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

Lecture - 07
Discriminant Function - II

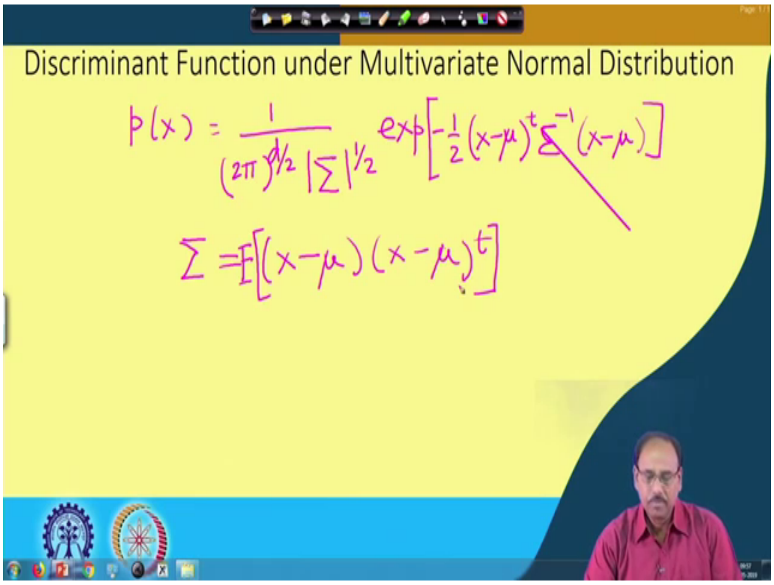
Hello, welcome to the NPTEL online certification course on Deep Learning.

(Refer Slide Time: 00:33)

You remember that in the previous class we had started our discussion on Discriminant function and the boundary between different classes. So, in the previous class we talked about the discriminant function under multivariate normal distribution and in today's class we are going to continue with our previous discussion and we will talk about, we will see that how the decision boundary under different tosses under various conditions of the covariance matrix we can have.

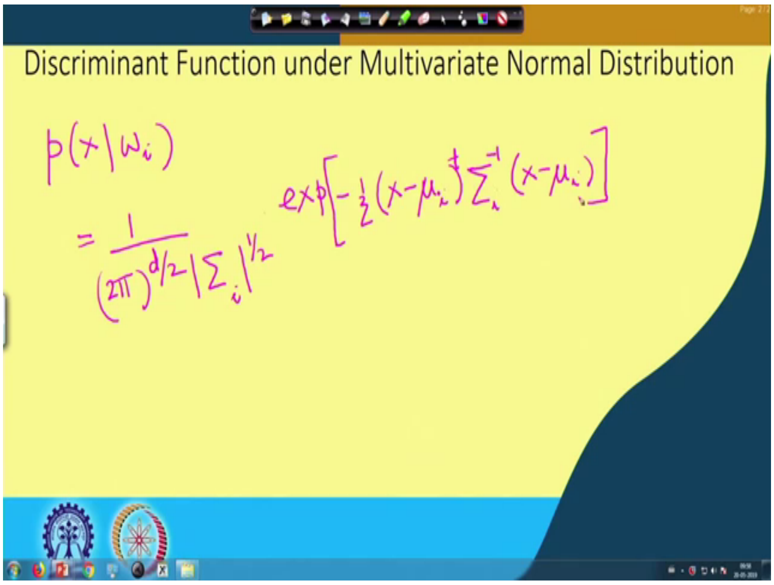
So, we can have linear boundaries, we can also have non-linear boundaries or quadratic boundaries and this will also illustrate with the help of some examples.

(Refer Slide Time: 01:24)



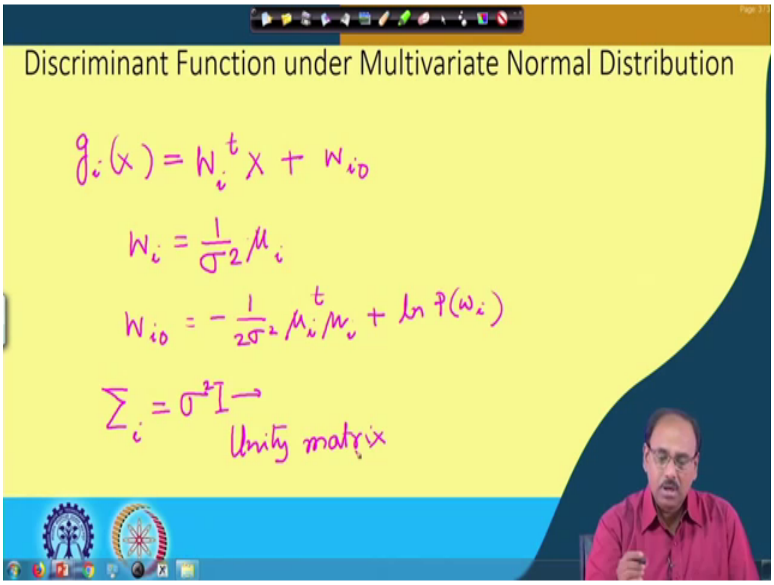
So, let us just try to start with what we had done in our previous class. So, in the previous class we have said that the multivariate normal distribution is given by p of X as 1 over 2 pi to the power d by 2 then sigma determinant to the power half and then exponential minus half X minus mu transpose sigma inverse into X minus mu. So, this was the normal distribution under multivariate case. What this covariance matrix sigma is nothing, but X minus mu into X minus mu transpose and the expectation value of this that is what is the covariance matrix and mu is the mean of all the samples all the vectors that we have.

(Refer Slide Time: 03:04)



So, now if we want to have represent this as class conditional probability density function, multivariate normal density function, then whatever they do is I have to get p of X given omega i which will be 1 upon 2 pi to the power d by 2 and as we said in our previous class that d is the dimensionality of the vector and then sigma now becomes sigma i because it is the covariance matrix of all the samples taken from class omega i and the other part remains exponential minus half X minus. Now mu becomes mu i because this is mean of the vectors taken from class omega i sigma becomes sigma i. So, this is sigma i inverse there of course transpose into X minus mu i. So, this is the cross conditional normal distribution.

(Refer Slide Time: 04:30)



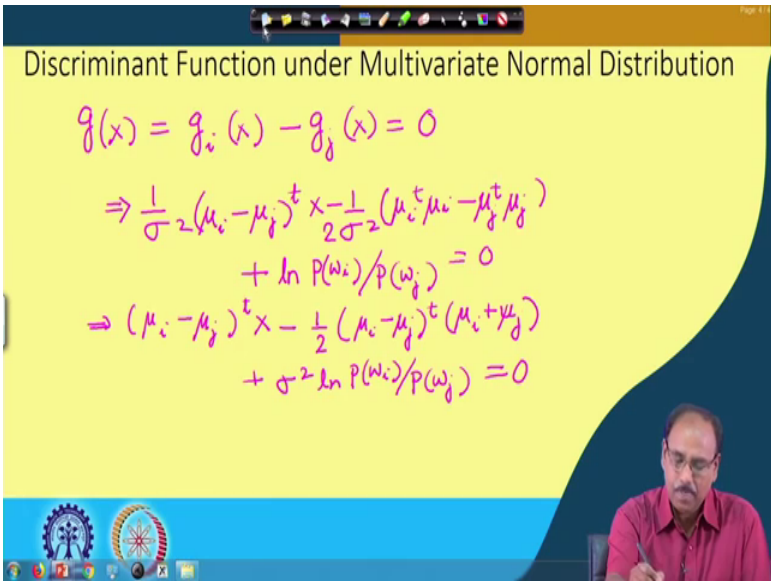
So, using this we have in our previous class found out that discriminant function for the i th class where the discriminant function was obtained as g i X that is the discriminant function for the i th class which was w i transpose X plus W i naught, where what is this W i, we had seen that this W i is nothing, but 1 upon sigma square into mu i and W i naught was minus 1 upon 2 sigma square mu i transpose mu i plus log of P of omega i. And this is the expression we have obtained under the assumption that sigma i that is the covariance matrix for all the classes is same and which is of the form sigma square I, which indicates that the covariance matrix is a diagonal matrix or all the off the diagonal elements are 0.

And all the diagonal elements of the covariance matrix is same as sigma square. What does it physically mean? It physically means that the different components of the feature vector that we have those components are statistically independent and every component has the same variance that is the variance of X 1.

The first component is same as the variance of X 2 which is the second component like this and every component the variance is same which is nothing, but sigma square. This indicates that the way the training vectors, the vectors are distributed in the feature space is like circle in case of two dimension, it is a sphere in three dimension and it is hyper sphere in multi dimension; where of course at the center that density is maximum and as you move away from the center, the density goes on reducing.

So, this is the physical significance of this sort of distribution where sigma i is of the form sigma square i where this i is nothing, but our unity matrix.

(Refer Slide Time: 07:16)



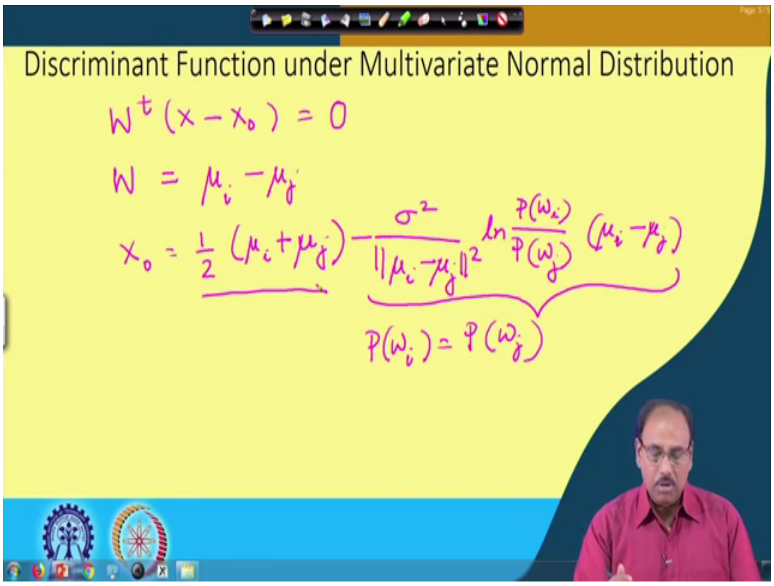
So, under this situation we have also stated in the previous class that if I have feature vectors, they slip test coming from two different classes the i th class and j th class I can find out g i X that is the discriminant function for the i th class, I can also find out g j X which is the discriminant function corresponding to the j th class. And given this if I want to find out what is the boundary, the decision boundary between these two classes that is we say that if the feature vectors falls on one side of the boundary, it belongs to one class I am it and if it belongs to falls on the other side of the boundary it belongs to some other class.

So, I can find out an expression of the boundary which separates these two classes and the expression of the boundary will be simply given by g X which is nothing, but g i X minus g j X and because on the boundary g i X and g j X both of them will be equal. So, this has to be equal to 0. So, our expression will be g X is equal to g i X minus g j X which will be equal to 0.

And this just from the expression of g i X that we have obtained you can find that g i X minus g j x, this will be simply given as 1 upon sigma square into mu i minus mu j transpose X minus 1 upon sigma square mu i transpose mu i minus mu j transpose mu j plus log of P omega i upon P omega j that will be equal to 0.

This can simply be put in the form mu i minus mu j transpose X minus this. There has to be a half right half of mu i minus mu j transpose into mu i plus mu j plus sigma squared log of P omega i upon P omega j that has to be equal to 0. And after simplification, you can find that this expressions can simply be written as.

(Refer Slide Time: 10:33)



W transpose X minus X naught that has to be equal to 0 where this W will be simply mu i minus mu j and X naught will be given by half of mu i plus mu j minus sigma square upon mu i minus mu j mod squared log of P omega i upon P omega j into mu i minus mu j.

So, see what is the significance of this particular expression. It says the equation of the boundary between the two classes i th class and j th class is given by w transpose X minus X naught equal to 0 which simply indicates that this boundary is a linear boundary. In case of three dimension it is a plane, in case of multi dimension it is a hyper plane. For the vector W is given by mu i minus mu j which is nothing, but a vector drawn from mu i to mu j where mu i is the mean of the vectors taken from class omega i and mu j is the mean of the vectors taken from class omega j and the expression for X naught and it says that because W transpose X minus X naught equal to 0.

So, the decision surface is obviously perpendicular orthogonal to the vector W and as W is the vector drawn from omega i to omega j. So, the decision surface is orthogonal to the line joining mu i and mu j and it passes to the point X naught. And here if you look at the expression for X naught and particularly under the case if I consider that mu P of omega i is equal to p of omega j, if I can use if I consider this condition that is both the classes omega i and omega j they are equally probable, the a priori probability is same under this case log of P omega i upon P omega j will be equal to 0. So, this term will be equal to 0 and in which case your X naught simply becomes half of mu i plus mu j. That means, X naught is a point which is at the middle of the vector the line joining mu i and mu j.

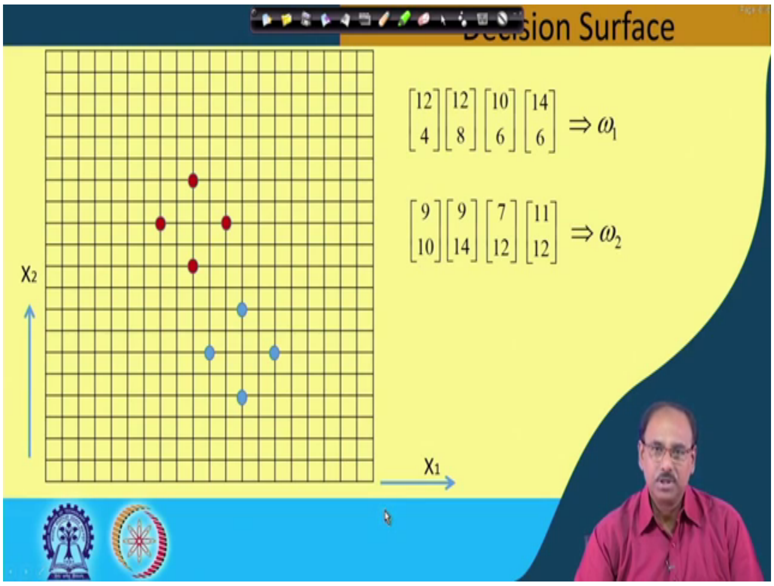
So, under such circumstances where the a priori probabilities of the two classes are same, your decision surface becomes an orthogonal bisector of the line joining mu i and mu j. And obviously, the kind of decision that you take in this case that given an unknown vector whether the vector should belong to class omega i or the vectors should belong to class omega j, that decision will be taken by simply taking the distance of that unknown vector from the mean vectors mu i and mu j.

So, whichever mean is nearest to this unknown vector x, the unknown vector will be classified to that corresponding class. And obviously, if p of omega i and p of omega j they are different that is a priori probabilities are different, that will give a bias in the decision. So, if a priori probability for omega i is greater than a priori probability for omega j, then your decision surface though it will be orthogonal to the line joining mu i and mu j, but it will be shifted towards mu j indicating that more space is allocated to class omega i.

Similarly, if mu j P of omega j is greater than P of omega i, in that case X naught will be shifted towards mu i indicating that more of space will be allocated to class omega i in which case your decision will be biased in favor of class biased in favor of class omega j.

So, now let us try to see an example that how this really works.

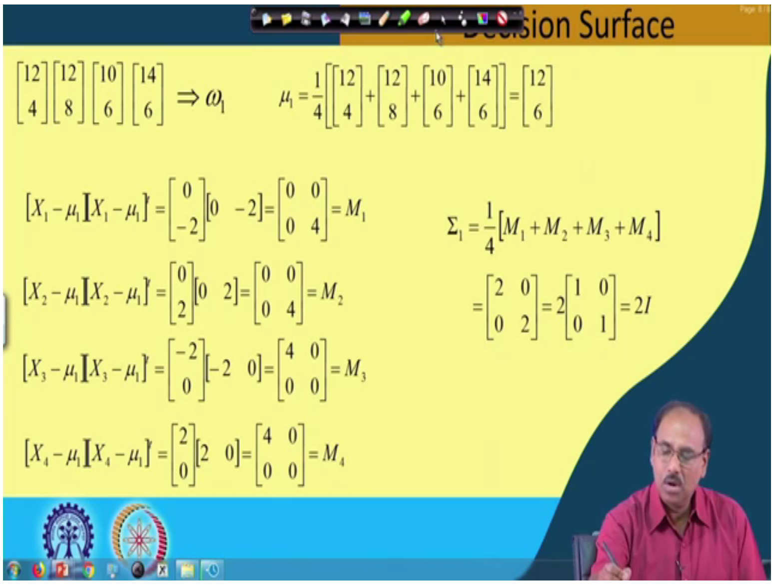
(Refer Slide Time: 15:31)



So, to illustrate this I take a set of points both from class omega 1 and class omega 2. So, first I take a point say 12 4. So, I am considering a two dimensional feature space and the features are say X 1 and X 2. So, first I considered a point 12 4, then I consider a point 12 8, then I consider a point 10 6 and say 14 6 and I assumed that these are the points which are taken from class omega 1.

Similarly, I also consider another set of points say 9 10, 9 14, 7 12 and 11 12 and I consider these points to be taken from class omega 2. And these are the points, so we call as training samples because using these feature vectors I am going to train my classifier and that is where what is the learning in this particular case. So, I have a set of points from 1 class omega 1 and another set of points from class omega 2.

(Refer Slide Time: 16:46)



Now given this set of points first I consider the points taken from class omega 1, I compute the mean vector mu 1 which is simply average or mean of all these vectors taken from class omega 1 and then, that mean vector comes out to be 12 6 and once I have this mean vector, then I have to compute the covariance matrix. And as we said that covariance matrix is nothing, but the expectation value of X minus mu into X minus mu transpose.

So, if I consider the first vector I call it say X 1 that is vector 12 4, I subtract the mean 12 6 from this. So, 12 minus 12 6 simply becomes 0 minus 2. So, a part of the partial covariance matrix simply X minus mu 1 into X minus mu 1 transpose in this case as is shown here X minus mu 1 into X minus mu 1 transpose simply becomes 0 0 0 4. So, this 0 minus 2 as column vector is my X minus mu 1 and 0 minus 2 as row vector is X minus mu 2.

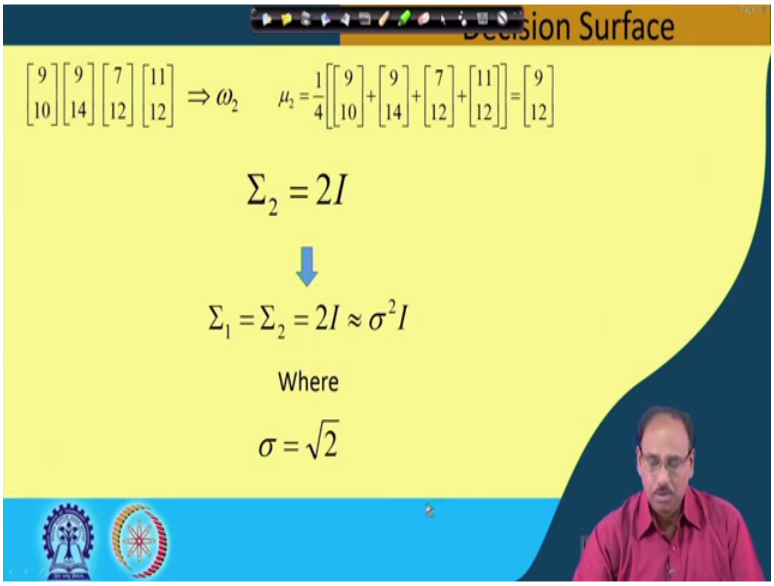
In the same manner I also compute X 2 minus mu 1 into X 2 minus mu 1 transpose which is m 2. That again comes out to be 0 0 0 4. I also compute X 3 minus mu 1 into X 3 minus mu 2 transpose; mu 1 transpose and that comes out to be 4 0 0 0.

Similarly, m 4 which is X 4 minus mu 1 into X 4 minus mu 1 transpose and that comes out to be again 4 0 0 0. So, once I have these four matrices now the covariance matrix is nothing, but the mean of all these four matrices.

So, the covariance matrix from class for class omega 1 which is sigma 1 is simply 1 upon 4 into M 1 plus M 2 plus M 3 plus M 4 and if you compute this, it will simply come out to be 2 0 0 2 which is nothing, but matrix of the form 2 into I where this I is an unity matrix.

So, this is what I get for class omega 1.

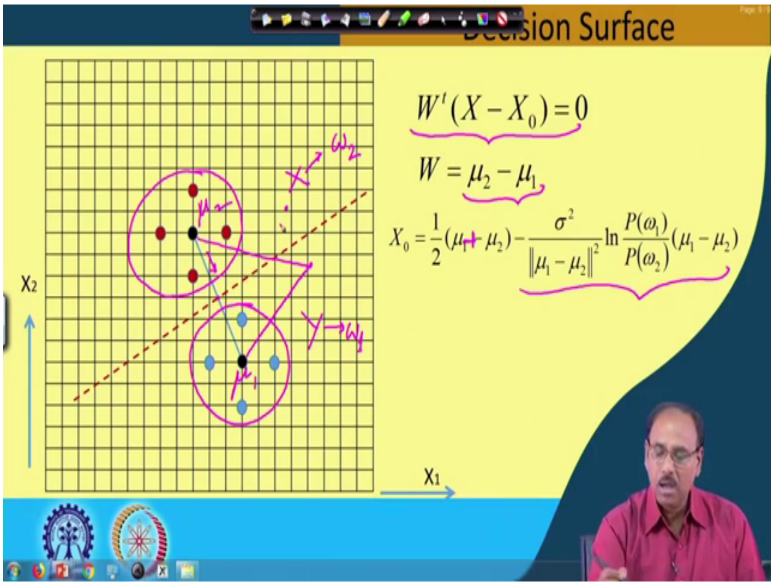
(Refer Slide Time: 19:38)



Similarly, for class omega 2, I consider all those feature vectors that have taken from class omega 2 which you remember that the feature vectors were 9 10, 9 14, 7 12 and 11 12. So, using these four feature vectors I compute the mean of the vectors belonging to class omega 2 and which in this case comes out to be 9 12. And once I have these mean vectors in the same manner as I have done previously, I am not repeating all those calculations. You can calculate the same way and you can find out that the covariance matrix for class omega 2 will come out to be sigma 2 which is nothing, but 2. I again I is the unity matrix.

So, here you find that if you remember that for class omega 1, we had sigma 1 which was 12 and for class omega 2, I have sigma 2 which is also 2 I. So, this is a case where as we said earlier that sigma i is equal to i sigma square I. So, it is the same condition and under which case we have seen that the discriminant functions will be linear and not only that the decision boundary between the two classes omega i and omega j that will also be linear. So, this is a perfect case of that where I have sigma 1 and sigma 2 to be equal and that is 2 into i which is of the form sigma square I. And what is sigma square? Sigma square in this case is 2. So, sigma is square root of 2. So, given this calculations now let us see how the decision boundary will look like.

(Refer Slide Time: 21:34)



So, come to the same set of points as we shown earlier that these points are belonging to class omega 2 and these were the set of points which are taken from class omega 1. And we have shown earlier that when the covariance matrix sigma i is of the form sigma square i for all the classes, the decision boundary takes the form w transpose X minus X naught equal to 0 where W is nothing, but mu 2 minus mu 1. That is the vector drawn from omega 2 to omega 1 right or omega 1 to omega 2. If I compute the other way that is g 1 X minus g 2 X here the computation was g 2 X minus g 1 X.

And the middle point will be given by there is a mistake it is not mu 1 minus mu 2 that it, but it should be mu 1 plus mu 2. And the midpoint the point on this decision boundary is given by X naught is equal to half of mu 1 plus mu 2 minus sigma squared mu 1 minus mu 2 square into log of P of omega 1 upon P of omega 2 into mu 1 minus mu 2.

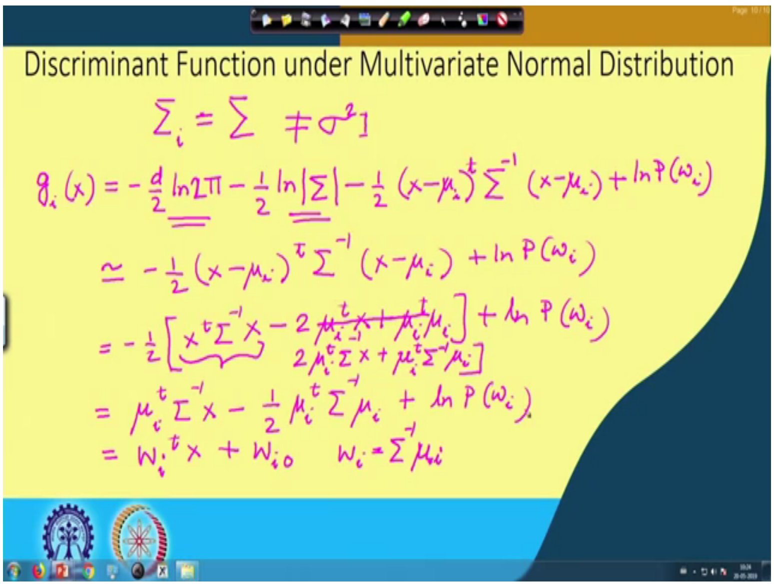
So, if you look at this if I consider the case that p of omega 1 and p of omega 2 that is a priori probabilities to be equal, then this term will be equal to 0 and X naught will be half of mu 1 plus mu 2 which is nothing, but the midpoint of the line joining mu 1 and mu 2. And the decision surface being orthogonal to w which is nothing, but a line joining mu 1 and mu 2 are vector drawn from mu 1 to mu 2. My decision surface will be orthogonal bisector of the line joining mu 1 and mu 2.

And that is what is this one, you have this blue line that is this which is the line joining mu 1 and mu 2 here, this was mu 1 and this was mu 2 and this red dotted line which is an orthogonal bisector of this blue line is the decision boundary between the two classes omega 1 and omega 2.

And as I said that as I am assuming omega 1 and P omega 2 to be equal so, this X naught is nothing, but midway between mu 1 and mu 2. So, given this if I have any point X which is falling on this side of the boundary, the X will be classified two class omega 2. Because I am considering these to be the vectors taken from class omega 2, whereas if I have an unknown vector y falling on this side of the boundary, this unknown vector will be classified to class omega 1. And as it is clear very clear from here the kind of classification rule that we have is nothing, but a minimum distance classification rule because for any point on this side of the boundary its distance from mu 1 will be less than its distance from mu 2.

Similarly, for any point on this side of the boundary its distance from mu 2 will be less than its distance from mu 1. So, that kind of classification rule that you have is a minimum distance classification. So, this is what we get for a simple case when I have a situation that mu i is equal to sigma square i.

(Refer Slide Time: 25:28)



Now, let us go to the next case that when I have a situation that mu i a sigma i is equal to sigma, but this may not be of the form sigma square i. What does it mean? It means that the covariance matrix of the samples belonging to all the classes are same, but the components of the vectors may not be statistically independent or in other words the off diagonal elements of this covariance matrix may not be 0, whereas in our earlier simplified case we had assumed that the off diagonal elements are 0, right.

So, given this situation my g i X if I compute from here as before the g i X was minus half log of 2 pi minus half log of mod of sigma i, but now sigma i is equal to sigma. So, it becomes minus half log of a determinant of sigma minus half X minus mu i transpose sigma inverse. Now sigma i is equal to sigma so, it simply becomes sigma inverse into X minus mu i plus log of P of omega i.

Again if I simplify this, you find that this is independent of the class, this is independent of the class so, these two do not participate in discrimination. So, my g i X now becomes same as minus half X minus mu i transpose sigma inverse X minus mu i plus log of P of omega i and if you expand this, it simply becomes minus half X transpose sigma inverse X minus 2 mu i transpose X plus mu i transpose mu i plus log of P of omega i.

As before you find that this X transpose sigma inverse X this is class independent, because sigma is same for all the classes ; so, I can ignore this, I can remove this form the discriminant function. So, the discriminant function now simply becomes mu i transpose, sorry here it should be there as a mistake. This should be 2 mu i transpose sigma inverse X plus mu i transpose sigma inverse mu i.

So, what I get is mu i transpose sigma inverse X minus half mu i transpose sigma inverse mu i plus log of P of omega i. And this you find that this is again of the form W i transpose X plus W i naught and this equation is again a linear equation, where this W i will now be sigma inverse mu i and W i naught will be minus half mu i transpose sigma inverse mu i plus log of P of omega i.

So, I will stop in this stuff here in this particular lecture. So, here we are what we have seen is starting from the discriminant function which you have seen that both in the case where sigma i is of the form of sigma square i whereas, and also sigma i is equal to sigma or decision or discriminant functions are linear. In case of sigma i is equal to sigma square i. We have seen that the decision surface is also linear and the decision surface is orthogonal to the line joining mu i and mu j and in case the a priori probabilities are equal that this decision surface becomes an orthogonal bisector of the line joining mu i and mu j.

And in the other case where my covariance matrix may not be a diagonal matrix and all the variances of all different components may not be equal, but even in that case the discriminant function g i X is a linear discriminant function. So, we will stop here today. In our next lecture we will start from this point.

Thank you.