

Indian Institute of Technology Madras
Presents

NPTEL
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING

Pattern Recognition

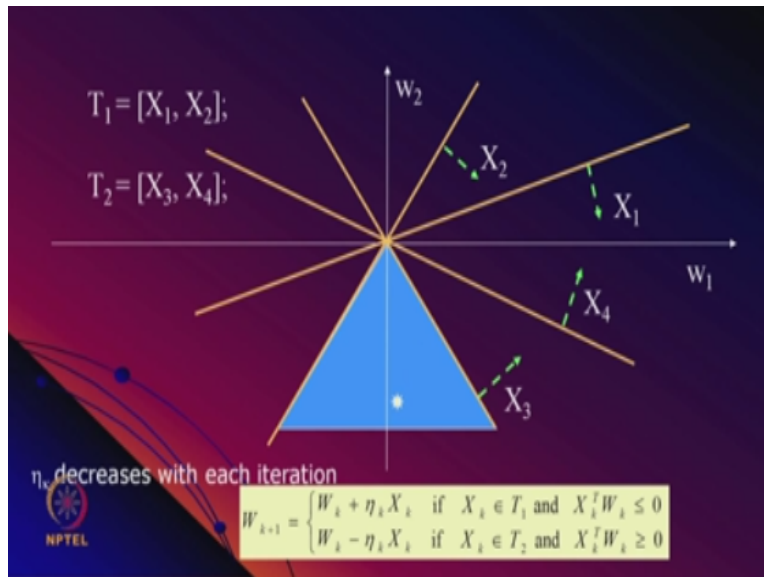
Module 02

Lecture 10

Perception Learning and
Decision Boundaries

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

(Refer Slide Time: 00:14)

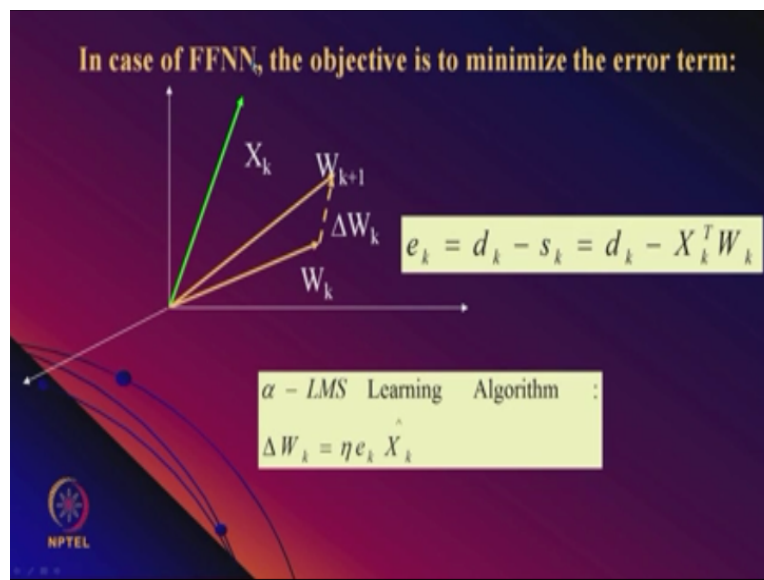


So this is where we stopped at the end of the last class and we were actually looking at the significance of the equations here for learning or updating the weights based on samples from two different classes and based on the inequalities here whether you are in the positive or negative side of the hyper plane you adjust the weights based on the sign of the direction of the tests training samples okay.

And we also discussed about the significance of the learning rate parameter as it is called and you start with the larger value and keep on reducing them as time progresses and you stop actually when the iteration converges and which basically means that these conditions are not met typically the condition of convergence in a learning algorithm is not based on the these inequalities.

Basically you find out the errors in predicting the classification of samples as given during the training and the error is very, very less or does not change over a certain amount of time you stop the iteration of the convergence to be very, very specific.

(Refer Slide Time: 01:33)



In the case of a feed-forward neural network where a back projection learning algorithm is used the objective is to basically minimize an error term okay. So this objective of the minimization of the error term is also along similar logics were given a certain weight W_k at any point of time of an iteration you adjust the weight and get the new weight W_{k+1} the change in the way W_k is basically given by this particular expression and member as shown in this diagram the change in the weight is along the direction of the vector of course it could be the negative side as well depending upon which class you are working on and the least mean square learning algorithm.

That the derivation which I am skipping here which says that the change in the way it should be η which is the learning rate parameter multiply by the direction vector of the sample multiply by an error term this is something new which you have not discussed earlier remember the logic

is the same that means you change it along the direction of the vector the ΔW_K is a vector as given here it is parallel to or along the direction of the test sample which is here multiply by the running mate parameter and there is another scalar quantity here which is an error term defined by this what is this error term indicating.

You look at the difference between the desired value of the discriminate function and the actual function of classification and then decide typically you want this you stop when they are in this error term is 0 this error term is 0 when the corresponding value of this G which the discriminate function here is the same as a desired value in this case in our case it is simply positive or negative but if in a multi-class situation this could actually indicate a value so when this value of the error.

That means the desired value for the classification required for the discriminant function G and the actual value of G which is X multiply by W was given in earlier class if this value is large you move you actually have a larger value of ΔW_K because this value will then be larger if this value is negligible or small the change will be also small so you see there are three factor direction given by X_K learning rate parameter which reduces over time and the error.

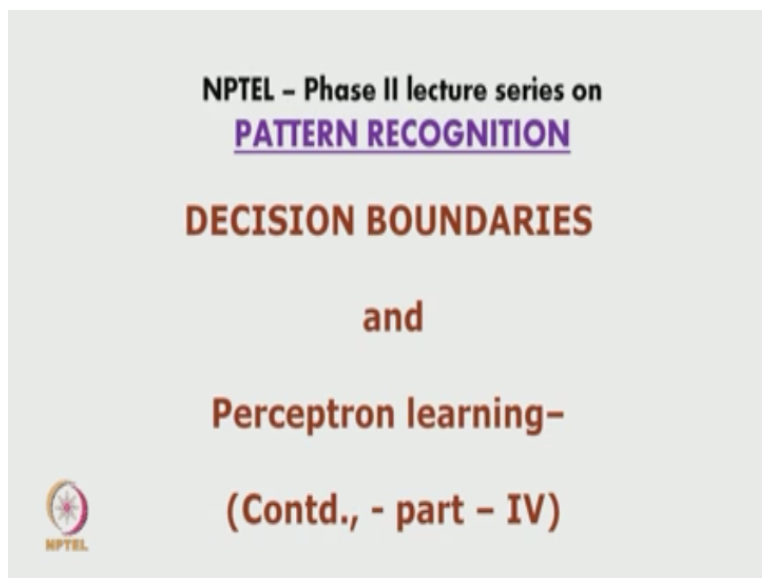
So the error is large you move larger in larger steps if the error is small you move with smaller steps and of course as iteration proceeds the iteration learning rate parameter will also go down forcing you to anyway move at smaller steps when do you converge now you have an error term K you look at the error term and say for all training samples if the error term is less or some of their terms is less our individual terms are small or they do not change over time then you can stop.

Because you have reached the solution space you have is the solution space because this term x multiply w is actually equal to DK and are very close to that giving to a null or a very, very small so this is the essence of the learning algorithm in case of a Perceptron as well as the feed-forward Neal neural network in general and this is applicable with minor modifications for a multi-layer feed forward neural network algorithm as well where the learning algorithm is actually called the back propagation learning law BP neural network back professional network very casual it is called but actually the error is actually back propagated the signal is moving from left to right from the signal input towards the output.

But the errors are propagated back which is there a top this is an example of such an error term which is propagated back to know how much of the weights should I adjust depending upon the error the error is back propagated but the signal flows forward from the input x the weights to the output to the next layer and so on in a similar case it is done so it is applicable for a Perceptron but a Perceptron you can actually use the simplistic form which you have discussed in the previous slide without the error term as well but if we have a layer of you know different you know Perceptron forming a single layer or multiple layers with different number of new neurons or Perceptron as they are called to form a multi-layer feed forward neural network learning law of this type is adopted where the errors are propagated back.

They are measured multiplied with an learning rate parameter η then how the weights are modified always along the direction of the training sample this is what we have learnt to the end of the last class and also today that it brings you in the weight space in the weights page towards the positive zone of G for of course the correct set of samples.

(Refer Slide Time: 06:38)



We move at and go back to our linear discussion of linear decision boundaries.

(Refer Slide Time: 06:43)


Linear Decision Boundaries

The decision region boundaries are determined by solving :

$$G_i(X) = G_j(X),$$

which gives: $(\omega_i^T - \omega_j^T)X + (\omega_{i0} - \omega_{j0}) = 0$

This is an expression of a hyperplane separating the decision regions in \mathbb{R}^d . The hyperplane will pass through the origin, if:

$$\omega_{i0} = \omega_{j0}$$


Remember in the last class we talked about that we will not discriminate between and boundaries and discriminant functions okay now we will discriminate it exactly formerly earlier we were not discriminating much one discriminant function is discriminating between one class with respect to the other or the rest and that is also acting as a decision boundary now we will actually do is what we will do is take to discriminant functions both of them being linear what are these two discriminant functions for two separate classes I, n, j 1 and 2 A and B examples again I repeat fruits and flowers cars vs trucks aircrafts vs. trains.

Let us say or two different landscapes in the case of remote sensing applications whatever you are trying to discriminate one class with respect to the other let us say you form discriminant functions for each class and we have seen expressions of that courtesy using the multivariate Gaussian distribution under the Bayes law that gives us a discriminant Function the mall on the base distance criteria under that when we took the covariance term equal to an aunt Ida matrix.

We it boil down to a linear discriminant function if we take two such linear discriminant functions as given here you get a decision region boundary by solving this, this is a general expression for a DB decision boundary between two regions if individually each of them are linear then you actually get an expression like this okay where these w 's are the corresponding weights these are the corresponding bias terms they are for two different classes I, n, j and each of them and this actually indicates a linear decision boundary precisely now discriminating between these two classes.

So you are using two discriminant functions for each class or this is actually represent also a group of classes put together if necessary for certain applications or I could be one class J could be a mixture of other classes but it is basically a binary class education problem two linear discriminant functions representing two groups or two sets of classes or even just two classes themselves or even one versus the rest whatever the case may be we are not worried about this at this point of time.

But if we put them under this expression then you get a linear instrument function and is a course an expression of a hyper plane a point in one day line in 2d planar surface in 3dhyper planes in higher dimensions separating the decision regions or dr's in high-dimensional space of course we know that the hyper plane will past through the origin if the bias term themselves are zero or they cancel out either they are zero or they cancel out such cases.

(Refer Slide Time: 09:47)

Generalized results (Gaussian case) of a discriminant function:

$$\begin{aligned} G_j(X) &= \log[P(X | C_j)] = \\ &= \log\left[\frac{1}{\sqrt{\det(\Sigma_j)} (2\pi)^d} \right] - \frac{(X - \mu_j)^T \Sigma_j^{-1} (X - \mu_j)}{2} \\ &= -\frac{1}{2} (X - \mu_j)^T \Sigma_j^{-1} (X - \mu_j) - \left(\frac{d}{2}\right) \log(2\pi) - \frac{1}{2} \log(\Sigma_j) \end{aligned}$$

The **mahalanobis distance** (quadratic term) spawns a number of different surfaces, depending on Σ^{-1} . It is basically a vector distance using a Σ^{-1} norm.



It is denoted as: $\|X - \mu_j\|_{\Sigma_j^{-1}}^2$

We get, get back to the expression of that capital G which gave us that discriminant function so we are looking at the generalized results of the Gaussian case of discriminant function G remember in we started with these two classes back we linearize it under the assumption that this is an identity matrix now we will not make identity matrix let this be a general covariance matrix but again we will take some special cases so this expression is not new this is the model numbers term.

This is the constant term and this is of course there is a class variant because if the Σ I changes from one class to another this could be separate and this modern resistance in general if this is not an identity matrix it gives you quadratic terms or quadratic expression it spawns a number of different type of quadratic surfaces some of these examples will take in the next class when we talk of nonlinear decision boundaries because we will still talk of linear discriminant functions under special cases of covariance matrix not equal to an identity matrix also.

We will talk about that today and still will be discussing in detail about linear decision boundaries but remember in general this is a malleable distance criteria it is a vector distance using the inverse of the covariance matrix term and it is denoted by this, this is the symbol we might use in certain cases when this remember if this is an identity matrix then this is the simple equivalent distance norm which we have used and then we are linearize as well by ignoring the class in very in term the quadratic term okay But in general this the is criteria okay.

(Refer Slide Time: 11:25)

Make the case of Baye's rule more general for class assignment. Earlier we has assumed that:

$$g_i(\vec{X}) = P(C_i | \vec{X}), \text{ assuming } P(C_i) = P(C_j), \forall i, j; i \neq j.$$


Now,

$$G_i(\vec{X}) = \log[P(C_i | \vec{X}) \cdot P(\vec{X})] = \log[P(\vec{X} | C_i)] + \log[P(C_i)]$$

$$G_i(X) = \log\left[\frac{1}{\sqrt{\det(\Sigma_i)}(2\pi)^d}\right] - \frac{(X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i)}{2} + \log[P(C_i)]$$

$$= -\frac{1}{2}(X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i) - \left(\frac{d}{2}\right) \log(2\pi) - \frac{1}{2} \log(\Sigma_i) + \log[P(C_i)]$$

$$= -\frac{1}{2}(X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i) - \frac{1}{2} \log(\Sigma_i) + \log[P(C_i)]$$

 **Neglecting the constant term**

So under the Bayes rule modern class assignments we have assumed that the g of x is equal to this what is this term called in the Bayes theorem I repeat we have done this, this is called there are four terms glass prior unconditional condition on a posterior which one is this the classical posterior given a sample X what is the class to which it belongs to correct so this is the left-hand side of the page expression that is our discriminant.

We started with that and we made several assumptions we ignored classifier of course you always ignore the unconditional denominator term okay but we are of course mainly concentrate on the conditional distribution function that is where we bought in the multivariate normal density distribution so we will assume that the class priors are still the same for the time being ok so the in some sense basically this is what we are talking about G of X well I just give it a different notation.

Because this is not the same as it so put the, the unconditional density function here product of these two is log of this Plus this and under this if the class priors are same this is the one the class

conditional function here is the one which is going to dictate and we take this to be our normal density function this one is this, this is nothing new we have all done this but I am just recollecting all of that so that we can put ourselves in perspective the log prior term is still here and if you break it in two parts and simplify this expression will give you this term and this term the log prior is still here this is the expression within the exponent sorry it was within the exponent it has already taken out.


So this still here it is still here so we have just ignored the constant and look at the final expression here what we get I leave this simple derivation here this what is the constant term we ignored this one this is the constant term we ignored when we went from this step to this step here the log prior still here the log of the covariance matrix is still here yes and the inverse this is the natural base resistance here okay.

(Refer Slide Time: 14:01)

Simpler case:
 $\Sigma_i = \sigma^2 \mathbf{I}$, and eliminating the class-independent bias, we have:

$$G_i(X) = -\frac{1}{2\sigma^2} (X - \mu_i)^T (X - \mu_i) + \log[P(C_i)]$$

These are loci of constant hyper-spheres, centered at class mean.
 More on this later on.....



So this is what we are working on and earlier we have assumed this to be an identity matrix so now let us take the next simplest case or the simpler case of the covariance matrix instead of taking it to be an annuity matrix we will take it to be a diagonal matrix not only diagonal all the terms variance terms are same and that to follow our classes physically if you want to interpret this what are you talking about you are distinguished between distinguishing between safe fruits and flowers two different classes.

And I have taken feature samples like color, weight, smell, size and so on and so forth what I have found is across the two different classes across the two different classes it seems I am having the same variance same scatter of all these features the variance of color feature across the fruits which I have taken is the same as the flowers well from my statement itself you can almost imagine how little bit quite a bit unrealistic this is this assumption but for the sake of mathematical argument will start to relax these constraints one after another in the previous case mind you.

We took the coverage to buy and read a matrix that means there were no variance at all a variance was equal to one now at least we are having some variance but the variance is same across all dimensions across all classes fruits are having the same variance of color as the case of flour the variance of weight is also the same size is also the same and so on and so forth not an idealistic assumption here but just to show that this may not be good but for the sake of mathematical argument yes we will still proceed and go on and we will eliminate the class independent bias which one did we eliminate in the previous expression first of all there were three terms the lock pride is there what did you eliminate can you guess and tell me from the previous one.

We will go back and have a look this term how could we eliminate this because the log of this matrix actually I should put a mode here because the determinant of this so predominant term is missing here so please put the determinant the determinant of the Σ okay for a diagonal matrix what is the determinant because we have made an assumption that it is strictly diagonal the diagonal matrix the determinant will be the product, product of all the diagonal terms correct product of all the diagonal terms.

So that is the same for across classes that is why the term could be ignored here and inverse of a diagonal matrix inverse of a diagonal matrix will be 1 by those elements okay and the elements are σ^2 so $1/\sigma^2$ now it is taken out and the factor of 2 actually is coming out anyway here this is already there so in fact this term remains with a $1/\sigma^2$ here this is ignored and the log pride is there so this is how you get this expression what do you do with this in general.

This expression indicates constant hyper sphere centered around the class mean where is the class mean here μ_i why I leave it for you to write this expression in two dimension write this in two dimension you will get the equation of a what is the geometry you will get go back and look

if you write this expression this part forget the constant term here this if you write in two dimension will actually give you the expression of a you should be able to extrapolate the statement which was there at the bottom of the slide.

We will go back to and look at the slide if these are constant hyper spheres in 2d this should indicate what is the projection of a sphere in 2d projection of a sphere in 2d louder you take a sphere in three dimension projected on the two-dimensional world simple projection take a ball think of a ball what will the figure indicate geometry in 2d it should be a circle okay so that is an equation of a circle which you get into D it will be a sphere in 3d and hyper spheres okay.

But more of this later on because you will still looking at so this is in general appears to give a linear additional Modric because a nonlinear term here the quadratic term here is what this will do okay.

(Refer Slide Time: 18:44)


If Σ is a diagonal matrix, with equal/unequal σ_{ii}^2 :

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_d^2 \end{bmatrix} \text{ and } \Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/\sigma_d^2 \end{bmatrix}$$

Considering the discriminant function:

$$G_i(X) = -\frac{1}{2}(X - \mu_i)^T \Sigma_i^{-1}(X - \mu_i) - \frac{1}{2} \log(\Sigma_i) + \log[P(C_i)]$$

This now will yield a weighted distance classifier.
Depending on the covariance term (*large or small spread/scatter*), we tend to put more emphasis on some feature vector components than the other.

 **This will give hyper-elliptical surfaces in R^d , for each class.**

But more of this later on so this is an example of the diagonal covariance matrix and this is the case when σ_1 is not equal to σ_2 then of course you have to write the inverse of the Covariance matrix in this form of course if we ignore the subscript that means the individual variants terms are same you can have the same term along all of this in fact I can take this constant term out and say this is an identity multiply by $1/\sigma^2$ which you have done little bit earlier.

So considering the instrument function we had ignored this a little bit earlier in general this will yield a weighted distance classifier and depending upon the covariance term large or small

scatter speed of this we tend to put more emphasis on some feature vector component than the other this will give in general hyper elliptical surfaces in our d in D dimensional space for each look at this expression.

Now with this covariance matrix how is this different from the previous covariance matrix σ_1 is equal to σ_2 equal to σ_3 and σ_d remember the σ is still same for all classes but the terms are different now in the previous case the diagonal terms were all same so those who are giving hyper spheres nonlinear additional boundaries we are approaching the case where we can get nonlinear decision boundary because of the quadratic term.

And the decision boundaries will be spheres in the case when you have equal, equal terms or identical terms as it is called identical terms along the diagonal if they are unequal as given here if they are unequal you will have hyper ellipsoidal surfaces hyper ellipsoidal surfaces if the values are unequal if these are all are same a special case ellipse becomes a sphere 2d so circle or ellipse in 3d.

You have sphere ellipsoids and in higher dimensions you have hyper spheres or hyper lapse and we will discuss this in detail in the next class of nonlinear edition boundaries with examples of non-linear at least in 2d.

We can show this sort of thing carrying on with the discussion of the decision boundaries let us assume if all the class priors are same well I must want you a little bit that sometimes we are writing VP of WI but W is wait so let us stick to see I for all k then eliminating the class independent term there which is basically if you look at this term so these two are ignored you are just left with the let us consider on this term which is actually telling you that this covariance term tells you where it should more give more importance towards a particular dimension or not because this is what will dictate which dimension has more importance okay.

(Refer Slide Time: 21:37)

More general decision boundaries

Take $P(C_i) = K$ for all i ,

and eliminating the class independent terms yield:

$$G_i(X) = (X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i)$$

$$\bar{d}_i^2 = (X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i) = -X^T \Sigma_i^{-1} X + 2\mu_i^T \Sigma_i^{-1} X - \mu_i^T \Sigma_i^{-1} \mu_i$$

$$G_i(X) = (\Sigma_i^{-1} \mu_i)^T X - \frac{1}{2} \mu_i^T \Sigma_i^{-1} \mu_i$$

as $\Sigma_i = \Sigma$, and are symmetric.

Thus, $G_i(X) = \omega_i^T X + \omega_{i0}$



where $\omega_i = \Sigma^{-1} \mu_i$ and $\omega_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i$

So this is the term and if you expand that in this particular form you can write this expression in this form and this can be now written in terms of g_{ix} as this provided you can switch off this term you can switch off this term if all the corresponding covariance matrixes are same for all classes now what you are saying is this is not a constraint which we are putting that the covariance matrix is diagonal not it can be any arbitrary covariance matrix mayor may not be diagonal what we are saying is it is same for all classes how unrealistic that can be just to show okay it may be possible in certain situations.

But what I am saying is features such as height weight color size spatial extent for two different types of fruits and flowers are all saying okay and they are symmetric of course it is symmetric matrix but they are all same so if it is same you can switch off this I and if you can switch off this I you can switch off this particular term you will be left with only these two terms here and typically GI is di by two.

So the two will go off you will not to have you have a minus half factor here so we will concentrate on this term it has come back again this is not a new term one or two classes back couple of classes back we are talking about this term the last class we discussed at length which gave us to the concept of Perceptron learning for linear de Chaumont is the only difference now is earlier this weight was based on only the mean now the covariance term has come and sit here the covariance term is coming in verse of the covenants matrix to be very precise is coming and sitting here.


These weights I have just changed the notation too short of a ω here small W you can treat this and this is an inverse of the covariance matrix okay so we have discussed this case earlier when this was an identity matrix and when this was the covariance matrix was an identity matrix we just had this μ as the weights and μ transpose ρ without this term is what we had for linear addition we are still in having linear discriminant functions we can still have linear decision boundaries but the covariance term is coming and sitting here it is maybe an arbitrary coverage metric just it is symmetric that is all.

(Refer Slide Time: 24:23)

Thus, $G_i(X) = \omega_i^T X + \omega_{i0}$
 where $\omega_i = \Sigma^{-1} \mu_i$ and $\omega_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i$

Thus the decision surfaces are hyperplanes and decision boundaries will also be linear (use $G_i(X) = G_j(X)$, as done earlier)

Beyond this, if a diagonal Σ is class-dependent or off-diagonal terms are non-zero, we get **non-linear DFs, DRs or DBs.**



So this is what we have and we analyze this for the rest of the class and move word to non linear decision boundaries in the next class so then it is the last discussion on last part of the disc on discussion on linear decision boundaries because these type of g_{ix} will give us dr's DVS which are hyper planes and they will be linear by exploiting this constraint as we have done earlier at the beginning of the class beyond.

This if you have whenever diagonal σ which is class dependent remember this is coming out of the constraint that day covariance matrix is class independent you will go back and look at this constraint look this is class in that means you keep changing go from class one to two I to J is the same covariance matrix.

So it is class independent it is class independent if it is class dependent or are they often of diagonal terms are non zero you will typically have non linear decision functions discriminant functions decision regions are decision boundaries non-linear discriminant functions and decision boundaries are the correct way of saying there is no nonlinear decision region but whatever it means is that discriminant functions and decision bond is actually give you dr's.

So the boundary of this regions will be nonlinear if you have class dependent covariance matrix and the off diagonal terms typically have more roles to play but as less you as long as you are having class independent covariance matrix term in a or not including identity matrix which is a special case of a diagonal covariance matrix you will have linear DF linear TV so again to repeat what are the special cases of covariance matrices in which you will have linear TVs or linear discriminant functions one is if the, if the covariance matrix is class independent that is all straightaway first of all even if this class dependent one well in some sense.

You can say a special case of what we consider the identity matrix that is also class independent so as long as it is class independent you will have linear discriminant functions if it is class dependent and diagonal you will still have hyper ellipsoids or hyper sphere which gives rise to the simplest case of non linear discriminant functions which we will discuss next.

(Refer Slide Time: 26:57)

**Effect of class Priors -
revisiting DBs in a more general case.**

$$P(w_i | \bar{X}) = \frac{P(\bar{X} | w_i)P(w_i)}{P(\bar{X})}$$

$$p(X | w_i) = \frac{1}{\sqrt{\det(\Sigma)(2\pi)^d}} \exp\left[-\frac{(X - \mu)^T \Sigma^{-1} (X - \mu)}{2}\right]$$

$$g_i(x) = \ln p(x|w_i) + \ln P(w_i)$$

$$g_i(x) = \frac{-1}{2}(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{d \ln 2\pi}{2} - \frac{1}{2} \ln |\Sigma| + \ln P(w_i)$$

Let us proceed with the discussion on the decision boundaries with their and discuss the effect of class priors in a more general case remember this to just revisit the equations this is what is this equation we have seen this many times base theorem okay under the Bayes theorem this is the posterior remember the C changes Wi be careful this is not the same as the class w so this the same as a class w but not the weight this is the conditional density function which we are talking about here.

This is the x this is the multivariate Gaussian density which we started few classes back the normal distribution okay and if you take GI to be this plus the cross prior that means you are taking the this term which is the class conditional density function as given by this plus the class prior you can expand it using this expression that means what you are doing this log prior is here and we have done this many, many times take the log of this expression you have resistance plus these two terms which one of them is a constant anyway.

The other could be class dependent or class independent since we are talking about linear decision boundaries what assumption did we make about the covariance matrix it is glass independent so there is no subscript here you can see although I have put a subscript here but the covariance matrix is same so we can switch it off here carrying on.

So we can switch off these two terms because they are class independent we are left with the Mahalanobis distance function that is Euclidean distance waited by the covariance matrix it is same for all classes plus the class file what is this term doing if it is the classifier is not same

remember what is class prior you are discriminated between two types of fruits say mangoes vs apple both are.

Let us say seasonal fruits depending upon certain time of the year you may have more apples let us say in winter in summer or rainy season you may have more mangoes so the class priors will change they may not be the same that is what we are saying what how does this affect the decision boundary.


(Refer Slide Time: 29:10)

Canceling in class-invariant terms:

$$G_i(X) = -\frac{1}{2}(X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i) + \log[P(W_i)]$$

CASE A. – Same diagonal Σ_i with identical diagonal elements.

$$g_i(X) = \frac{-1}{2\sigma^2}[(X - \mu_i)^T (X - \mu_i)] + \ln P(w_i)$$

$$g_i(X) = \frac{-1}{2\sigma^2}[X^T X - 2\mu_i^T X + \mu_i^T \mu_i] + \ln P(w_i)$$


Let us look canceling the class invariant terms we have just left with these two remember this is the main one which is responsible for all my linear or non-linear decision boundaries. But under the case same diagonal σ that means we are switching of I diagonal elements class independent what do we have for the expression so we are switching of this diagonal this is the expression which we Plus lock prior this is the case a we are discussing diagonal we have done this already this is a diagonal matrix.

We can take out this half is here inverse of the covariance matrix will this term and then we have the class prior when we break open the expression this is the expression which we have nonlinear

term here the linear term particularly here as a function of X we will analyze this expression in the next class in the next slide.

(Refer Slide Time: 30:00)


$$g_i(X) = \frac{-1}{2\sigma^2} [\cancel{X^T X} - 2\mu_i^T X + \mu_i^T \mu_i] + \ln P(w_i)$$

Thus, $g_i(X) = \omega_i^T X + \omega_{i0}$

where $\omega_i = \mu_i / \sigma^2$ and $\omega_{i0} = -\frac{\mu_i^T \mu_i}{2\sigma^2} + \ln P(w_i)$

The linear DB is thus: $g_k(X) = g_i(X), k \neq i$

which is:

$$(\omega_k^T - \omega_i^T)X + (\omega_{k0} - \omega_{i0}) = 0;$$


Analyzing this expression again the nonlinear term here the linear part here why are you switching of this because if you move from one class to the other change I to J you say this will change this will change as well this will change this will not change. So this is a class independent term again it is a quadratic term but class independent term okay and why it has come because the class independent covariance matrix the same σ which is diagonal okay.

So once you do this, this is what you are and this is same expression which we had just a few slides back sometime back but instead of writing inverse of the covariance matrix here multiplied by the mean you are able to right now this is - σ^2 sorry divided by the variance, divided by the variance here the same thing which comes here remember the two and two, two cancels out here so this divided by σ^2 you will transpose in general you will have it as a inverse of the covariance matrix here basically this term is indicating the inverse of the covariance matrix but the class power is still here earlier we are ignored this the linear test in boundary is now this we have seen at the beginning of the class today.

That it can be written in this where individually each of these terms are given by this expression so it is $W^T X - d$ where K and L at for the two different classes they are not identical and the bias term is also written here.

(Refer Slide Time: 31:32)

The linear DB is thus: $g_k(X) = g_l(X), k \neq l$

which is:

$$(\omega_k^T - \omega_l^T)X + (\omega_{k0} - \omega_{l0}) = 0;$$


Prove that the 2nd constant term:

$$(\omega_{k0} - \omega_{l0}) = (\omega_l - \omega_k)^T X_0; \text{ where}$$

$$X_0 = \frac{1}{2}(\mu_k + \mu_l) - \sigma^2 \frac{\mu_k - \mu_l}{\|\mu_k - \mu_l\|^2} \ln \frac{P(\omega_k)}{P(\omega_l)}$$

Thus the linear DB is: $W^T (X - X_0) = 0;$

where, $W = \mu_k - \mu_l$ **Nothing new, seen earlier**



Let us observe this expression in the next slide the linear this decision boundary is given by this which is basically this and I leave this is an exercise for you to prove that this difference in the bias term can also be written as a function of this which is like this where the X_0 this is interesting is given as this particular term look at this term I am leaving this is an exercise for you to prove it analytically you can prove this with a little bit of jiggle maybe about a half a page of derivation okay.

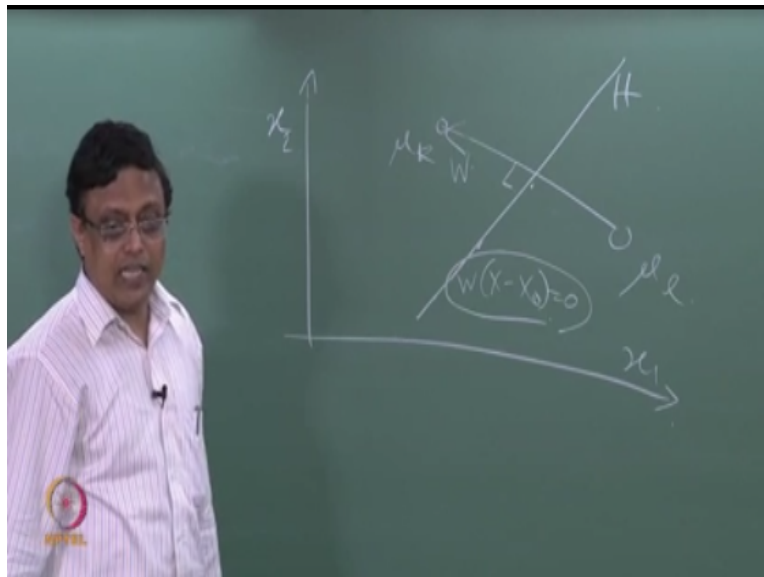
And henceforth if you can write this term in terms of this and substitute back here this expression can be written this is nothing new we had this to two classes back w into x minus d is equal to zero from where we derive the Perceptron the hyper plane this is a plane passing through the origin or plane passing through a particular point the only thing is what is this w it is the it is the difference of these two main vectors what is the mean vectors.

If there are two main vectors this is the vector difference between the two mean vectors that is the w s will be remember the w is the normal to the hyper plane the normal to the hyper plane is

now the line joining the two means and if it is so the hyper plane will be normal to the line joining the two means.

I repeat again the W which you see here we had seen this earlier two classes back that if this is a hyper plane this is the W and now what you are saying is this w is the vector joining the two class means so it is a class mean one here in the class mean to hear this is the YW this is the hyper plane will have it is in a diagram coming up in the next slide but I hope you got the justification of that maybe I will go to the board and just draw that for you because I may not have a slight ready.

(Refer Slide Time: 33:37)



So if this is a μ_k and this is μ_L then $\mu_k - \mu_L$ okay this will be the w the vector direction will be here or here although it strictly does not matter what do you think it is given as μ_k the expression says it is $\mu_k - \mu_L$, $\mu_k - \mu_L$ it will be pointing towards this towards me this will be W and we said that the W is orthogonal to the hyper plane.

So this is the hyper plane H I am drawing in 2D this is my attend space or feature space not the weight space like we did in per case of Perceptron so this should be x_1 this will be x_2 dimensional feature space you can visualize in third dimension also this is my hyper plane and which I wrote that equation here $W \cdot X - X_0$ hyper plane which is equal to if you select points here in this airplane this will be equal to 0.

So this is the expression which defines this hyper plane this δw is normal to that okay and you could ask me where is this X_0 it is a point here it is a point here on this plane okay and let us look at the expression for X_0 now in this slide look at that expression of the X_0 at X_0 t is a point on H and if you look at this expression of X_0 look at this term can you tell me where is this term it is the average of the two means go back to the board these are the two means me uk μ_L average of the two means will be the central point.

Let us say so for the time being I let this up it is not till the some average value of the two means and it is seems to be an approximation because it is not the full expression this containing the first term in the slide so let us go back to the slide so what I have marked there and the board is basically the yeah this term here but there is another factor which is based on the class mean we will explore that now we will explore that now what is the effect of class priors on that point X_0 and X_0 is a point which is actually lying on the hyper plane okay.

(Refer Slide Time: 36:01)

CASE - A. - Same diagonal Σ , with identical diagonal elements (Contd.)

Case of Linear DB:

$$W^T (X - X_0) = 0;$$

where, $W = \mu_k - \mu_l$

$$X_0 = \frac{1}{2}(\mu_k + \mu_l) - \sigma^2 \frac{\mu_k - \mu_l}{\|\mu_k - \mu_l\|^2} \ln \frac{P(\omega_k)}{P(\omega_l)}$$



So this is the X_0 so we had this w is very simply this is the vector it is the line joining the two means vector from one mean to the other one orthogonal to the hyper plane and hyper plane passing through X know not because that will actually decide my decision boundary it should have been in between X_0 will be equal to this under what condition tell me what condition the expression of the second term do not tell me σ equal to zero what, what constraint will make this vanish do not tell me μ should be equal to μ_L do not tell me this will be equal to zero.

These two probability terms if their priors are same class priors are same that means I get equal number of mangoes as equal number of apples and if I can do that then if that happens this term vanishes forcing X_0 will be equal to the mean of these two classes that means my hyper plane is strictly in the middle of the two class means in this case hyper plane H is the decision boundary earlier it was discriminant function.

(Refer Slide Time: 37:14)

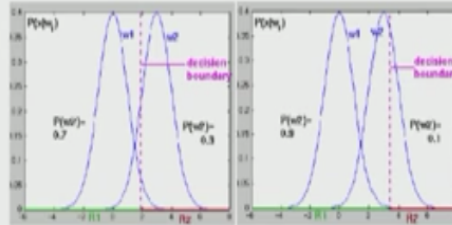
CASE - A. - Same diagonal Σ , with identical diagonal elements (Contd.)

Case of Linear DB:

$$W^T (X - X_0) = 0;$$

where, $W = \mu_k - \mu_l$

$$X_0 = \frac{1}{2} (\mu_k + \mu_l) - \sigma^2 \frac{\mu_k - \mu_l}{\|\mu_k - \mu_l\|^2} \ln \frac{P(\omega_k)}{P(\omega_l)}$$



But now it is a difference of 2 g's which is a linear decision boundary we will have some examples let us look at the simple example which I have borrowed from the website I will give that reference a little bit later on look these are two Gaussian functions drawn here and if you look at the Bayes theorem ignoring the class priors this point at the center of intersection of these two should have been the actual value of X_0 . X_0 is a point here mind you average of the two means look at this is one mean this is the other mean class 1 w_1 class 2 w_2 two Gaussian distributions intersection of this at the middle should have been the point X_0 corresponding to this expression.

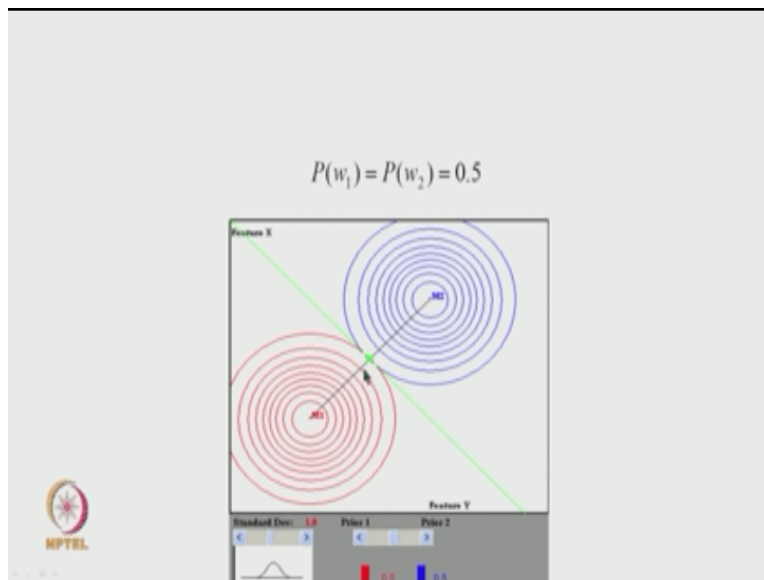
But in this figure 1 class prior is more than the other one so this term will not vanish it pushes the decision boundary more towards the other class ideally I would have loved to have the decision boundary exactly in the middle of the two Gaussian functions provided the class priors are the same but if the class priors are more what does it mean in reality there are more apples than mangoes there are more flowers than fruits there are more men than women there are more forests than mountains okay.

There are more a landscape there is more water than deserts okay there are more people smiling then more people anger I am talking about discrimination of expressions in the human face these are examples only and in certain cases of class priors could be more class prior pushes the decision boundary this will be an additive corrective term to this value of x not the point is push there it could go to the other side if this value of the class player of class two would have been

more increase the class prior more you can see how far it goes why because this value is .9 this is 0.1 this is 0.7 and 0.3.

So this value of class power is now more than this it pushes the decision boundary from this point here more towards the other side why because this term will be now larger because this numerator is more than it is nominative making this term more for this figure than the left hand side this is a case of a DBE with same diagonal σ same for all classes all right identical diagonal elements that is why you could have taken out that σ but one class prior is more than the other one the decision bond is not lying in the middle it is a point in 1d we will show examples now in 2d what will become a line and you can extend that to a plane in three dimension hyper planes in higher dimension.

(Refer Slide Time: 40:00)



This is a java applet available as an open source which you can play and manipulate the individual terms of variances and the class priors so there is a slight bar to change the standard, standard deviation is equal to one for both, both equation a deviation identical covariance matrix 2d but what happens both identical class spares.

So this is this is now look at the DB here this is the type of plane H which is a DB where it is sitting this is that X_0 Y because it is exactly the midpoint of these two means μ_1 μ_2 m_1 and m_2 so assume this to be μ_1 μ_2 classes equal priors .5, .5 each as given in this two bars at the bottom class fires are same so ideal condition so class projects are same.

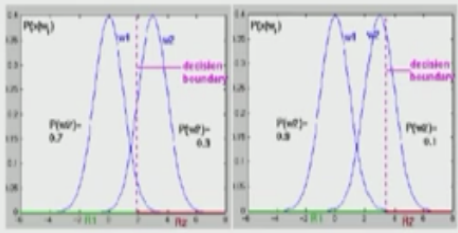
(Refer Slide Time: 40:53)

CASE - A. - Same diagonal Σ with identical diagonal elements (Contd.)

Case of Linear DB:

$$W^T (X - X_0) = 0;$$

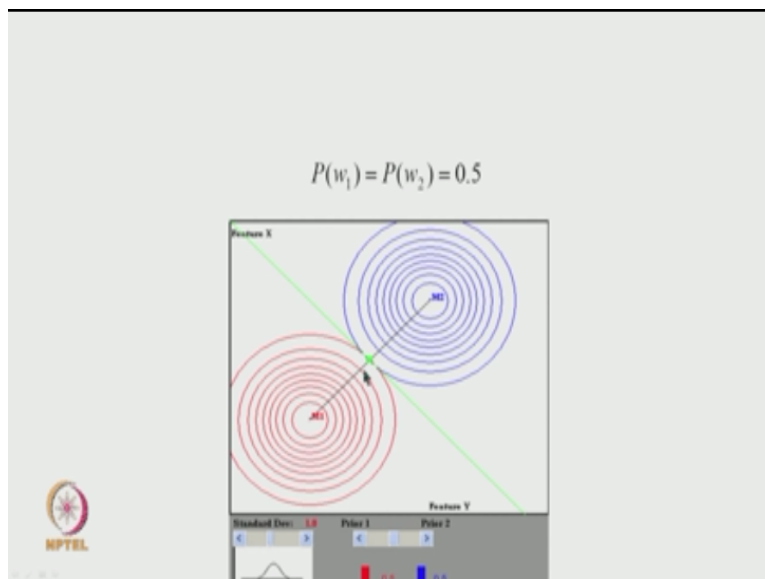
where, $W = \mu_k - \mu_l$

$$X_0 = \frac{1}{2}(\mu_k + \mu_l) - \sigma^2 \frac{\mu_k - \mu_l}{\|\mu_k - \mu_l\|^2} \ln \frac{P(\omega_k)}{P(\omega_l)}$$


NPTEL

If you go back this terms cancels out you exactly have the X_0 at the bisector point with μ_1 .

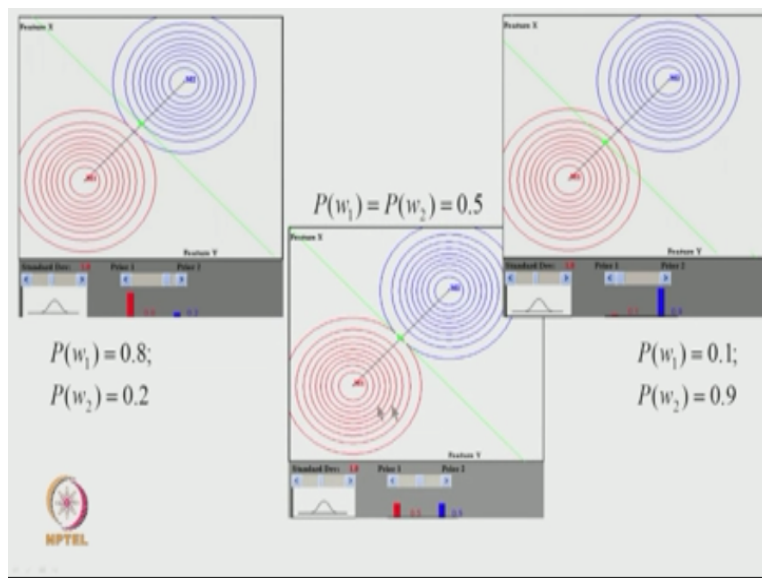
(Refer Slide Time: 41:01)



And now DB is actually a perpendicular bisector of the line joining the means DB is a perpendicular bisector as a special case law of the class joining the class means if the class priors

are same identical covariance matrices that means class independent covariance matrix what I will do now first I will change the class priors and see what happens to this diagram so now what we will do is we will first change the class priors and see what does it cause in what the effect is the effect in the location of the DB that is number one.

(Refer Slide Time: 41:46)



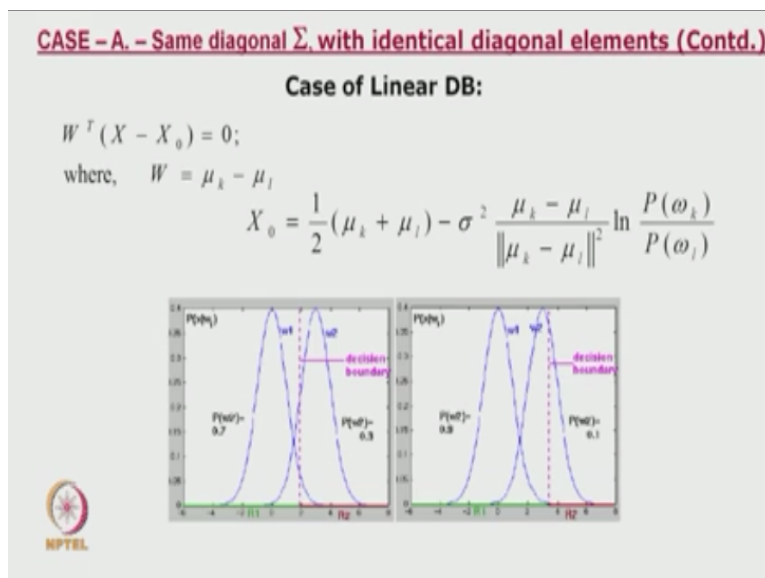
And then we also of course change the variance terms and see if that has an effect as well it should have an effect as well let us look at this diagram on the right hand side what I have done is maintained the same value of the variance which is equal to one both here as well as there but I have changed one of the class prior look at the value here it is given and this blue bar indicating that the value is much large is 9. And the other class prior is .1 earlier the both were identical what it has done earlier the decision boundary was at this point or the line here as given in this diagram.

It has pushed it towards which class it had pushed it towards the other class which is a less class prior class one had the less class DB has moved towards that earlier it is at the center because the class pass were same the class prior for class 1 is now .8 indicated by the big red bar there the other value is point to which is much less here look what it has done because the class pair of class 1 is higher it is push the decision boundary from this point has given here to more towards the cluster why the second term the decision boundary will now change its sign and move in the other direction.

Remember it was a log ratio of class priors so depending upon which term is more you will have a positive value or a negative numerical value pushing the, the separating hyper plane as the DB towards one particular class which particular class does it move towards the class mean which is having me less class prior you see the less class prior the DB is moving towards their class mean here it is moving towards class to mean because that has the less class prior so this is the effect of class prior in base decision rule under the normal distribution all that has been taken all right the class prior is not allowing the decision boundary to be strictly at the perpendicular bisector.

Whether it is a line or a plane or even a point in 1d remember the previous slide where it pushed it towards the class mean or even further away.

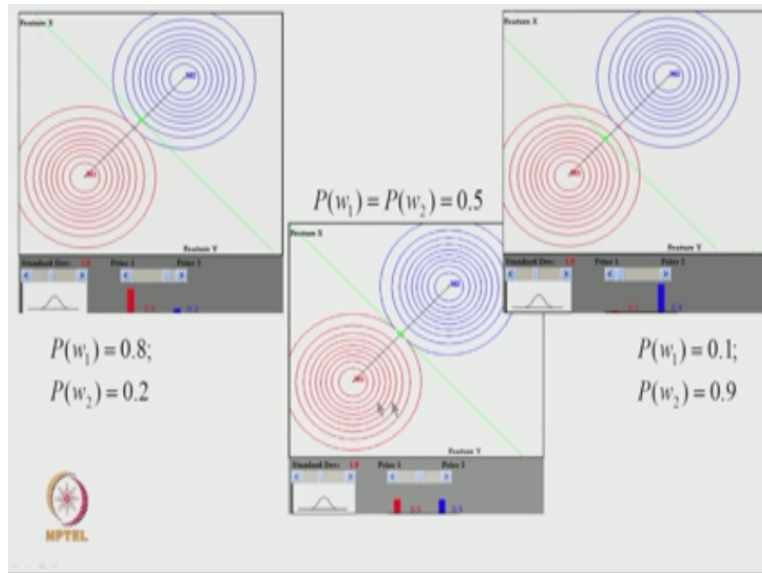
(Refer Slide Time: 43:56)



Let us go back you look at this example here you look at this it has even gone past the class mean of class 2 because it is really very, very large the class prior for class I think there is a typo

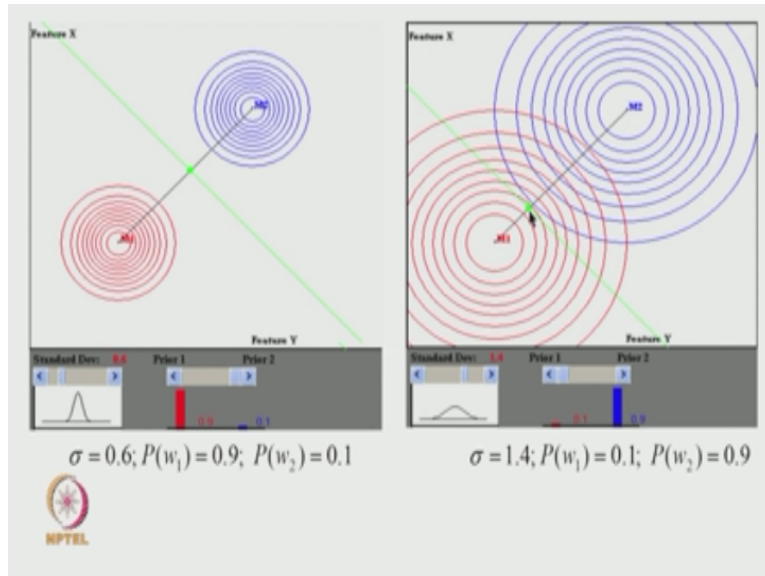
here this should be w one class pair for w one is .7 this is .9 so it is gone pass the class prior for the second class itself though the mean is here somewhere cannot pass then so if it is too big it can even go past the class means whether it will go pass the class mean or not depends on the individual variants.

(Refer Slide Time: 44:37)



So what I will do now is keeping these unequal class priors we will change the class variance now the scatter we will reduce in the next slide.

(Refer Slide Time: 44:44)



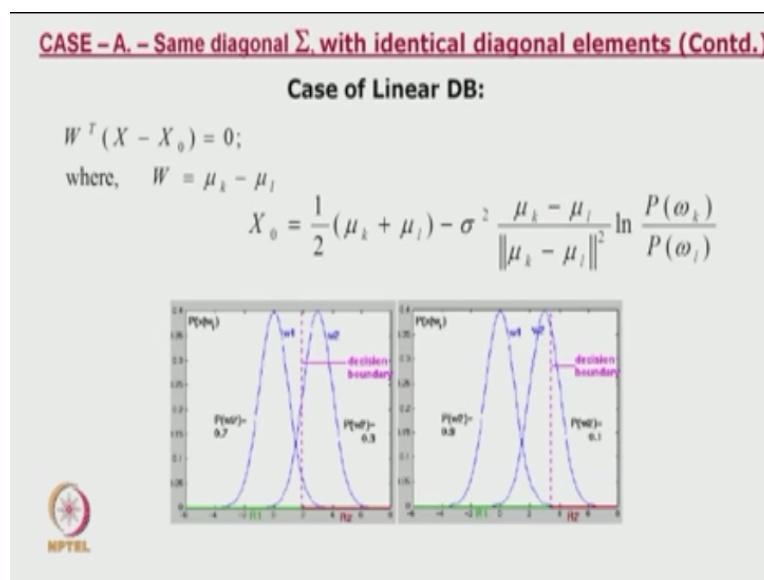
Look this is the effect of class priors unequal cross priors but a lesser scattered this has moved but not that much that means the variance has a role to play this is standard deviation let us go back to the expression there look this is multiplying factor so if this is larger this effectively more pronounced if this is less this will be lost pronounced so that is why in these cases the effect is more and if you enlarge it more if the variance is more than will be a class overlap plus the decision boundary will skip moving more towards the class mean.

And even further away remember it will always orthogonal the plane will be orthogonal the separating hyper plane or the DB in this case will be normal to the line or vector joining the two class means or the vector joining the two classrooms will be orthogonal to the plane that will always be the case the plane will remain orthogonal to this vector but it may go away if the ratio of this is higher or the variance terms increases here we have reduced the variance the shift is there is not at the perpendicular bisector.

Because its prior is less let us look at this case here this is the case where the variance is really large it has come the separating hyper plane has come more closer to the class minute may go further away if you increase this or reduce this further and change this increase it more towards a value 1 so this shows an effect of where the decision boundary linearization boundary will be located depending upon two factors the spread or scatter of the individual scatter matrices or the variance of the features.

And the class pass this will decide where the decision boundary will be located remember always it will be the separating hyper plane will be orthogonal to the line joining the two class means or the line joining to class planes will be orthogonal to the hyper plane but its position dictated by X_0 will be dictated by the second term in general it will be in the perpendicular bisector at the center point of the line joining the two class means but it will go up and down more towards the class mean depending upon the ratio of the class prior multiplied by a factor which also depends on the variance plus there is another factor in between.

(Refer Slide Time: 47:15)



Let us go back to the expression normalized by see this is the vector this is also a vector ton so this, this is a unit vector along the line joining the two class means so the numerical value dictated by the log prior show and the variance or the scatter or spread dictates where you are we will stop with the discussion today on linear decision boundaries will move towards nonlinear, non-linear additional boundaries more where we will have we will have class dependent variances or covariance matrices and the effect of off diagonal terms also as long as it is unequal as long as it is unequal we will have non-linear decision boundaries thank you will come back to the next class.

Online Video Editing /Post Production

K.R.Mahendra Babu
Soju Francis
S.Pradeepa

S.Subash

Camera

Selvam
Robert Joseph
Karthikeyan
Ram Kumar
Ramganes
Sathiaraj

Studio Assistants

Krishankumar
Linuselman
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 Krishnamurthy

IIT Madras Production

Funded By
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in

Copyrights Reserved