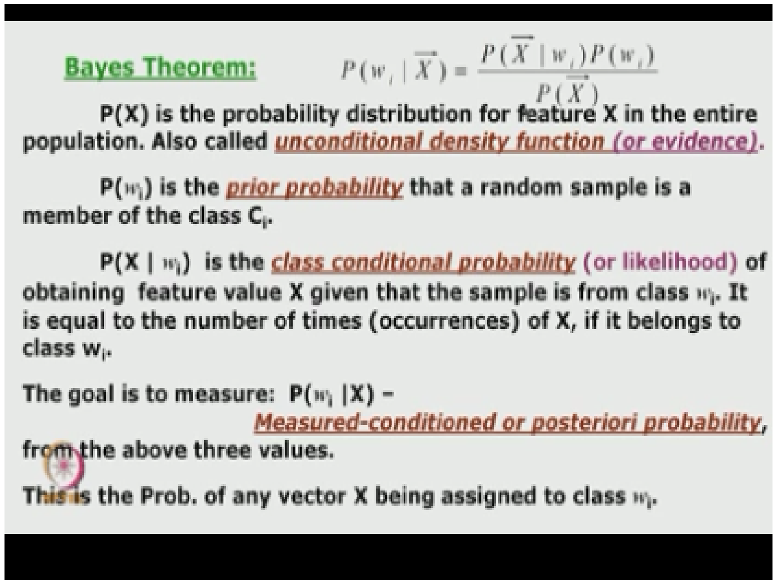
(Refer Slide Time: 01:14)



So Bayes theorem the best decision rule has been discussed in an earlier class where the Bayes theorem is given by the expression as you see on the top so we are trying to classify a sample X this is what the capital X indicates a feature vector with an arrow and the WI belongs to indicates the class I okay WI indicates the class I so what Bayes rule or the Bayes theorem say that it requests three inputs the three terms which you see on the right hand side one two and three are all given here and this is the output what are the individual terms which are here now.

There are different books which will actually use these terms in different ways and I am going to use as much of the common terms which are used to describe these corresponding terms within the Bayes theorem okay which is used for later on used for class assignment okay so what are them this is what you get.

(Refer Slide Time: 02:14)



P of X the denominator term is basically called the unconditional density function or evidence some books will casually call it evidence but I think we will stick to the word called unconditional density function or unconditional probability it is the prior distribution of a feature X in the entire population that is what it means if you are taking color as a feature to discriminate between flowers and fruit I want to find out how much of red flowers are there it is not a question of trying to identify.

Whether a flower is a tulip or a separate category of a flower okay or the classical logic on whatever you want to whatever color it has or the rose okay how many flowers have a colored so if color red or if color is a feature color is a feature the redness of the color for that particular sample week that can be calculated from a used set of samples okay if you want to take some other example let us say the height of a particular person you take a group of individuals you may try to categorize individuals based on different patterns such as say language dress color.

Whatever the case may be food habits but let us say I want to take height as a feature and we use it for a classification whatever the case name but height is a feature so if you take a group of 100 individuals you would like to find out how there are how many people with the height let us say exactly 5 feet or 5 feet 6 inches that will indicate the unconditional density function because you are sampling the data without worrying about the class under which category the rate of all that is P of X P of Ω I or WI as a symbol indicates.

You can use it as a WI as the prior probability that a random sample is a member of the class here index is the same the prior probability that a random sample is a member of the class CI this is sitting in the numerator of the expression here of the base what is called as the casually the class prior okay let us take an example of a two class problem discriminating between apples and oranges I give you a bag of fruits.

And there are ten fruits six of them are apples and four of them are oranges I repeat there are ten samples six of them are apples four of them are oranges so I can compute P of W one for Apple's P of W tube for oranges very simple answer what should be P of W for apples the prior for apples it will be six by ten or whatever it is 0.6 and the other will be of course one minus that will be point four very simple.

So though that is an example of class prior you need to find out because actually what happens is in the case of fruit classification the fruits are sometimes seasonal so it is good to find out for a particular season what is there a certain type of seasonal food which is available in large quantities say mango in summer or could be jackfruit or some other food which could be apples could be typically more or cheaper lettuce at least in the in winter oranges also are seasonal fruit but of course these days we are having almost fruits all through this is an accept mangoes probably not available all that ok let us go back the other term.

Which is the most significant term in the numerator so we are discussed what is the class prior and unconditional probability the term here which is the most significant one is called the class conditional probability or likelihood given a particular class WI how many times does this feature X occurs okay it is the likelihood of obtaining feature value x given that the sample is from a class WI it is equal to the number of times or occurrences of X if it belongs to class WI that means if you pick apples as a class how many fruits will you have which have the color red well that probability could be higher.

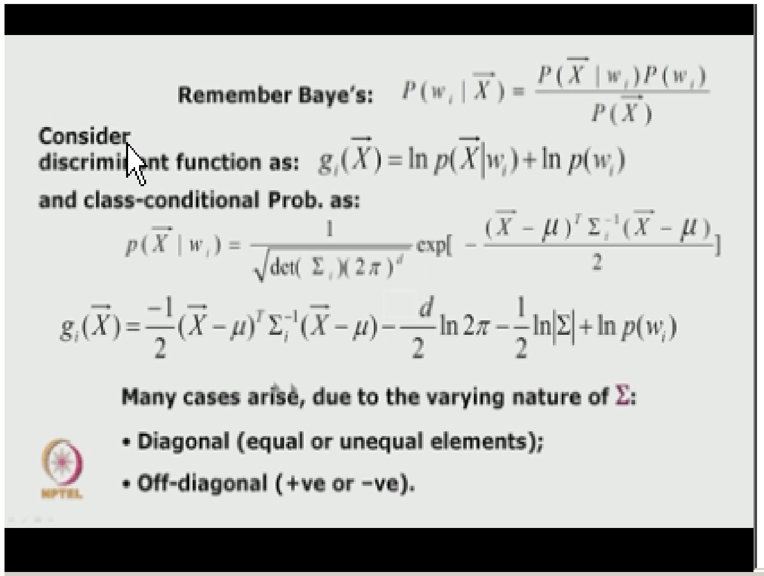
You would like to compute how many red fruits will you have if you are given only oranges now you know how many times an orange can become red or some other color that is a which is not orange which is a black or green well orange is if it can be green but let us take some other color red or black so you can compute the likelihood or class conditional probability as it is called for a particular feature given a particular car.

So given all these three terms on the right hand side a very simple probabilistic estimate will give you the left-hand side quantity which is called the measure condition or the posterior probability and this is the probability for any feature vector X being to be a sign for a class WI. So you assign X to a WI if this measured condition probability is higher for a particular class the rule is the same the rule for assignment is the same whichever the corresponding probability posterior probability under Bayes satisfies.

The maximum likelihood you will assign it but how to compute the likelihood that is done using this formula so in the numerator you have just remember class conditional probability you have a class prior and you have an unconditional can you repeat this with me on the numerator you have class conditional likelihood then second this one is class prior denominator is unconditional profit using this three you compute.

The Bayes theorem so assuming you know this formula let us go back to our discussion on decision boundaries and decision regions to be obtained under the Bayes paradigm so now look back you remember base this is the expression.

(Refer Slide Time: 08:48)



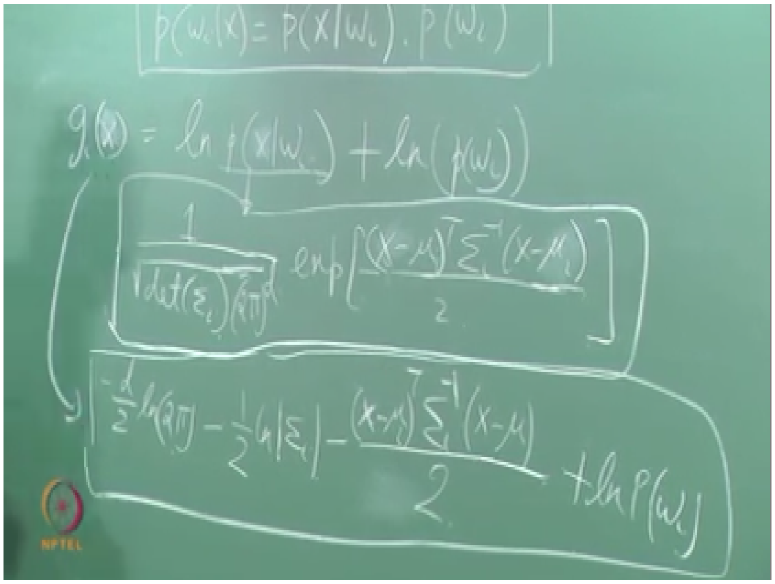
Which you just discuss now okay now consider a disturbance function GI which we have been talking about in the last class a log of the numerator of the expression numerator of the Bayes expression you see these two terms if I take the log this is what I will get and I am ignoring the eliminator what was the denominator here which we are ignoring we discussed that just a few minutes back.

It is the unconditional probability which we are ignoring and we are taking the two numerator terms and taking a log of that expression and within that what we are bringing in now is this term class conditional probability or casually also called the likelihood correct class culture prodigy or likelihood.

This is the class prior okay this is called the class prior okay and this is called the class conditional probability or likelihood is given by this famous expression we have seen this couple of classes back also in the last class beginning this is the what is this expression multivariate Gaussian density function so that pins in a two dimensional space this is the distribution look in the expression you have remember there is a vector sign which I have kept consistent here but after some time I may not keep this vector notation I have tried to be consistent as much as possible throughout my slides actually ideally this should also have a vector sign.

But this is a vector mean vector the sample vector covariance matrix here normalizing term here so using this expression using this expression if we incorporate inside the log can we write the expression of G I will work this out in the board so what did we have we had the Bayes rule.

(Refer Slide Time: 10:35)



I am just writing the two numerator terms can you just tell them what they are the first is the class conditional probability or likelihood what is that multiplied by class prior correct so that is that is my base so what I am ignoring here is the denominator term and what I am writing now is G of I of X my first discriminate function.

In my class okay so what I am writing I am taking a log of this basically okay simple log off so it will be log of this plus log of the classifier very simply and this particular term I am writing here this I am saying will be a let us say this is a multivariate Gaussian density function so this will be root over determinant of ∑ I for the corresponding eyuth class correct so there is a symbol here the subscript I then 2 π to the power D correct the square it will be part of this correct TB overall this is the normalizing term.

We see that is the idea then this is the main thing exponential minus will be there here correct divided by 2 X okay, I should be careful here that X is a vector mind you I had this vector sign inside my slides for the sake of convenience I am ignoring that but if you are copy you can put the vector sign here because both of these are vectors then you have the covariance matrix inverse of the covariance matrix and then X- UI correct will be I here okay so what is this is the this is the O so if you substitute that take the log so this there will be this factor which will come out plus the log prior at the end the interesting part will this factor here.

So if you check it out I think it will be -D by 2 log 2 π here into this you know tell me what I will get here- 1 by 2 log determinant oh yeah and the log and exponential will be cancelling out so there will be a minus sign here so it will be a minus 1/2 by this one not so X -μ I transpose covariance inverse they keep getting used to this term plus look so this is what you get this is your GI this is your GI okay to be very precise you just correct it here this is X this check out if they are running miss mistakes of notations here okay.

Now if we look at this is what we get in this slide as well let us look back in the expression to see that whatever we have derived as you see here it is the same alright look at the expression at the bottom as given here and the one which is now in this slide at the bottom here the only thing in the slide is that I have brought in this covariance term at the beginning.

The rest of it is basically the same Organa in this slightly different manner that is all if you look into this expression here these can be visualized to be constant terms which are outside they are not function of X this is the one which only varies. If we change X of course if you tell you I get means go from one particular class to the other there'll be another term which could be also changing as well as the class prior unless the class prior themselves are constants or uniform across different classes.

That means you have same number of arranges as the number of apples when you go for go to the market or the stores grocery if you want to purchase two types of flowers you have as new number of rows as number of tulips or something like that so in that case class priors also could be same or different covariance matrices also could be same or different across different classes but given a particular class ie., when you change the sample of the instant.

Which is being tested the only term which is varying is this one and this is the one actually which gives you the distance D the distance D which we have been talking about since about one or two classes back the distance D distance of a sample from the class mean nearest neighborhood classifier is actually a special case of this expression with the covariance term suppressed you only have these two you get that D which we talked about and we will see that again as a special case so this is the distance plus these are some constant terms which do not vary with X but if you go from one particular class.

To the other yes there will be some changes which will taking place and we will discuss that but what may happen also is that the covariance term and the class price also could be same as a special case across classes so in fact the concentration henceforth will be mainly on this particular term let us go back to the slide so to recapitulate what we are just done now that within the purview or using the Bayes theorem incorporating the multivariate Gaussian function as the class conditional probability for a particular class we have got a discriminate function expression for GI which contains.

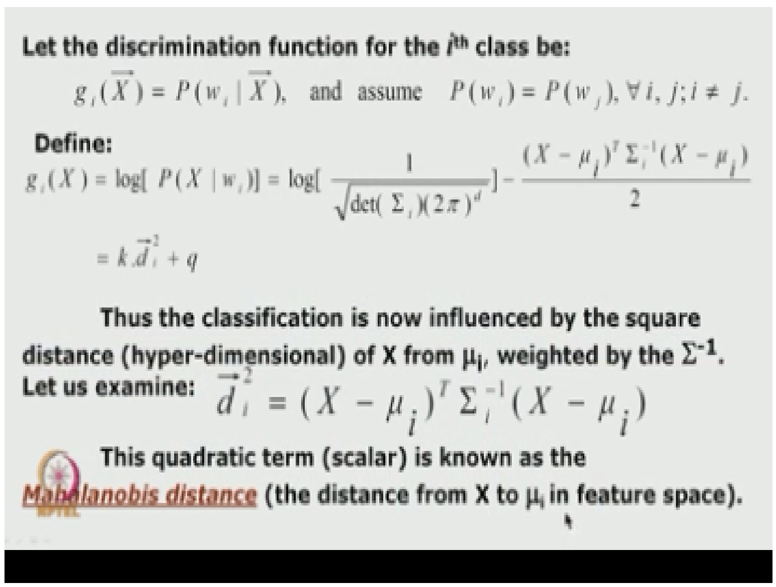
The class priors the class PDFs the class PDFs in turn contain the covariance matrix which is very important which you are just discussed about now in terms of distance about a couple of classes back and of course a few constant terms which is again based on the dimensionality of the problem and the covariance matrix the determinant of the covariance matrix and if you look back into the expression in this slide as given here it is the term which in the left hand side of this expression for G of I of X is the one which dictates your distance from or which easily dictates your classification.

The job of classification and also it is giving a distance measures because all of these the rest of it are not functions of X so when you vary X within this expression it is this term which is changing and it actually it gives a value of distance many cases may arise due to the varying nature of the covariance matrix $\sum$ it could with equal or unequal elements there may be off diagonal terms positive or negative we have already seen some of these variations when we are looking at the ISO contour plots in 2d for a particular Gaussian example he remembered two classes back we had talked about isometric Gaussians in fact.

In the last class we have talked about asymmetric Gaussians oriented Gaussians what we depend on it was the fact that if we had off diagonal term zero or not what the diagonal terms one was equal to the other or not these are all the factors which dictated or whether the Gaussian was a symmetric or not whether it was oriented or not the same thing is going to happen with the diagonal of diagonal terms of the covariance matrix is going to dictate how this particular expression in fact it is an expression of a distance already.

Which we have got using a discriminated functions and this will give rise to DRS which in turn will give rise to D B's let us look at some special cases of this particular function remember this expression of G I which you have just derived it in the class today.

(Refer Slide Time: 19:43)



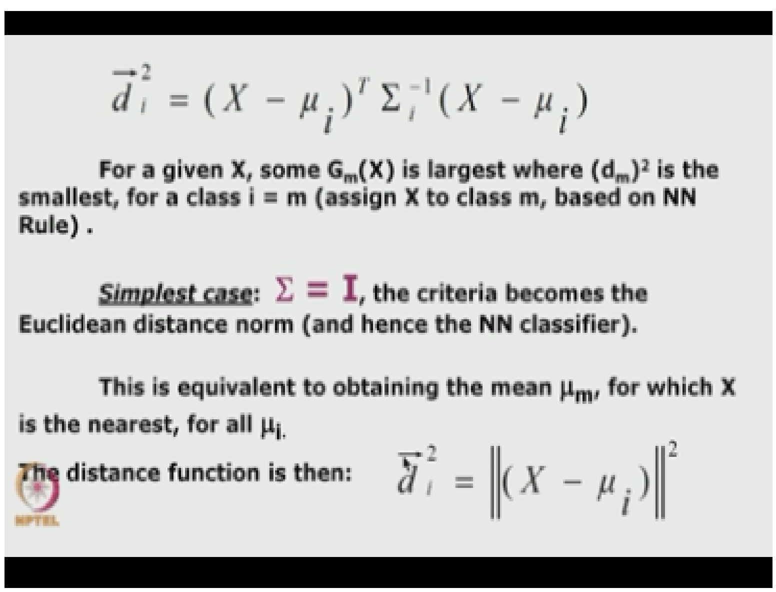
Let the discriminated function for the eyuth class be the same G of X so assuming class spires are same for two different classes I or J in fact if it is a two class problem but if it is arbitrary see different classes all the class plans are same that is what we are assuming then the only term which remains actually is this probability function and then based on this we can write a simple expression like this which you have done this is the term and we can write that this constant term here which you have actually broken in two parts are broken into two parts in the last slide is given by some constant Q.

Why it is constant because not a function of X anymore and this is a constant multiplied by a distance vector well distance usually a scalar okay distance usually a scalar okay so this is a scalar quantity mind you DI squares it is a norm of a distance what is basically taken and this case simple one by two it is a simple constant 1 by 2this factor and this DI is basically given by this expression here which is going to be the most important part of the discussion for the rest of this part of the core on distance measures with multivariate Gaussian functions for classification for decision boundaries estimation.

So the classification is now going to be astir because why this term is going to be only there for Gi this is a constant the classifiers classification is now influenced by the square of distance in hyper dimension space of X from me why weighted by the inverse of the covariance matrix and we're going to examine this a little bit more detail in the rest of the class today and henceforth and this term has already been introduced earlier may not be explicitly this is actually the Mahalanobis distance the distance from x to μ I in feature space weighted by the covariance matrix.

So it is a weighted distance if you are just take an X -μ which was the nearest neighbor classifier or the nearest neighbor classifier or minimum distance classifier then you did not pay attention to the class priors as well as the class PDFs of the class conditional probability density functions in such a case the Muhammad's distance take the simple role of a squared equilibrium distance but if it is weighted by the covariance matrix specifically. It is inverse to be precise then you are talking it as a Mahalanobis distance criteria for a quadratic term okay.

(Refer Slide Time: 22:22)



So look into the slide this is the one which are going to discuss now and so forth so it is all ball boiled down to the covariance matrix and the distance to the mean the rest of the class price they are relevant in certain examples but we will bring them back when we discuss a few things together a little bit later on.

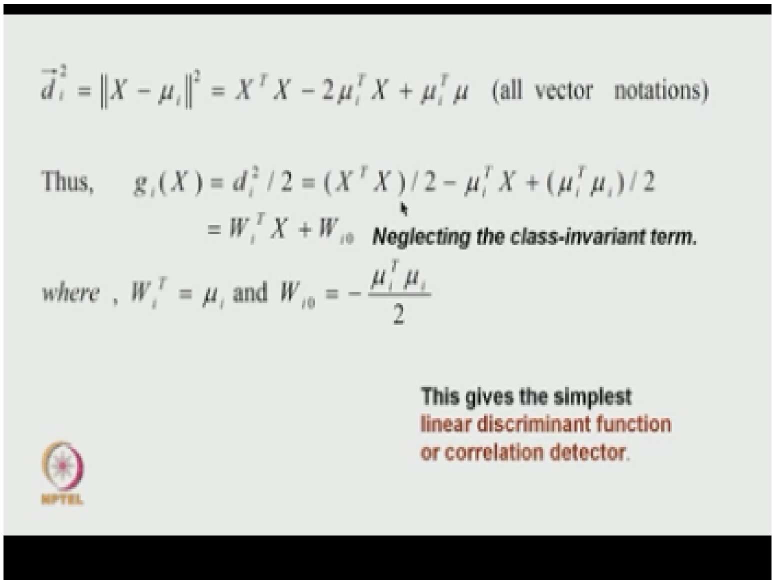
So for a given X some for some arbitrary I is equal to M the value of G well it is the same small G I have written in capital here is largest when the square of this distance is the smallest for a class I = M remember if you go back to the previous expression we are talking about a constant minus this K remember.

There is a minus 1/2 on the K so if the D value goes down the GI value will be maximum remember the class assignment rule assign it to the class for which the determinant function becomes the maximum so maximum value of GI if you look back to the expression the maximum value will be assuming it to be constant if this to be maximum this has to go down the distance is to go down the distance must be minimum for this to be maximized so that is what is written here so GM must be largest where this is the smallest the distance is the smallest for a particular class I = M.

And for that particular class M assign X to that particular class based on the nearest neighbor role okay the simplest case of course is that the covariance matrix is equal to an identity matrix I this criteria of course yield we have just been talking in the last few minutes that this yields our Euclidean distance norm the nearest neighbor classifier or the minimum distance classifier this is equivalent to just taking the mean of a particular class for which X is the nearest for all μ and the resistance function is very simply this there is just substitute an identity matrix.

Here you will have these two terms which actually will you this square norm of the distance from the in D dimension space of course easier going to be your d square but remember here all those d indicator has a vector but you are taking the square norm says basically the scalar quantity here.

(Refer Slide Time: 24:31)



Okay so in all vector notations this is how it can be expanded this is how it can be expanded you can write you can write expression in this particular form and then this G of X remember this is D square by 2 why this because there is a constant here which I am bringing in here for normalized sake of normalization and this is this is very simply substituting this overall by2 you will get this expression and what how can I write this in terms of this where look at this what term WI 0 which is taking care of this term the W transpose is this multiplied by X it seems I have been owning this strong.

If I ignore this term I can write this expression as something like this I can do this only when well I won't say X is negligible that is not the basic idea I am ignoring this term which is called a class invariant term it is a class invariant term why because if I keep changing the value of I this term does not change why am I trying to do this remember I am interested in classification discriminated function I have formed now.

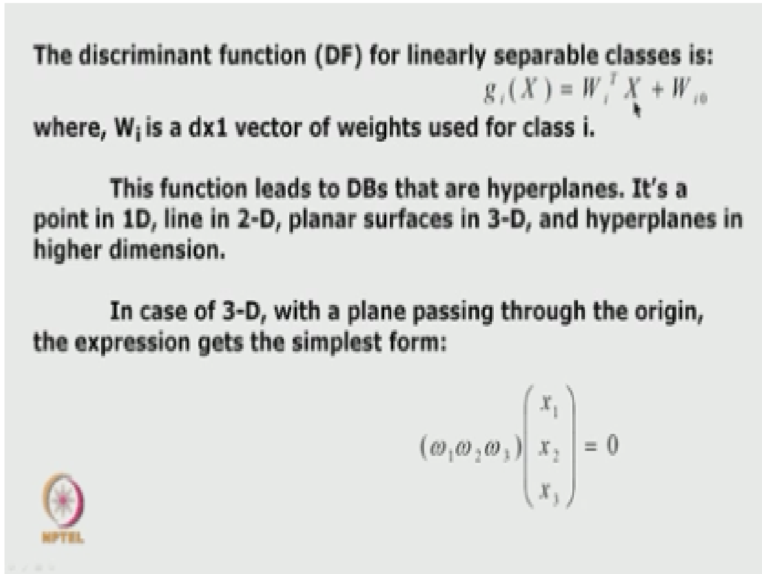
I am trying to find outreach discriminated function produces a decision region two discriminated, functions for two different values of subscript I and J they will create two different the dry sub decision region then there is a decision boundary between this pair I and J we will have a decision boundary between this pair I and J so if you move from class I to J to compute this G of I and look back into the expression what will change this expression here for the same sample X if you change I to J this is the term.

Which is going to change not this so hence this is called a class invariant table because it does not change between classes here the same value of x because there is a test sample which you are trying to classify after you have learned the discriminated function you move from class 1 to class 2 to class 3or Class A to Class B to Class C the first quadratic term which you see herein this expression this does not change across classes this does not have a subscript I it is the other two terms which are so I am ignoring this for the sake of comparing fix across classes across classes then.

I am having this and I am writing this the only thing which you need to worry about is that I have probably taken GI to be minus of this because the sign has been reversed okay so we what we are doing here is as if fix will be a constant - this part which is given here you can see that there is a negative sign which is been introduced here this negative sign has vanished so do not get this flip is not automatic it is the way you have taken this year because I want to maximize some function by minimizing the distance so this will help you in doing this is actually called in a very simple sentence.

I have liberalized the discriminated function this will actually give you linear decision boundaries this is also very casually called a correlation detector but that is not our main aim we will say that this is the simplest form.

(Refer Slide Time: 28:16)



Of a linear discriminantfunction the related discriminant function can also be written actually is this is the expression which we had in the last night if we go back let us see here this is what we had it is um it is sample x multiplied by a weight vector plus a bias term this is basically a bias term which base which is big which depends on the class mean I repeat again that this is the called the bias term which depends on the class mean and is the weight vector.

Which also depends on the class mean weight vector depends on the class mean and the bias term rule everything is depending on the class beans because the covariance term is ignored where did we ignore the covariance term we have taken the covariance matrix equal to identity matrix that is why we could write this in the first place and only which whatever remains is as the μ that is why we are able to get a linear expression of this so the linear discriminate function for a separable class is given by this which you have seen in the previous slide.

And the W vector is it d dimensional vector depending upon the dimensionality of the problem it is a vector of weights used for class I and we have the expression of W which is simply the class mean see this is itself a dimensional vector the same thing which holds good here this function leads to decision boundaries that are hyper planes in higher dimension and I have talked about this in the last class that it is a point in1d line in 2d planar surfaces in 3d and of course hyper planes in higher dimensions we will examine this linear addition boundary in very great detail okay and in the simplest case of a theory.

Which is a plane of course in 2d it will become a line it will become a pain passing through the origin and the expression gets in the simple form what has been written here is now basically this expression okay the W matrix is given by the three components here do not confuse these W eyes with the class levels which you have used probably earlier and the three-dimensional space ax is a three dimensional vector is a simple 0product here simple dot product here is what has been written in this particular case this is a plane passing through the origin but of course that is a special case you may have a line also passing.

Through the origin or somewhere in 2d space plane in three-dimensional space pearl place in an n-dimensional space once these weights are learnt is just simple and it is always okay, that I pickup these weights class means assign it to my W eyes assign the W bias term Wiser based on the me. A transpose and the classic above the classifier is done well in the case of linearly separable problems where problems can be separable by a linear hyper plane align in 2d or Apple in 3d this will work we will see that with an example.

Now we will see that with an example now we will take examples from geometry in two-dimensional space and understand the significance of this linear decision boundary which will lead us later on to an important concept of linear discriminate functions LD a linear discriminate analysis are a fairly a little bit later on not an immediate extrapolation right now because what we will do is we will learn the importance and significance of linear decision boundaries then bring in the which one we have ignored to get this linear decision boundary.

We have ignored the in the GI expression of GI we have ignore done important factor what was that factor the covariance matrix we will bring that covariance matrix we will see the significance of that covariance matrix which will might which into bring non-linearity into the picture of the decision boundary not only due to the diagonal elements but off diagonal elements and see some examples of those and then again come back to you in addition boundary using the official indiscriminant analysis our FLD a criteria DA which will lead us to the popular method of supervised classification called LDA a linear or fisher's linear discriminate criteria analysis that is going to be over all the discussion ends forth in the next few lectures we will stop here.

Online Video Editing /Post Production

K.R.Mahendra Babu

Soju Francis

S.Pradeepa

Camera
Selvam
Robert Joseph
Karthikeyan
Ram Kumar
Ramganesh
Sathiaraj

Studio Assistants
Krishankumar
Linuselvan
Saranraj

Animations

Anushree Santhosh
Pradeep Valan .S.L

NPTEL Web &Faculty Assistance Team

Allen Jacob Dinesh
Bharathi Balaji
Deepa Venkatraman
Dianis Bertin
Gayathri
Gurumoorthi
Jason Prasad
Jayanthi
Kamala Ramakrishnan
Lakshmi Priya
Malarvizhi
Manikandasivam
Mohana Sundari
Muthu Kumaran
Naveen Kumar
Palani
Salomi
Senthil
Sridharan
Suriyakumari

Administrative Assistant

Janakiraman.K.S

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