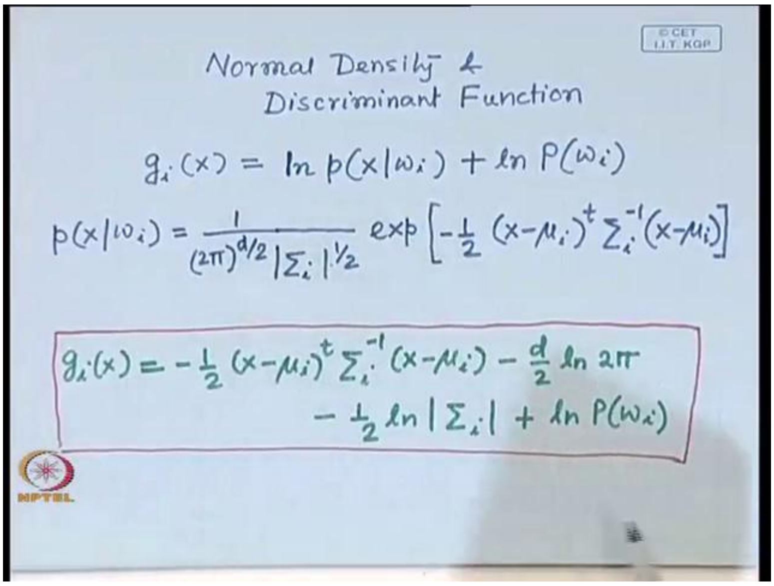
Pattern Recognition and Applications
Prof. P. K. Biswas
Department of Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Lecture - 8
Normal Density and Discriminant Function (Cond.)

Good morning, so we are going to start continue with our discussion on normal density and discriminant function which we started in our last class.

(Refer Slide Time: 00:40)

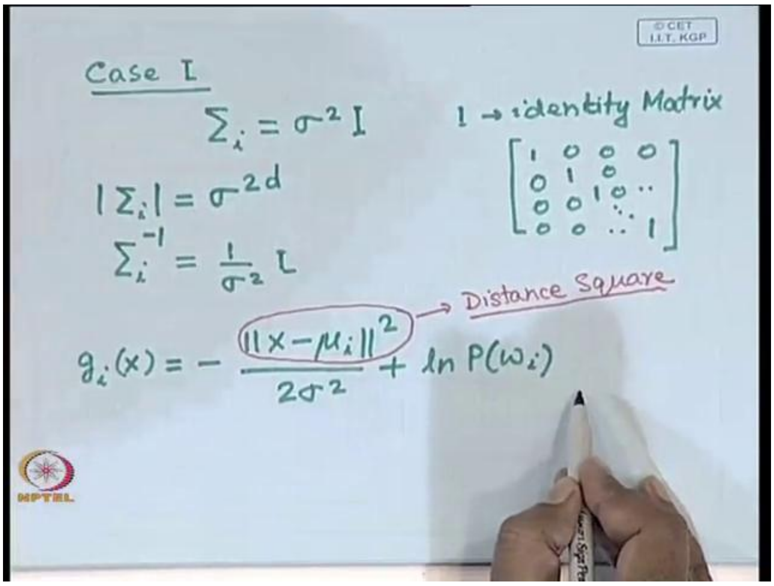


So, then the in the last class we said that the discriminant function is given by g i X is p l n p x given omega i plus log of actuary probability that is P of omega i, where omega i is the i th class. If we replace this probability density function P x given omega i by a normal density having the covariance matrix at as sigma i and the mean vector as mu i that is p of x given omega i.

If we write in this form 1 over 2 pi to the power d by 2, so this is a multivariate normal density function and the feature vector x are actually the dimensional vectors. So, taking this form of probability density function normal probability density function the discriminate function g i x becomes of the form minus half into x minus mu i transpose sigma i inverse into x minus mu i minus d by 2.

So, this becomes the discriminate function when we assume that the probability density function the associate probability density function is a multivariate normal distribution normal density function. We have said that depending upon this covariance matrix sigma i, we can have different types of discriminate function, so we will see those cases that if this covariance matrix sigma i takes different form. Then what are the forms of a discriminate function or what is the architecture of the decision boundary between two different classes vary depending upon the covariance matrix sigma i for the different classes. So, the first one that we will consider is the simplest form.

(Refer Slide Time: 04:08)



So, we will put that has case one when sigma i that is the covariance matrix is of the form sigma square I where this I is nothing but an identity matrix that is it is a diagonal matrix where only the diagonal elements are 1's and the rest of the elements are 0's. So, obviously you find that when this covariance matrix sigma i is of the form sigma square into I where I is the identity matrix. That means we have only diagonal elements which keeps only the variance of the individual components. The variance of all the components are same which is equal to sigma square and the off diagonal elements are equal to 0 which indicates that sigma i g is equal to 0.

That means the components different components are statistically independent, so when a covariance matrix becomes like this, so is nothing but the case when the feature of vector. So, the points belonging to the same class, they actually cluster in a hyper spherical space and the shape and size of all different clusters, they are same because the variance system. So, all the different clusters belong to different classes, so the points belong to the hyper spherical space and they are of the same size.

So, that is what is meant by this short of covariance matrix, now here we find that if the covariance matrix is like this, then the determinant of sigma i that is for empty class. It is nothing but sigma to the power 2 d, where this sigma is the standard deviation of independent components. It is simply because i have a d by d dimensional identity matrix oaky which is multiplied by sigma square, so I have a diagonal matrix where only the diagonal elements are equal to sigma square and the off diagonal element are all 0's.

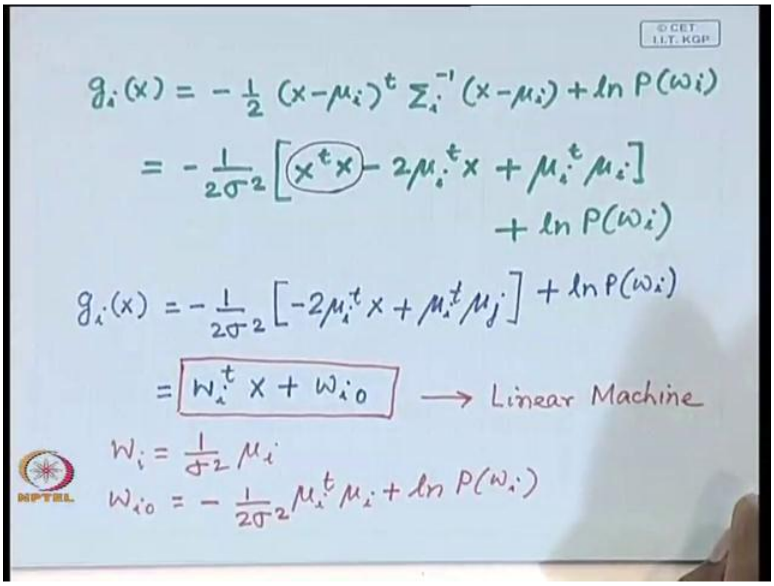
I have d number of diagonal elements, so the value of the determinant is nothing but the product of all the diagonal elements which is nothing but sigma to the power 2 d. It is also quite obvious that inverse of the covariance matrix that is sigma inverse sigma i inverse will be nothing but 1 upon sigma square into identity matrix I, so when I have in this case one that covariance matrix is same for all different classes. In this case having this simplified form this discriminate function g i x, you can compute it will simply become minus x minus mu i square into divided by 2 sigma square plus l n p of omega i.

So, what is this term this is nothing but distance square, so this is this is nothing but the squared distance of sample x from the mean vector mu i. So, in the situation that is all the classes are equally probable that is l n p omega i is there for all the classes, then this term becomes irrelevant in the discriminate function. So, my discriminate function is simply this equation and this being the distance square from the mean of the vectors, you find that it becomes a minimum distance plus square because this term will be maximum.

This term is minimum because it is negative and x minus mu i square being the squared distance from the mean vector and all the classes having same variance sigma square you find that this simply becomes the minimum distance plus square. So, given a feature vector x, you simply compute its distance from the mean vectors of all different classes and whichever distance is the minimum that is the nearest mean of the classes. This x will be classified to that particular class, now if we that even in this P of omega i is present.

That means, the probability of different classes are different they are not same then for a sample point which is equidistant from all the cluster means all the class mean. The decision will actually be biased by this term log of P of omega i.

(Refer Slide Time: 10:22)



Now, if I go for expansion of the term discriminate function, then you find that this g i X the discriminate function is simply of the form minus half x minus mu i transpose sigma i inverse into X minus mu i plus log of P omega i. Why i am writing like this, is in this expression you find that minus d by 2 log of 2 pi, this is independent of class this is a constant term. So, I can simply ignore this from the expression of the discriminate function because this term anyway is not going to influence our decision because for all the classes for the discriminate functions of all the classes this term is going to be present.

Similarly, this determinant of sigma i that is covariance matrix in this particular case which is nothing but sigma to the power 2 d, this is also same for all the classes. So, this also becomes irrelevant in the discriminate function, because for all the discriminate function the same term will be present and the value of this term will be same.

So, we will discriminate among the classes is the first term and the last term, so we have this simplified discriminate expression when sigma i is equal to sigma square into identity matrix. So, if I explain this term you find that this expression can be written as minus 1 upon 2 sigma square into x transpose x minus 2 mu i transpose x plus mu i transpose mu i plus log of p omega i. This is nothing but expansion of this term where sigma i inverse, we have said is 1 upon sigma square into identity matrix, so that is why here I get 1 upon 2 sigma square.

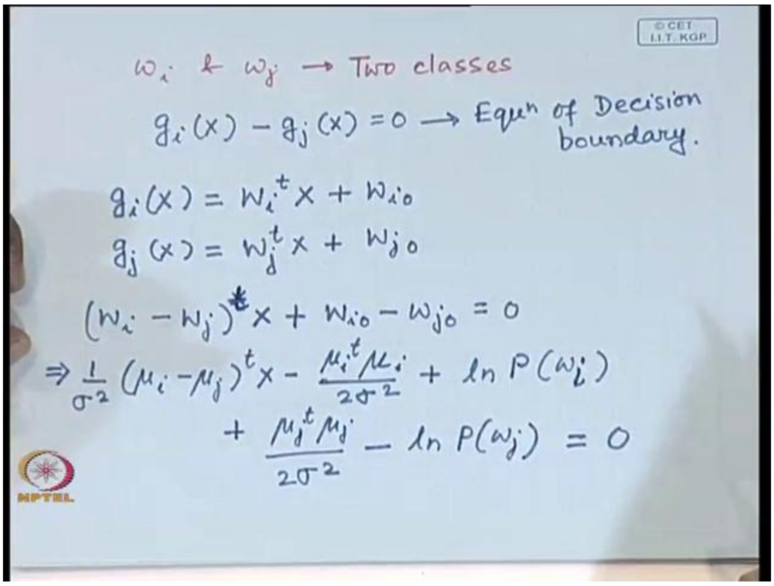
If I simply expand this I get this particular expansion, now again here we will find that the first term within the bracket that is this X transpose X. This is also independent of class for all values of i in all the discriminate functions g i X, the same term will be present and this sigma square being same for all the classes. This term X transpose X upon 2 sigma square that also becomes undetermined term it is not really deciding to which class this, so I can ignore this term from the discriminate function.

So, by ignoring this discriminate function becomes a simplified form like this g i x is equal to minus 1 upon 2 sigma square into minus 2 mu i transpose x plus mu i transpose mu j plus log of p omega i. This we find that I can write this in the form W i transpose x plus W i naught where to find we will find that this W i is simply 1 upon sigma square mu i and W i naught is nothing but 1 upon 2 sigma square mu i transpose mu i plus log of p omega i.

So, this particular expression can be written in this form W i transpose x plus w i naught where this w i is nothing but 1 upon sigma square into mu i and w i naught is minus 1 upon 2 sigma square mu i transpose mu i plus log of p omega i. So, if you look at this expression that finally, how we get the discriminate function expression for g i x which is w i transpose x plus w i naught this is nothing but a linear equation. So, in this simplified case, where all the covariance matrices covariance matrices for all the classes are same and is of the form sigma square into i the discriminate function simply becomes a linear equation.

So, this is if we have a classifier for the discriminate function which classify based on this linear equation that is called a linear machine. So, a classifier which uses the linear discriminate functions to decide about the class belongingness of an unknown vector that is called a linear machine, so this will form a linear machine, so that is about the discriminate function of an individual class or i th class. Now, if I want to find out the decision boundary between two different classes, so I have two class omega i and class omega j.

(Refer Slide Time: 17:14)

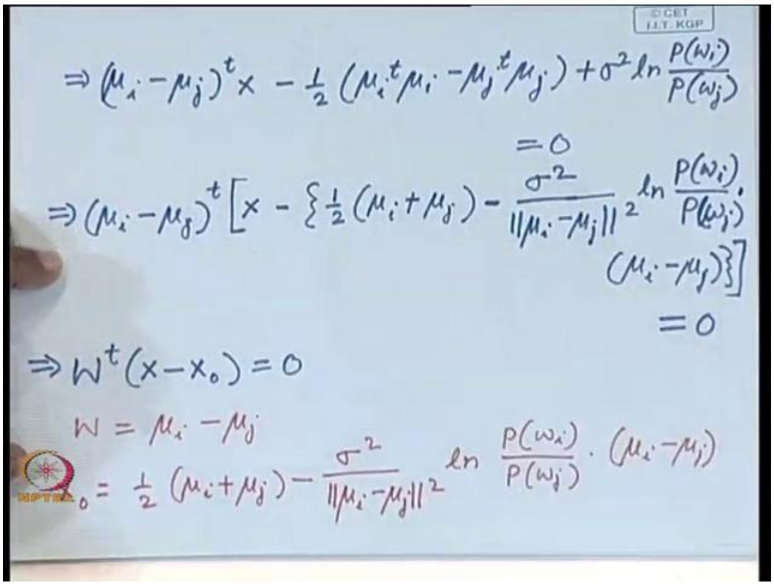


So i have two different classes so what we do to the nature of the decision boundary between two these two classes the i th class and the j th class. So, here we find that if I want to find out that decision boundary that decision boundary g i X for a featured vector lying on the decision boundary g i X will be equal to g j X. So, if that is the case, I mean that is the situation which differentiates between two different classes if g i X is greater than g j X, X will belong to class i if g j X is greater than g i X then x belong to class j.

The situation that g i X is equal to g j X that is took a decision of it that means the equation of the decision boundary is simply given by g i X minus g j X that is equal to 0 this is the equation of decision boundary. We have just seen that g i X is nothing but w i transpose x it is the linear equation plus w i 0. Similarly, g j X will be W j transpose x plus W j 0, so in this expression if I replace g i X by this and g j X by this. So, I simply get the equation as w i minus w j transpose x sorry transpose x plus w i naught minus w j naught that is equal to 0.

Here, if I replace W i by the value of W i which is nothing but 1 upon sigma square mu i and W i naught by the value of W i naught which is nothing but 1 upon 2 sigma square mu i transpose mu i plus log of P i P of omega i. Similarly, for W j and W j naught, so by replacing these values of W i W j and w i naught W j naught. This expression will simply become 1 upon sigma square mu i minus mu j transpose X minus mu i transpose mu i upon 2 sigma square plus log of P of omega i plus mu j transpose mu j upon 2 sigma square minus log of P omega j this will be equal to 0.

(Refer Slide Time: 21:45)

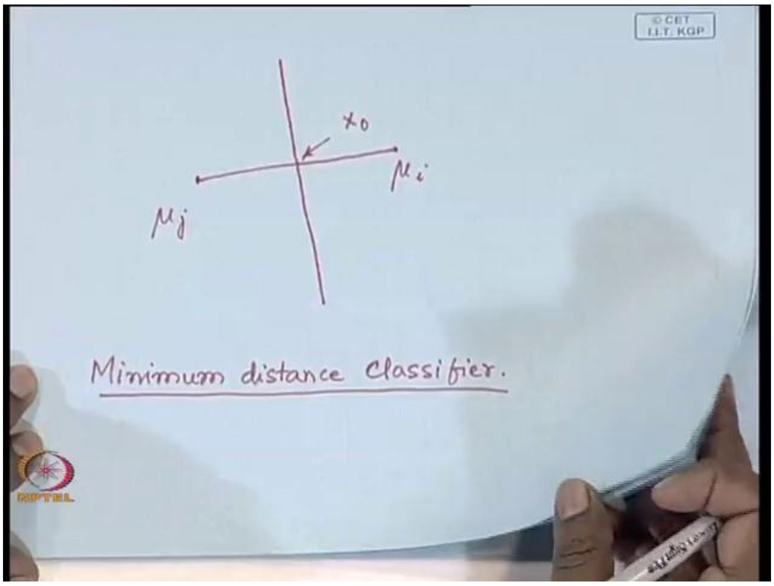


This same expression if I go on simplifying, this will lead to a form mu i minus mu j transpose x minus half it is simply the same expression. We have grouped together mu i transpose mu i upon 2 sigma square 2 sigma square and mu j transpose mu j upon 2 sigma square and grouped together, this log of p omega i and log of p omega i. So, I get this particular equation and this expression can further be simplified so I can write it in the form mu i minus mu j transpose into x minus half of mu i plus mu j minus sigma square upon mu i minus mu j square log of this is equal to 0.

I can convert this expression into this form where you find that this mu i minus mu j transpose that has been taken out of all the terms. So, when i can write in this form this is simply of the form w transpose x minus x naught is equal to 0 where this term w is nothing but mu i minus mu j. Here, mu i is the mean vector of the i th class and mu j is the mean vector of the j th class and x naught is simply half of mu i plus mu j minus sigma square.

So, what information we get from this here, you find that this is W transpose x minus x naught is equal to 0 that is the decision boundary between the i th class and j th class, but this w is given by mu i minus mu j. That means it is a vector drawn from the j th mean vector mu j to the i th mean vector mu i of the line joining mu i and mu j. So, that is what is W and as the decision surface is W transpose X minus X naught equal to 0, so the decision surface is orthogonal to the vector joining mu i mu j.

(Refer Slide Time: 26:14)

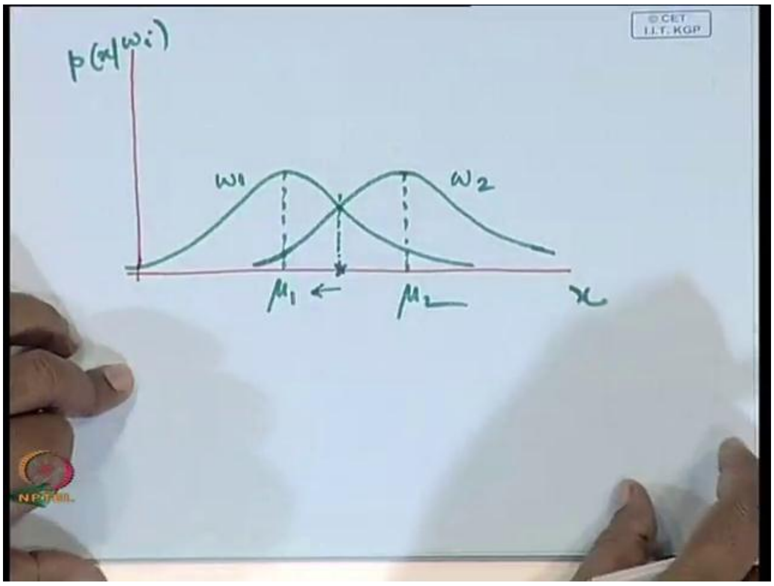


So, the situation is simply like this that if two vector positions are say I have mu i over here and mu j over here and this is the line joining mu i and mu j. The decision surface will be orthogonal to mu i and mu j and because it is the linear equation, so the decision surface is a hyperplane which is orthogonal to the line joining mu i and mu j. If this term is equal to 0 that is when the actual probability is P omega i is equal to P omega j, then this term is equal to 0.

This orthogonal plane it passes to point x naught which is on the line mu i and mu j, so this is point x naught if this term is 0 in case that probability is P omega i is same as P omega j then x naught becomes half of mu i plus mu j. So, this x naught in that case is midway between mu i and mu j and I have this hyperplane which in this case is nothing but orthogonal bisector of the line joining mu i and mu j and because it is the orthogonal bisector.

Again, I came back to the same situation that are kind of classifier that I have is the minimum distance classifier because if I have a have an orthogonal bisector of the line joining mu i and mu j. Then for all the points which are falling on the side of mu i, mu i its distance from mu i will be less than its distance from mu j. Similarly, for all the points which are falling on the side of mu j, its distance from mu j will be less than its distance from mu I. We put the sample x to that particular class from which its distance is minimum. So, when it is an orthogonal bisector or the probability P omega i and P omega j, they are same then effectively what I get is a minimum distance classifier.

(Refer Slide Time: 29:12)

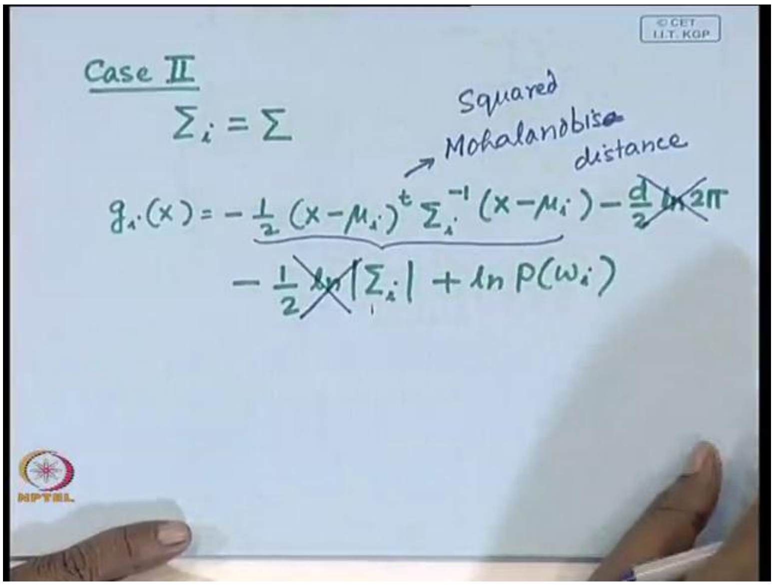


So, effectively it says that if the probability density functions the normal probability functions like this again plotting in you need variant case, so in this side I put x and this is p of x given omega i and suppose this is for omega 1 and this is for omega 2. So, if the probability is P omega 1 and P omega 2 are same and this is what is mu 1and this is what is mu 2. So, the probabilities are same my decision will be decision surface will be at the middle if the probabilities are more if P omega 1 is more than p omega 2 probability.

Then, the decision surface will be pushed towards P omega 2 that is away from mu 1 if p of x omega 1 P omega 1 is greater than p of omega 2, then this decision surface will be pushed away from mu 1 if p of omega 1 is less than p of omega 2. Then the decision surface will be pulled towards mu 1, so in the simplified case when all the covariance matrices are same and they are of the form sigma square into i, I get such simplified linear classifiers.

The decision boundary between the two classes will be orthogonal line joining the mean corresponding vectors if the two classes are equally probable that is if the probabilities are same. Then the decision surface is nothing but a orthogonal bisector of the line joining mu 1 and mu 2, in this case the classifier effectively becomes a minimum distance classifier.

(Refer Slide Time: 31:36)



Let us take another case that is case two here I assume again that all the classes have the same covariance matrix that sigma i is equal to sigma. In a first case we have said that this sigma is of the form sigma square into i, but now we are not putting that restriction the sigma or the covariance matrix is arbitrary, but all the classes samples belonging to all the classes have the same covariance matrix though it is arbitrary. So, in the previous case we said that the samples are clustered into hyper spherical spaces of equal size in this case, the sample space is clustered into hyper ellipsoidal spaces of same shape and same size.

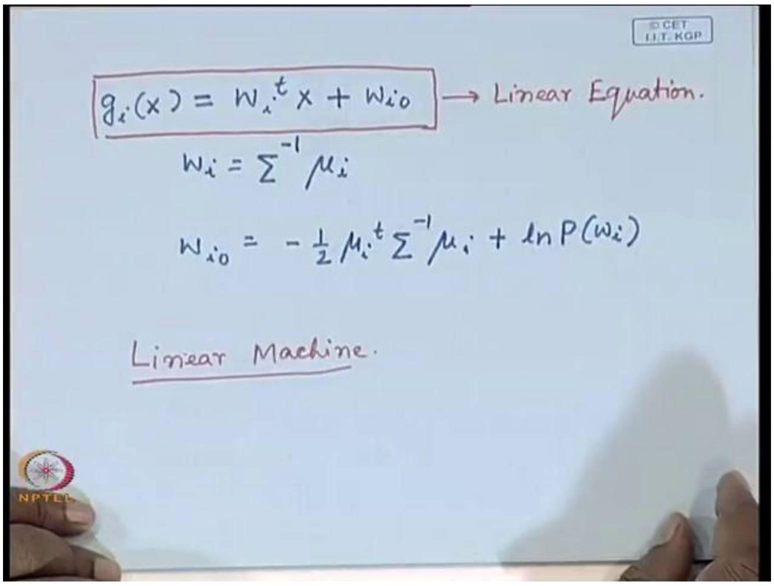
It will be same shape and same size because of its covariance matrix system and as the covariance matrix is arbitrary, so in general it will be a hyper ellipsoidal space. So, the samples belonging to different classes the cluster into ellipsoidal space of same shape and same size is that so that is the second case. So, here again we find that out of this expression of the discriminate function the general expression, in the general expression we had g i x is equal to minus half x minus mu i transpose sigma i inverse into x minus mu i minus d by 2 log of 2 pi.

So, this is the general expression of the discriminate function, now here we say we are saying that sigma i or the covariance matrix is same for all the classes so if it is same for all the classes. Then obviously this term obviously this is a constant term, so this does not give you any discriminate power and again this term this is same for all the classes because this sigma i is same for all the classes. So, this also becomes irrelevant, so we are left with g i X is equal to minus half X minus mu i transpose sigma i inverse X minus mu i plus log of P omega i as before.

Here, again we find that if this log of P omega I, this also becomes irrelevant that is if all the classes are equally probable, then my discriminate function simply is g i X is equal to minus half X minus mu i transpose sigma inverse x minus mu i. We said that this is nothing but a distance function which is squared Mahalanobi's distance squared. So, effectively the classifier again becomes the minimum distance classifier because this being a distance function if g i x has to be maximum.

Then, this has to be minimum and X will be put into that particular class for which this term is minimum, so again it becomes a minimum distance classifier, but the distance that we have to compute in this case is not the distance, but the Mahalanobi's distance. So, considering the fact that we will compute Mahalanobi's distance again, we have a minimum distance classifier, so as I have done before if we simply expand this term.

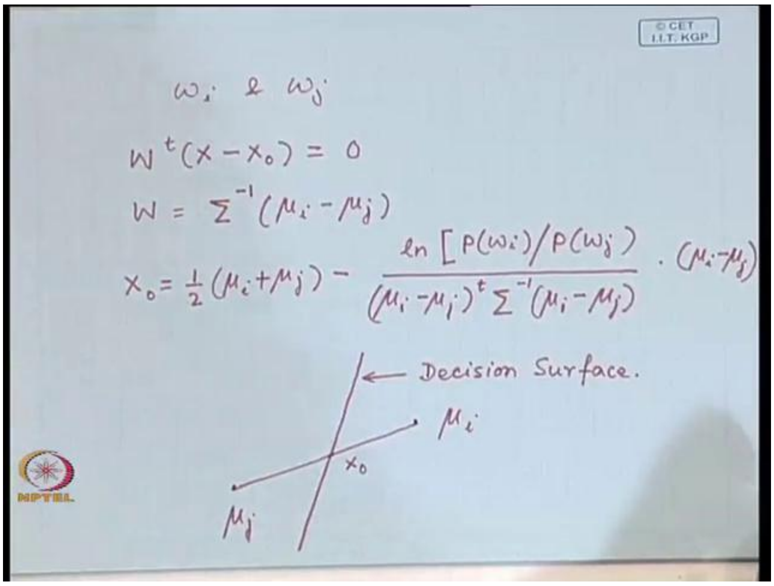
(Refer Slide Time: 36:41)



Then, again we will find that the discriminate function g i X will again be of the form W i transpose X plus W i naught where in this case. In the earlier case, this W i had a very simple form that is one upon sigma square into mu i, now this W i will be of the form sigma inverse into mu i. Earlier this sigma inverse was replaced by one of the sigma square in the first case, now it becomes sigma inverse into mu i and W i naught is simply minus half mu i transpose sigma inverse mu i plus log of p of omega i.

So, again we find that this term the nature of the discriminant function this is again a linear equation plus this is W i transpose x plus W i naught, so this is again a linear equation. So, our classifier making use of such linear discriminate functions that again becomes a linear machine now using these linear discriminate functions if I want to find out as I have done before.

(Refer Slide Time: 38:43)



We want to find out the nature of or the structure of the decision surface between two classes omega i and omega j, so again we will we will follow the same procedure that i x minus g j x equal to 0. That is the equation of the decision surface separating the two classes omega i and omega j and if I follow the same steps again. Here, we will find that the equation will come in the form W transpose X minus X naught is equal to 0, but this W is given by sigma inverse into mu i minus mu j and X naught is given by half of mu i plus mu j minus log of p omega i upon p omega j.

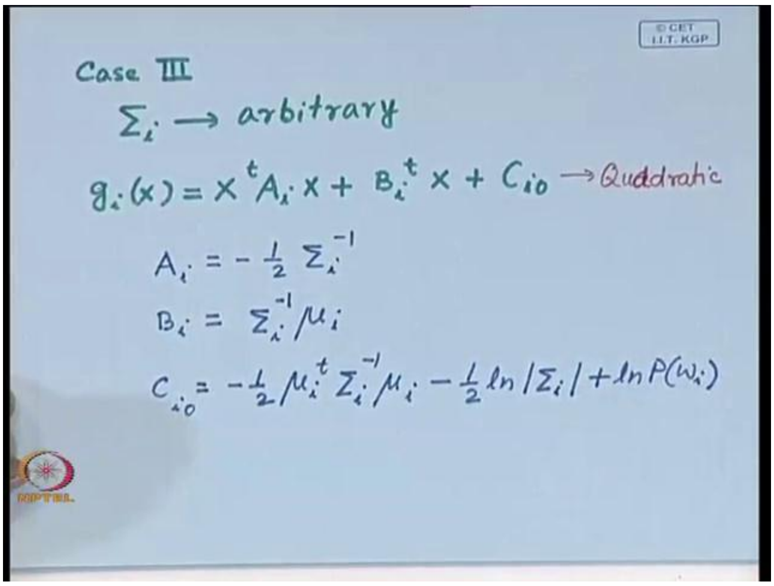
So, the expression becomes more or less same I get similar expression W transpose X minus x naught equal to 0, but this w is nothing but sigma inverse mu i minus mu j. If you remember earlier case our W was simply mu i minus mu j and mu i minus mu j is the line of the vector from mu j to mu i. So, W had the same direction as the vector mu i mu j nut sigma inverse mu i mu j is not in general in the direction of mu i and mu j. So, as a result this vector W which is sigma inverse mu i mu j this is not in general in the same direction of the vector mu i mu j.

So, we find that from this expression though my decision surface is orthogonal to W in the previous case W was in the direction of the vector mu i mu j. So, the decision surface was orthogonal to the line joining mu i and mu j and in this case because W is not in the direction of line joining mu i mu j. So, the decision surface is not in general orthogonal to the line joining mu i mu j nut the decision surface passes through the point X naught where the equation expression for X naught is given by this.

So, here again you find that if P of omega i and P of omega j they are same then this term becomes 0, when this term becomes 0, then X naught is half of mu i plus mu j that means it is halfway between mu i and mu j. So, when the classes are equally probable the decision surface or the hyper plane passes through the midpoint of the line joining mu i and mu j, but it is in general not a orthogonal bisector it is a bisector, but not orthogonal.

So, effectively the kind of situation that we will have is if i have these two means mu i and mu j this is the line joining mu i and mu j and the decision surface will be something like this. This is point X naught and this is and this is your decision surface, so we still have the linear classifiers the decision surface two classes are still hyper planes, but this is not a orthogonal bisector or this is not orthogonal to the line joining mu i mu j.

(Refer Slide Time: 43:43)



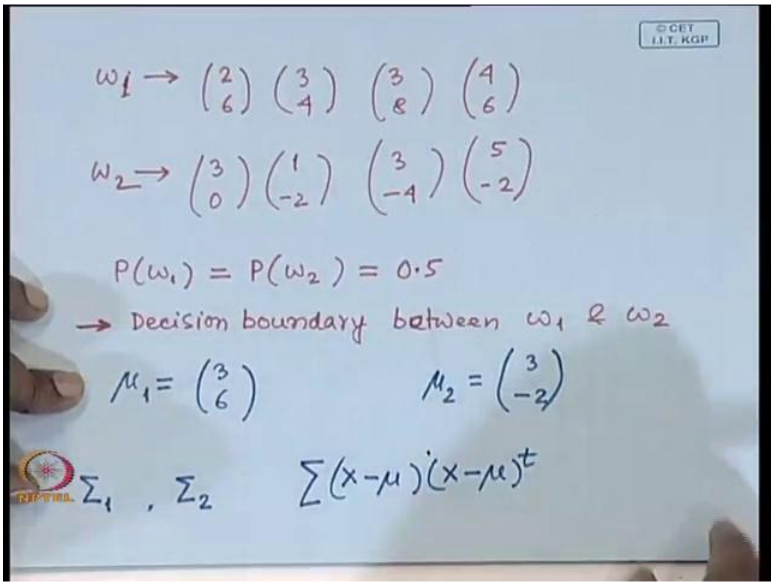
Now, let us take the third case which is the most general one that is the covariance matrix mu i is totally arbitrary that means different classes have different covariance matrices. In case two we considered that different classes have arbitrary covariance matrices, but covariance matrices of all the classes are same though arbitrary. The first case was the simplest case where we assume that the components different components are statistically independent and individual components are the same variance. So, this is the most general case where the different classes will have the different covariance matrices.

So, here I do not have any other option i have to keep all the terms, so g i X if I keep all the terms. You will find that it will be of the form x transpose a i X plus B i transpose X plus C i 0 where this a i is minus half sigma i inverse B i is sigma i inverse into mu i and c i 0 is given by minus half mu i transpose sigma i inverse mu i minus half. So, if we find that in the earlier cases where we could remove the quadratic terms that is term involved in X transpose X or X transpose a X we could have removed term because X transpose.

X does not give you any discriminating power, but over here this is X transpose a i X where a i X depends upon the value of sigma i that is the covariance matrix and different classes have different covariance matrices. So, this term x transpose a i X, now becomes class dependent because it is class dependent it really contribute to decide whether X belong to class omega y i or X will belong to class omega j.

So, I cannot remove any of the terms and this expression that I have it is nothing but a quadratic expression, so giving you a quadratic classifier in the earlier two cases we had linear classifiers in this case we get a quadratic classifier. Using this quadratic classifier if I want to find out the decision surface between two classes omega i and omega j, the decision surface will be a hyper quadrics. So, it is a quadratic surface in a multi-dimensional space, so these are the different cases that we can have assuming multi variant normal density function for the samples belonging to different classes.

(Refer Slide Time: 48:00)



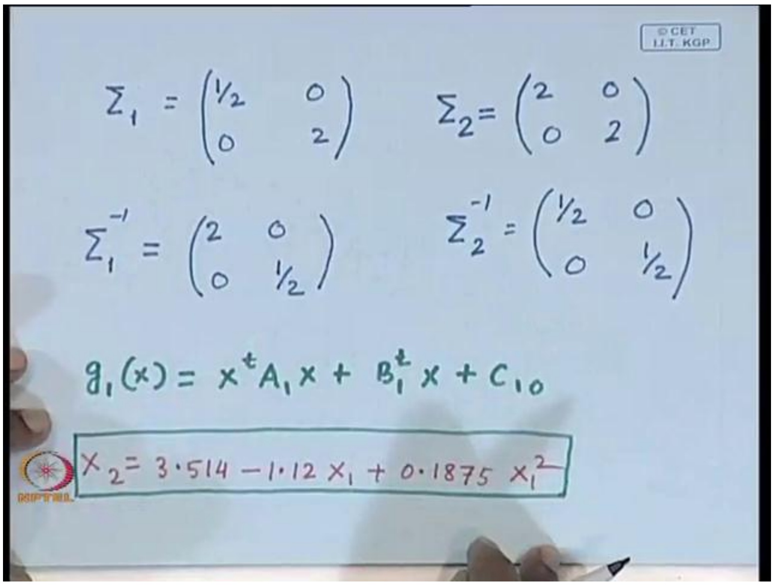
Let us see an example suppose we have points belonging to two classes one is omega 1, class omega 1 and class omega 2, the training samples which are provided for these two classes are like this. For class omega 1 it is 2, 6, 3, 4, 3, 8 and 4, 6, so these are the training samples which are given for class omega 1 and the training samples for omega 2 are given as 3 0. So, naturally we are considering it two dimensional space and as I said that the examples typically I will take in two dimension because I can draw it on it.

Then, they were so this is 3, 0, 1 minus 2, 3 minus 4 and 5 minus 2, so these are the training samples for the belonging to two classes omega 1 and omega 2 which is given. It is also said that the probability is P of omega 1 which is equal to P of omega 2 and naturally this is equal to 0.5. So, what we have to do is we have to find out that the decision boundary between plus omega 1 and plus omega 2, so to find out the decision boundary i have to find out the discriminate function between two classes omega 1 and omega 2.

Then, subtract that equate that to 0 or in general because i have to compute sigma i that is the covariance matrix for independent classes and I do not know what is the nature of sigma i. So, in general I can make use of this expression and go ahead with it assuming that my factors will be quadratic in case the surface is not quadratic all the quadratic terms will get cancelled.

So, first what I have to do is I have to find out the mean vector which is mu 1 and mu 1 is nothing but I will take all these vectors and find out their means. So, this mu 1 in this case will come out to be 3, 6 you can verify this, similarly for class two mu 2 will come out to be 3 minus 2, so that is the mean vector for class two we also have to find out the covariance matrix sigma 1. The covariance matrix sigma 2 you know how to compute the covariance matrix once I get the mean vector. Then what I have to do is I have to find out x minus mu into x minus mu transpose take the summation over all the vectors, and then normalize by the number of vectors not taken.

(Refer Slide Time: 51:56)



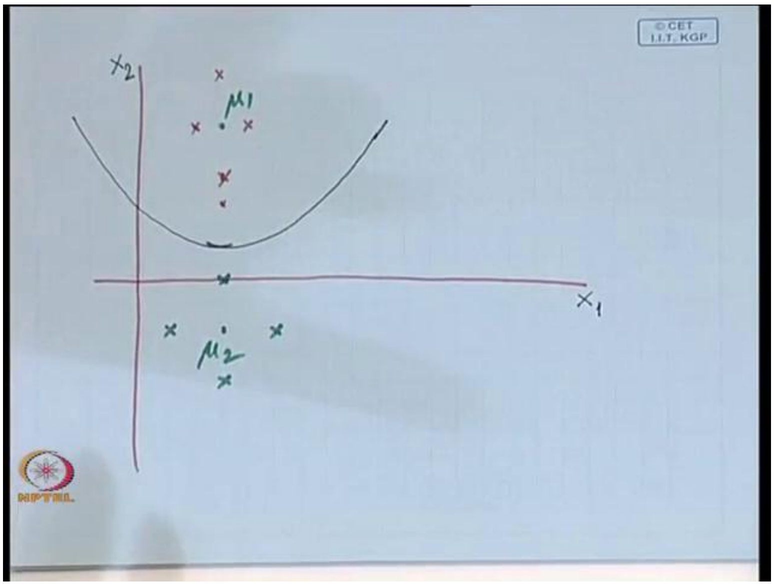
So, that gives you the covariance matrices sigma 1 sigma 2, so if you do it that way you will find that sigma 1 comes out to be half 0, 0, 2. So, this is the covariance matrix of class one and sigma 2, that is the covariance matrix of class two comes out to be 2, 0, 0, 2. So, we find that the covariance matrices of the two classes are different and as the covariance matrices of the two classes are different the classifier is not a linear classifier. It will be quadratic classifier; from here I can compute sigma 1 inverse which is nothing but 2, 0, 0 and half.

From here, you can compute sigma 2 inverse which is nothing but half 0, 0 and half, so these are the inverses of the covariance matrices. Now, once I get this I find out g 1 X which is of the form X transpose A 1 X plus B 1 transpose X plus C 1 0 and the expressions. We have given over here in a quadratic classifier, but the value of A 1 will be nothing but half of sigma one inverse value of B 1 will be sigma inverse mu 1 and value of C 1 0 will be minus half of mu 1 transpose sigma 1 inverse mu 1 minus half log of determinant of sigma 1 plus log P of omega 1.

Now, in this case P of omega 1 is same as P of omega 2, so this term can be negated even if i keep it on both sides they will get cancelled. So, I do not have to really compute this, so if you compute all these terms and put in this expression. Similarly, you find out g 2 X following similar expressions and then equate g 1 X is equal to g 2 X or g 1 X minus g 2 X equal to 0.

Then, we will find that finally the decision surface I am not going to the detailed calculation, let us look a simple one the final decision surface will come out to be X 2 is equal to 3.514 minus 1.12 X 1 plus 0.1875 X 1 square. So, we find that the decision surface is linear quadratic in this case it is the curve because I am we are in two dimensional space. So, the decision boundary is really a quadratic curve, now if I want to find out how this quadratic curve looks like I will simply plot these points on paper and find out how the surfaces look like.

(Refer Slide Time: 55:50)



So, I have points 2,6, then I have points 2, 4, sorry 3, 4 I have 3, 3, 8, 4 is here, this I have 3, 8, so this is one point, this is one point, this is one point and I have 4, 6 that is this one. So, these are the points which belong to class omega one, similarly for omega 2 I have 3, 0, this is a point belonging to class omega 2, I have 1 minus 2, 1 minus 1, one is over here. I have 3 minus 4 that is this one. I have 5 minus 2 that is this one, so these are the point which belongs to class omega 2, so all the reds belong to class omega 1 and the greens belong to class omega 2 mu.

One is somewhere over here, this is mu 1 and mu 2 is over here and if you plot this quadratic curve into this space, so this quadratic curve be something like this where I am assuming that I have two components one is x 1 other one is x 2. So, we find that we really get a quadratic curve separating the two classes omega 1 and omega 2, so I will stop here today and we will continue with other classifiers next time.

Thank you.