Deep Learning
Prof. Prabir Kumar Biswas
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Lecture - 06
Discriminant Function - I

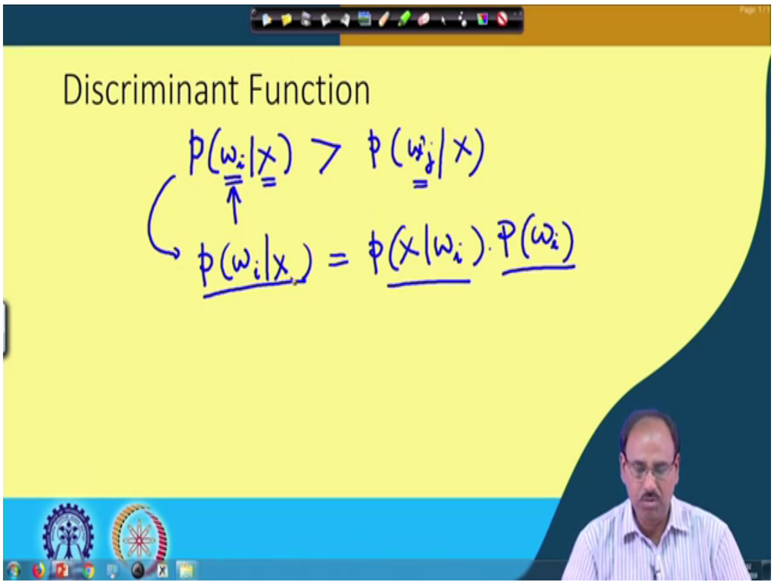
Hello welcome to the NPTEL Online Certification Course on Deep Learning.

(Refer Slide Time: 00:35)

In the previous class we have talked about topics like different types of pacifists like Bayes minimum error classifier and Bayes minimum risk classifier. And, we have also seen that Bayes minimum risk classifier under a specific case of 0 1 loss that is when the loss function for a correct decision is taken to be 0 and the loss function for an incorrect decision is taken to be 1 under that situation Bayes minimum risk classifier and Bayes minimum error classifier, they are identical.

So, in today's lecture we will talk about we will start from those Bayes classifiers and then we will move on to what is known as discriminant function. And, then using the Discriminant Function we will also try to derive and we will also try to demonstrate the decision boundary between different classes.

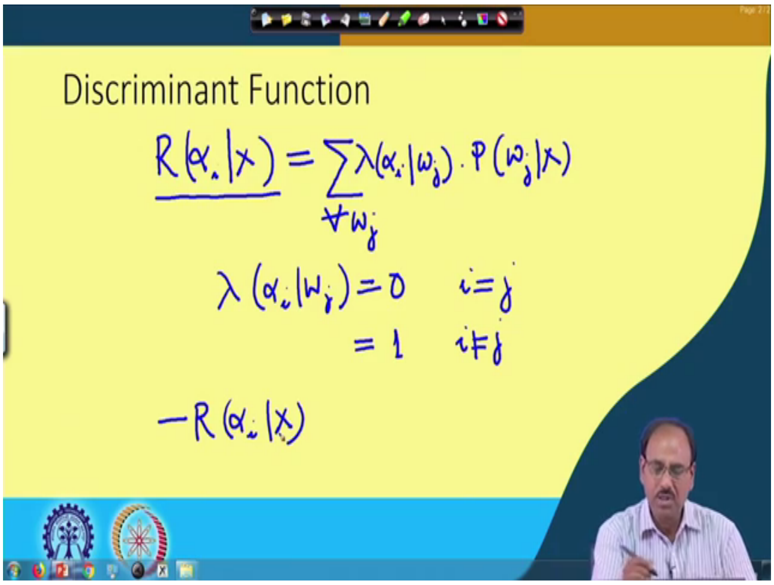
(Refer Slide Time: 01:43)



So, when we talk about a discriminant function you remember from the previous class in case of Bayes minimum risk classifier or Bayes minimum error classifier what we said is that for Bayes minimum error classifier if P of say omega i given X is greater than P of omega j given X where omega i and omega j are two different classes and X is the unknown input vector. In that case we classify the X to this class omega i and if I expand this P of omega i given X is nothing, but P of X given omega i multiplied by the a priori probability P of omega i, where P of X given omega i is what is known as the class conditional probability density.

P of omega i is the a priori probability and P of omega i given X is the a posteriori probability based on which we make the decision that whether this unknown vector X should be classified to class omega i or it should be classified to class omega j. So, obviously if P of omega i given X is greater than P of omega j given x, then it is more likely or more probable that your unknown feature vector belongs to class omega i. And, this is what we had derived in our previous lectures using Bayes minimum error classification rule.

(Refer Slide Time: 03:43)



And then Bayes minimum risk classification what we had said is for an unknown feature vector X, if we take an action alpha i, then the risk involved is given by R of alpha i given X. And, which we said that this is nothing, but lambda alpha i given omega j into P of omega j given x. You take the summation over this for all the classes omega j.

So, for every action alpha i if I have say c number of actions, so i varies from 1 to c. So, for every such action I have to compute this risk function. And. for whichever action the risk the estimated risk R of alpha i given X is minimum, I have to take that corresponding action. And, as we said that under a specific case when this lambda of alpha i given omega j is equal to 0 for i is equal to j and if I take this equal to 1 whenever i is not equal to j. That means, for every correct decision the loss incurred is 0 and for every incorrect decision the loss incurred is 1 under that situation we had shown in the previous class that Bayes minimum risk classifier and Bayes minimum error classifier, they turn out to be identical.

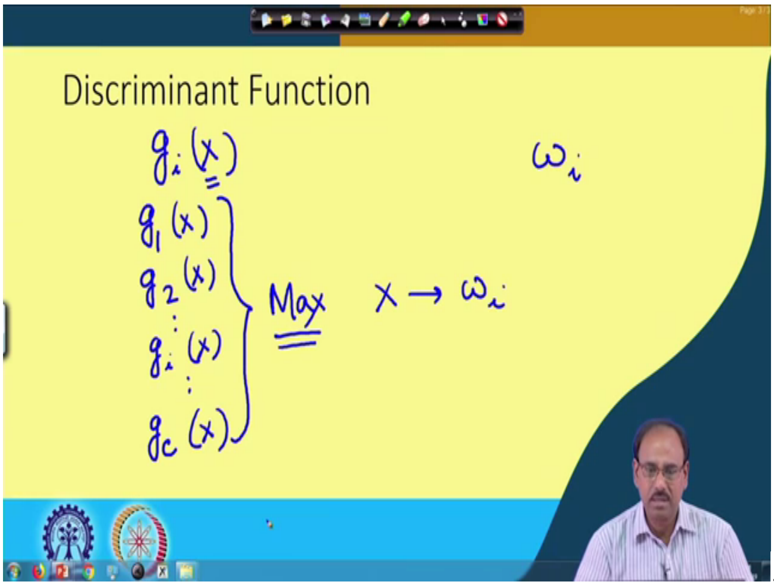
Now, starting from here we can define something called discriminant function because every time you find that whether I go for Bayes minimum error classification or Bayes minimum risk classification, in case of Bayes minimum error classification for every class I am computing P of omega i given x, where i varies from 1 to c where c is the number of classes that I have and for whichever i P of omega i given X turns out to be maximum i classify X to that corresponding class.

Similarly, in case of Bayes minimum risk classification for every class I compute R of alpha i given X and then for whichever class R of alpha i given X is minimum that is for whichever class for whichever action the risk involved is minimum, I am taking the corresponding action or I am classifying X to that to that corresponding class. And, I can say that in each of this case I am taking an action based on certain maximum criteria that is in case of Bayes minimum error classification for whichever class the a posterior ability is maximum, I am taking that corresponding action or classifying X to that corresponding class.

Similarly, for Bayes minimum risk classification for whichever class R of alpha i given x, so for whichever value of i are all R of alpha i given X turns out to be minimum or in other case I can say that instead of considering R of alpha i given X, I will consider minus R of alpha i given X. So, if R of alpha i given X is minimum, then obviously R of alpha I given X minus R of alpha i given X will be maximum.

So, for whichever action this negative of the risk of value turns out to be maximum, I am taking that corresponding decision or I am classifying X to that corresponding class.

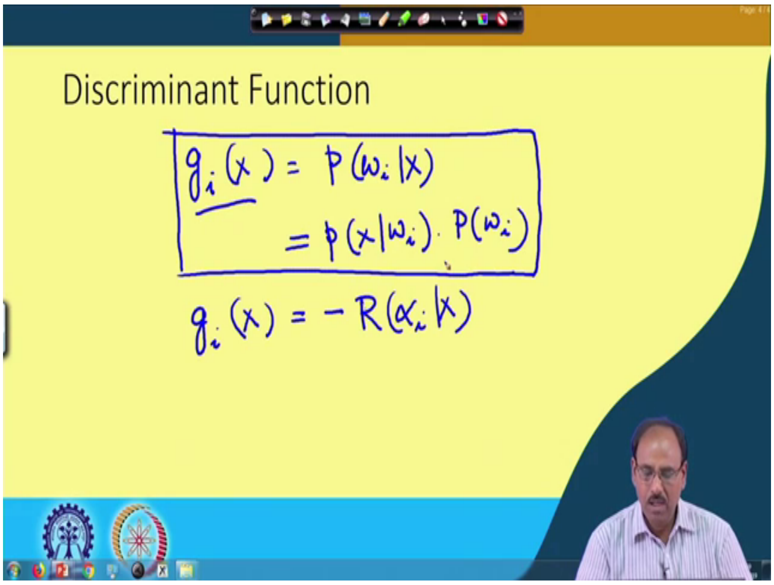
(Refer Slide Time: 08:00)



So, or in other words I can say that I can define a function say g i X for class say omega i. So, here X is the unknown feature vector and for every class I every class omega i, I am computing a function g i X and for whichever i this g i X turns out to be maximum I take decision in favor of that particular class or that particular omega i.

So, what I am doing is for an unknown feature vector X I will compute g 1 of X, I will compute g 2 of X, I will compute g i of X and if there are c number of classes I will compute g c of X. And, then I will try to find out that out of all these functional values whichever is maximum. So, I take maximum of all of this and for whichever i this turns out to be maximum I classify X to that corresponding class omega i. So, this is a function that g i X i want to design for every class omega i. So, what are the possible options that I can have g i X?

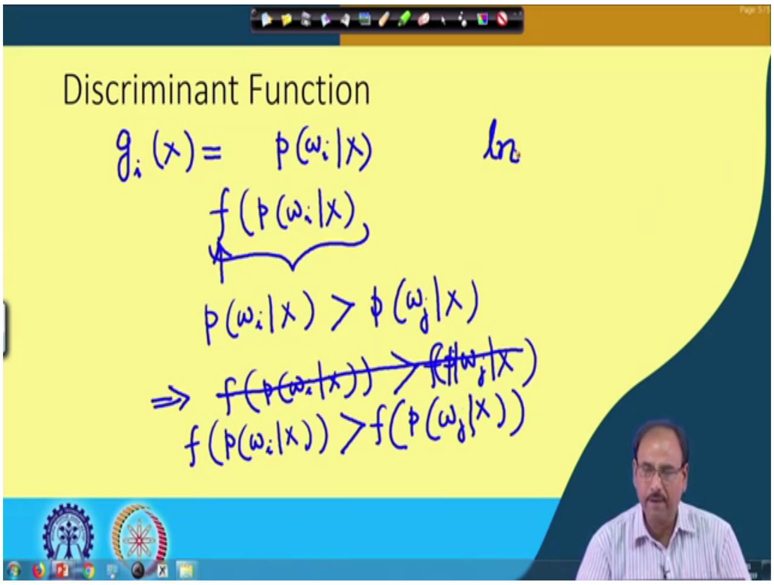
(Refer Slide Time: 09:29)



One of the option is obviously I can have g i X to be is equal to P of omega i given X that is staright forward which is nothing, but P of X given omega i into a priori probability P of omega i. So, this is a straightforward definition of g i X or I can also say that I will use g i X to be minus R of alpha i given x.

So, here also if g i X is maximum I take that corresponding decision here also if g i X is maximum I take that corresponding decision. So, out of these two options possible options because there might be other options as well. Out of these two we will try to explore this that is based on Bayes minimum error classification rule.

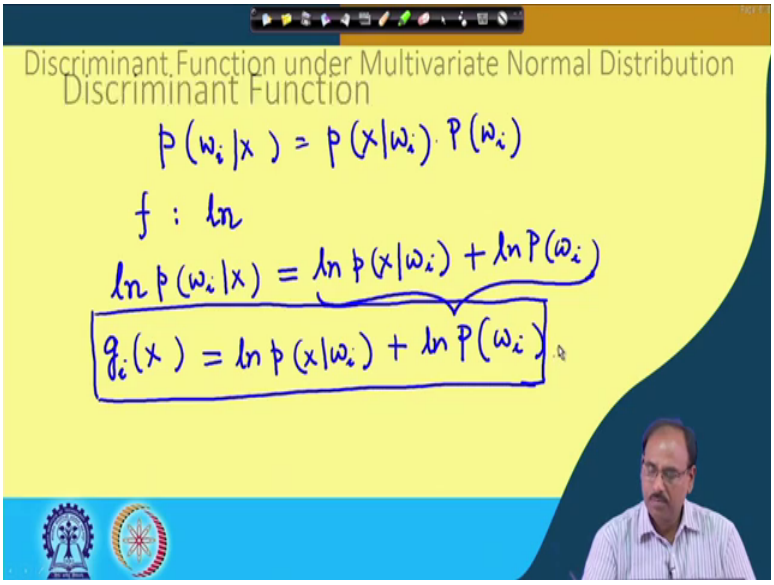
(Refer Slide Time: 10:37)



So, I will assume that we will use this discriminant function g i X where this discriminant function is defined as P of omega i given X or it is also possible that instead of P of omega i given X if I use a function of P of omega i given X where this function f has to be a monotonically increasing function. That means, if P i given X is greater than P of omega j given X, this should imply the f of P of omega i given X should be greater than f of P of omega j given x. Let me rewrite this. So, this implies f of P of omega i given X has to be greater than f of P of omega j given x.

So, that is f has to be a monotonically increasing function, then this form that is g i X as f of P of omega i given X that can also be used as a discriminative function because whenever g i X is maximum, the f of P of omega i X will also be maximum. So, given this you will find that one very convenient function that can be used is logarithmic function. So, at or I can use natural logarithm l n. What is the advantage?

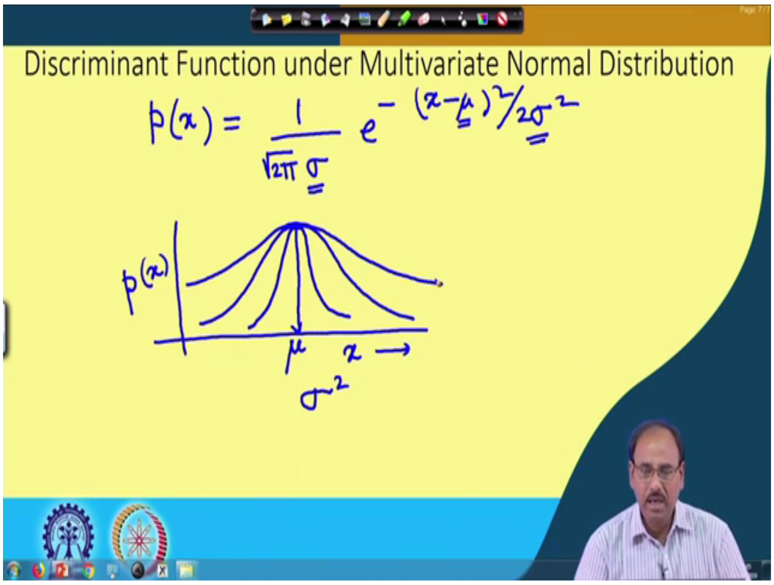
(Refer Slide Time: 12:39)



The advantage is because P of omega i given X is nothing, but P of X given omega i which is the class conditional probability density function that can be estimated experimentally into P of omega i which is the a priori probability of class omega i that is also pre-computed. And, now if this f the function I used as the logarithmic function ln, then the advantage that I get is l n P of omega i given X that turns out to be l n P of X given omega i plus log of P of omega i. So, this multiplication state way is converted to an addition operation and which is very very advantageous in many computational purposes.

So, I will use this particular form that g i X is nothing, but log of P of X given omega i plus log of a priori probability P of omega i. So, this is the discriminant function form that we will use in the remaining part of this lecture, right. So, given this let us try to see that if I assume a particular form or a particular distribution function probability density function which usually we use as a normal probability density function, then what form of expression of the discriminant function that we get or what form of the decision boundary between classes that we get?

(Refer Slide Time: 14:54)

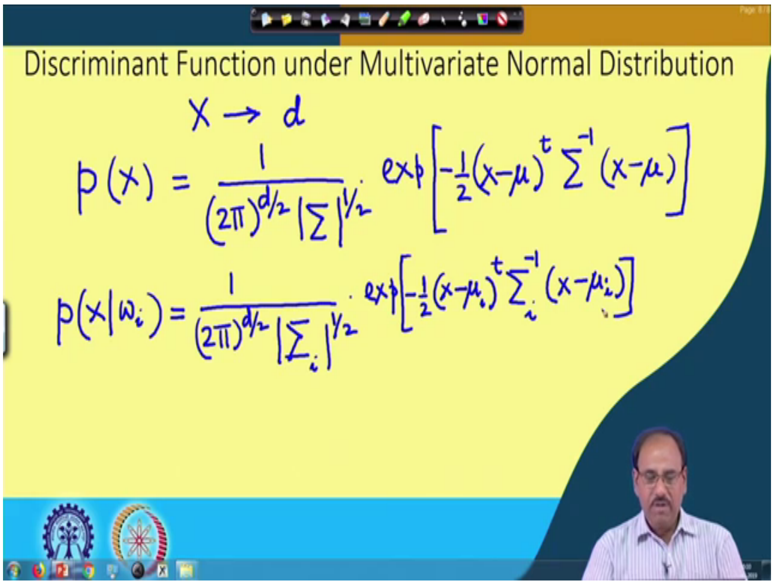


So, let us talk about this discriminant function under multivariate normal distribution. So, we all know that the normal distribution if I have a single variable say p x is given by 1 over square root of 2 pi sigma e to the power minus x minus mu square upon 2 sigma square. This is the normal density in case of a single variable x or scalar variable x where this sigma is nothing, but standard deviation and sigma squared is the variance and mu is the mean.

And, you all know that the typical form of this is if I plot x and p x, the typical form is like this where this is what is your the mean of x that is mu and the value of this envelope depends upon the value of sigma or sigma square.

So, if sigma is lowered, the sigma square is low then I will have a distribution something like this. If the sigma squared is high, then the distribution will be flat of this form. So, this is the form that I get when x is a single variable or it is a scalar variable.

(Refer Slide Time: 16:41)



But in our case since we are talking about feature vector which describes an object and the feature vector consists of multiple number of features where every feature captured some property or some attribute of the object. So, those features may be computed from the shape of the object, they might be computed from the color of the object. They might be computed from the intensity of the object, they might be computed from the texture of the object and various such different properties are put together in the form of a vector or a feature vector. So, the type of distribution that is important in our case is not a single variant distribution, but it is a multivariate distribution.

So, in our case I have a feature vector X and let me assume that the dimension of the feature vector is d. So, it is a d dimensional feature vector right; d is the number of components or the number of features which are packed into this feature vector X. So, given this the multivariate property density is now given by P of X which is 1 over 2 pi to the power d by 2, then instead of variants now I have multiple variables. So, what I have is a covariance matrix.

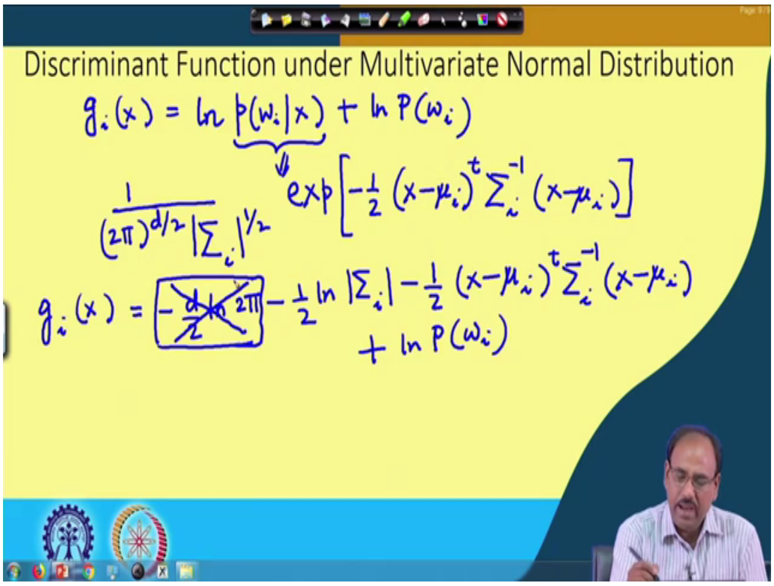
So, sigma is the covariance matrix. You take the determinant of that and square root of the determinant into exponential minus half X minus mu transpose sigma inverse into X minus mu that is what is the normal distribution form of normal distribution in case of multivariates or in case of vectors.

Now, what we are interested in or the expression that we have that contains X given omega i. That is the class conditional probability density and we said earlier that we get this class conditional probability density by taking the feature vectors X from class omega i. So, when I take feature vectors X from class omega i, so for those feature vectors the mean that I will get is dependent and I will represent that by mu i. Similarly the covariance sigma, the covariance matrix sigma that I compute will also be on for that particular class omega i. So, I will also represent this as covariance matrix sigma i.

So, what I will do is I will put this class conditional probability density function, express it in the form 2 pi to the power d by 2. Now, this sigma actually becomes sigma i because this is for class omega i square root of that into exponent minus half X minus. Now, this mu becomes mu i. It is for ith class transpose sigma becomes sigma i. So, it is sigma i inverse into X minus mu i. So, this is my multivariate probability density function where, sigma i is the covariance matrix computed over all the feature vectors which we call as feature vectors because using those vectors, I am computing sigma i and mu i.

So, it is the covariance matrix computed using those feature vectors taken from class omega i mu i is the mean of those feature vectors taken from class omega i.

(Refer Slide Time: 21:07)

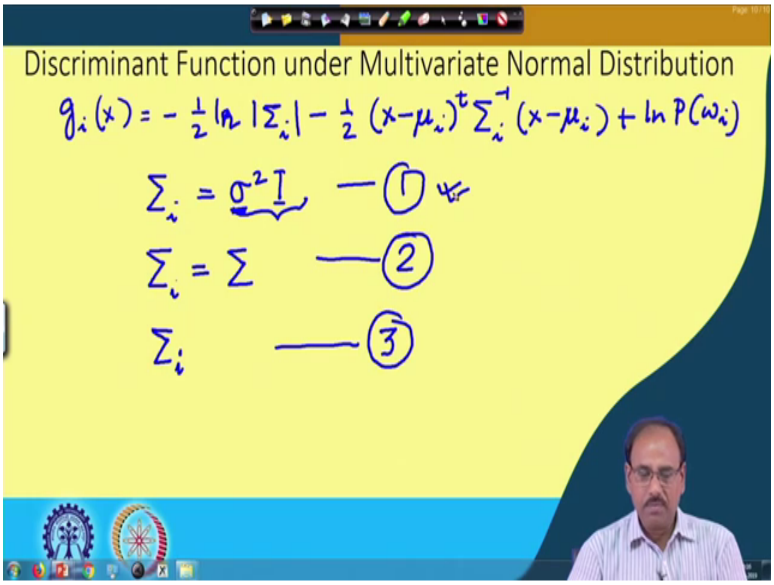


Now, given this the way we have defined g i X is equal to log of P omega i given X plus log of a priori probability P of omega i. Now you find that P of omega i given X is nothing, but 1 over 2 pi to the power d by 2, then sigma i square root of this exponential minus half X minus mu i transpose. So, this is what is P of omega i given X.

So, once I use this logarithm, then my g i X it simply becomes minus d by 2 log of 2 pi minus half log of sigma i minus half of X minus mu i transpose sigma i inverse X minus mu i plus of course I have this log of P of omega i. From here you will find that d by 2 log of 2 pi this particular term is independent of the class because there is no term like subscript i over here.

So, this minus d by 2 log of 2 pi, this does not differentiate between an ith class and jth class. So, easily I can conveniently ignore this particular term minus d by 2 log of 2 pi.

(Refer Slide Time: 23:47)



So, that simplifies my g i X as minus half log of P sigma i minus half X minus mu i transpose sigma inverse X minus mu i plus log of a priori probability P of omega i. Now, here I can have different cases; say for example this covariance matrix sigma i in a particular case in a specific case if all the components the components of the feature vector X, they are statistically independent. Then covariance matrix that we get will be a diagonal matrix and if every component has same variance, then this covariance matrix sigma i will be of the form sigma squared I.

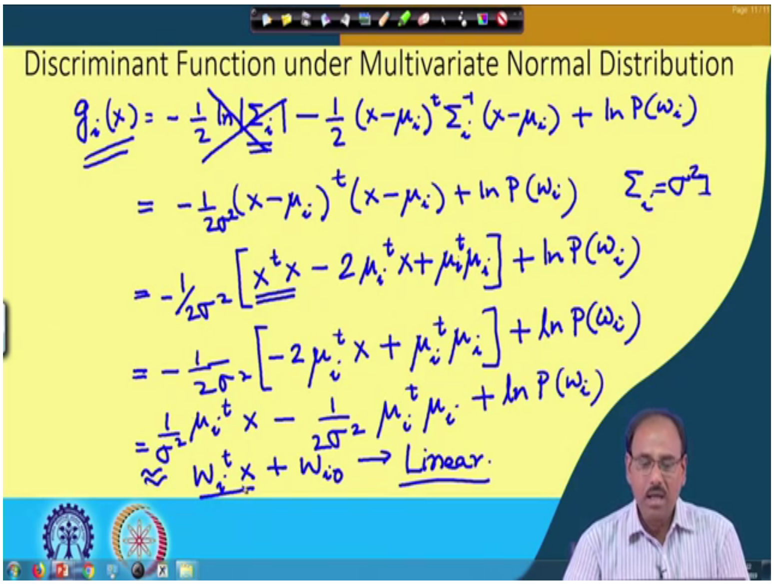
So, what I am assuming here that for all the classes the feature vectors that you obtain the components of the feature vectors are statistically independent. So, that means if I try to compute the variance involving say ith component and jth component because they are statistically independent. So, that variance will be equal to 0 which leads to the covariance matrix to be a diagonal matrix, where only I will have diagonal elements to be non-zero and all the off diagonal elements will be 0. And, then again if I assume that all those components, for all those components the variance is same in that case all the diagonal elements which are non-zero, they will be equal.

So, that ultimately leads to the covariance matrix to be of the form the sigma square I and I am assuming this to be same for all the classes. That means, for every sigma I, I have this covariance matrix for every omega i, the covariance matrix sigma i is of the form sigma square I. So, the sigma squared is again same for all the features across the classes. So, that is one of the simplified assumption that I can make the other assumption that can be used is where sigma i is of the form sigma.

So, in this case it is not necessary that the different components of the feature vector will be statistically independent, not even necessary that every component will have the same variance, but what I am assuming is that whatever is the covariance matrix, the same covariance matrix is valid for all the classes. So, this is a simplified condition, condition 2 and the third one where I have the most general case that every class will have its own covariance matrix that is the covariance matrix of one class need not be same as covariance matrix of other classes. So, that is the most general case which is case 3.

So, initially I will try to see that how this discriminant function look like when I assume the first case that is covariance matrix of every class is of the form sigma square I.

(Refer Slide Time: 27:58)



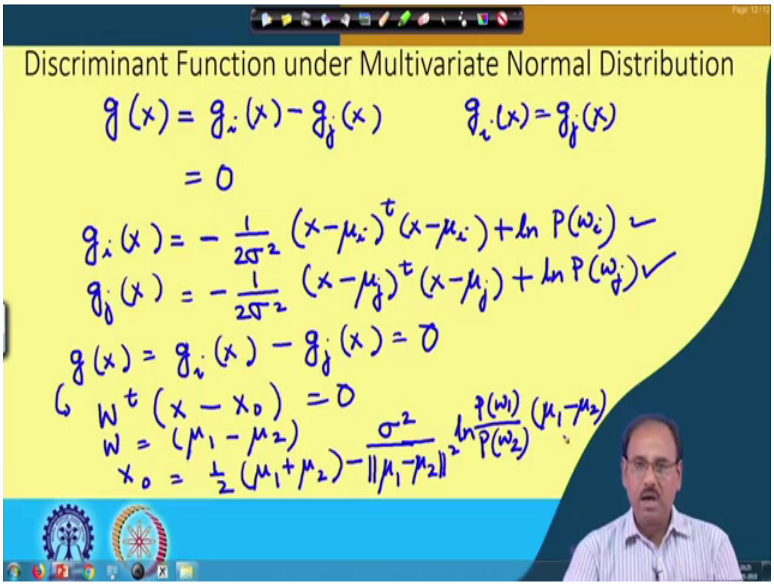
So, let us see this. So, what I have is g i X is equal to minus half log of sigma i minus half X minus mu i transpose sigma i inverse X minus mu i plus log of P of omega i. So, here as I am assuming that this sigma i the covariance matrix is same for all the classes. So, this minus half log of determinant sigma i, again this does not have any role in discriminating among different classes. So, I can simply ignore this term from the function of from the expression of the discriminant function.

So, my g i X now simply becomes minus half X minus mu i transpose and sigma i because it is sigma square I. So, the sigma i is of the form sigma square I. So, this sigma i inverse is simply 1 upon sigma square. So, what I will do is I simply put it as 1 upon 2 sigma squared X minus mu i transpose into X minus mu i plus log of P of omega i which simplify comes minus 1 upon 2 sigma square. If I expand this it becomes X transpose X minus twice mu i transpose X plus mu i transpose mu i plus log of P of omega i. In this expression again this X transpose X is class independent, right. So, again this term does not contribute to discrimination.

So, I further simplify this as minus 1 by 2 sigma squared. What I have within the bracket is minus 2 mu i transpose X plus mu i transpose mu i plus log of P omega i. You will simplify this, it simply becomes mu i transpose 1 upon sigma squared mu i transpose X minus 1 upon 2 sigma squared mu i transpose mu i plus log of P of omega i which I can write in the form W i transpose X plus W i naught where this W i is nothing, but 1 upon sigma squared mu i and W i naught is 1 upon 2 sigma squared minus 1 upon 2 sigma squared mu i transpose mu i plus log of P of omega i.

So, find that the expression that you get is a linear expression. That means, under the simplified case when all the components all when the components of the feature vectors are statistically independent, all the components have the same variance sigma square and this is same for all the classes. Or, in this particular case I am not assuming it is same for all the classes, but I am considering only a particular class omega i. The g i X is simply of the form of W transpose X plus W i transpose X plus W i naught which is a simply a linear expression, right. So, from here I can try to find out that what is the boundary between two different classes omega i and omega j.

(Refer Slide Time: 33:19)



So, in order to do that let me again try to find out. So, that boundary I can simply defined as g X and on a boundary I must have g i X is equal to g j X. That is the discriminant functional value for ith class and for jth class; they should be same on the boundary.

So, the equation of the boundary can simply be written as g X is equal to g i X minus g j X which is equal to 0. So, this is simply the equation of the boundary between two classes omega i and omega j. And, what we have seen is that for g i X under this simplified assumption, we have seen that g i X is nothing, but minus 1 upon 2 sigma squared into X minus mu i transpose X minus mu i plus log of P of omega i. Similarly for g j X I will also have the case that it is minus 1 upon 2 sigma squared into X minus. Now it will be mu j transpose X minus mu j plus log of P of omega j.

And, if I equate these two if I make g X is equal to g i X minus g j X to be equal to 0, then we will find that by putting g i X, this expression and in place of g j X this expression you will find that this g i X equal to 0. This will take a form W transpose X minus X naught is equal to 0 where, you will find that this W is nothing, but mu 1 minus mu 2. And, X naught will be same as half of mu 1 plus mu 2 minus sigma square upon mod of mu 1 minus mu 2 square into log of P omega 1 upon P omega 2 into mu 1 minus mu 2.

So, this is the expression that I will get for the boundary between the two different classes. So, I will derive the expression of this boundary under this simplified case in our next lecture.

Thank you.