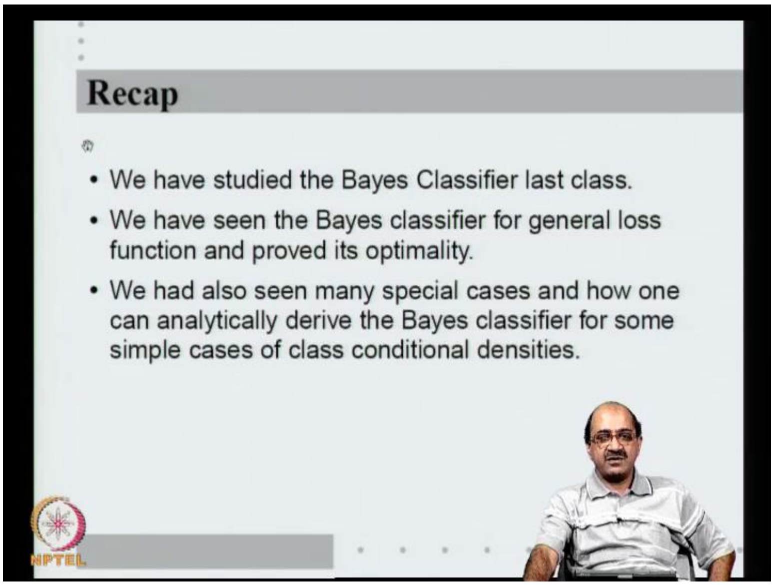
Pattern Recognition
Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 04
Estimating Bayes Error Minimax and Neymann - Pearson Classifiers

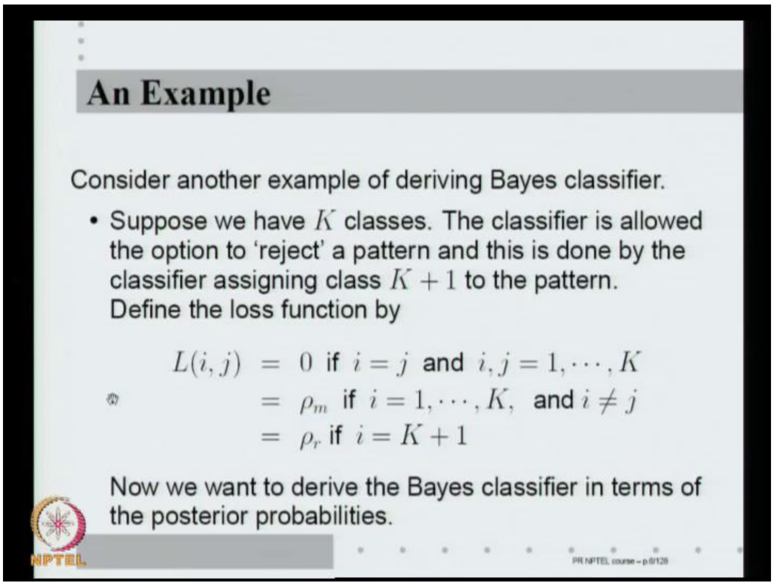
Hello, let us continue with the next class just briefly recapitulate last class, we have studied the Bayes classifier in detail.

(Refer Slide Time: 00:29)



We have derived the Bayes classifier for the general loss function and we have also proved the optimality of the general Bayes classifiers, optimality in the sense that no other classifier can achieve lower risk than Bayes classifiers. So for risk minimization, we showed that the Bayes classifier is the optimal classifier and we have also seen several special cases of what the classifier looks like, for example for 0,1 loss function how it looks like and also for special classes of lass conditional density such as normal and so on. We have analytically derived the Bayes classifiers, so we seen some examples of how one derives Bayes classifier. So, this class we will start with another simple example of deriving Bayes classifier in a slightly different setting. Then we will look at Bayes error, how to calculate the error of Bayes classifier and then, we will move on to a few other criteria for classification.

(Refer Slide Time: 01:26)

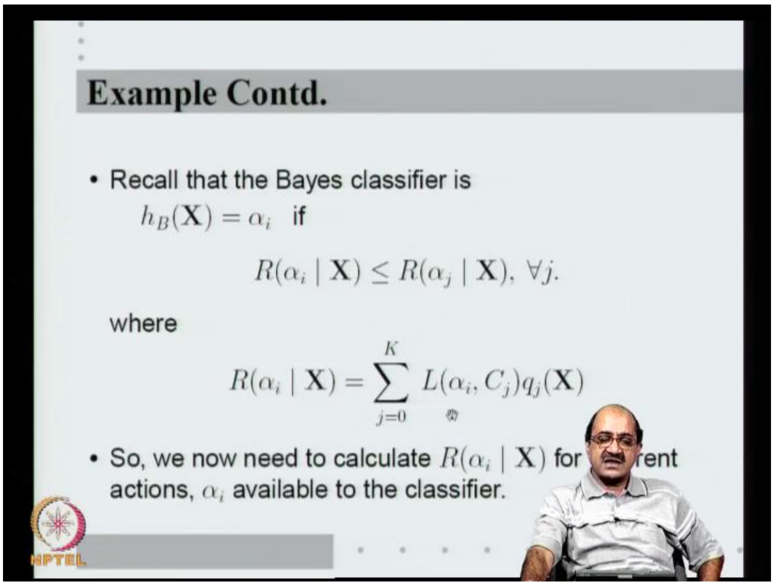


So let us start with an example, let us say we have K classes and as I said when, we did the Bayes classifier last class that the actions of the classifier need not necessarily have to be class labels they can be more than the class labels. So, let us take an example where, the classifier is allowed the option to reject a pattern that is you look at the pattern and say no I can not classify and let us assume that this is done by the classifier assigning a class K plus 1 to the pattern.

So, this K classes, so class label takes values 1 to K where, the classifier actions take values from 1 to K plus 1 and will interpret the action K plus 1 of the classifier as the classifier rejecting the pattern. It is rejecting the pattern because, may be is does not have enough confidence to classify, so now for this case let us put some loss function. So, as you know the loss function has 2 arguments, the first argument is what I call, i here is the actions of the classifier, second argument of the class labels. So, I can take values 1 to k plus 1 where as, j takes vales only 1 to K. So, is this is something like a 0, 1 loss function. So, if I did correct, so if both i and j belong 1 to K, that means, the class the classifier has classified the pattern to one of the classes and if i is equal to j then loss is 0.

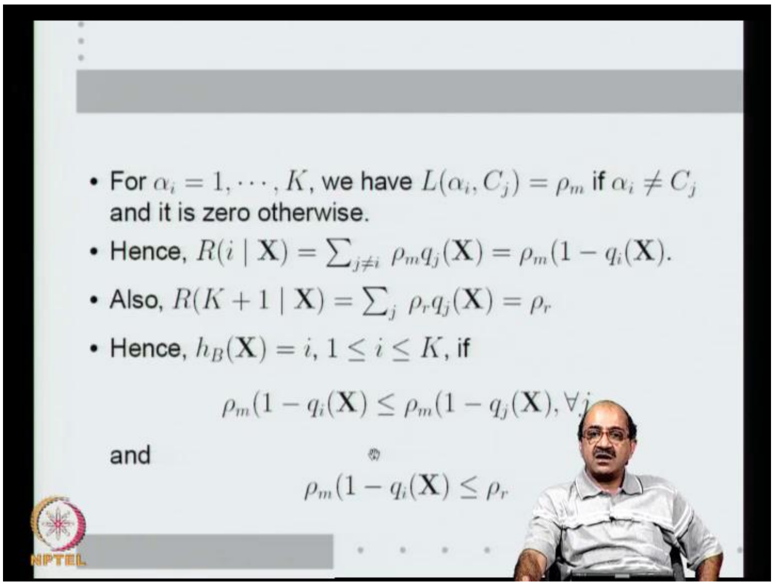
Similarly, if i is 1 to K that is the classifier has decided to call a particular class. But, i is not equal to j then, we have misclassified let us call that cos rho m, m for misclassification and there had irrespective of j, if i is K plus 1, that is the classifier has decided to reject the pattern the loss is rho r. So, this is the loss function, so if I misclassify a pattern i i i suffer a loss of rho m, if I reject a pattern i suffer a loss of rho r, ofcourse, correct classification is 0 loss given this can we now derive the Bayes classifier in terms of the posterior probabilities as usual.

(Refer Slide Time: 03:24)



Recall that the general Bayes classifiers is given an X, the Bayes classifier will say alpha i or i, if the risk of that action is less than the risk of all other actions, where risk of that action given X is defined as, L of alpha is C j q j X, where j the or C j are the class labels. So, j goes from not 0 to 1, I am sorry 1 to K, it is not 0 to K, but 1 to K, I am sorry about that. So, that K classes now, we have we we will calculate this risk for various actions and say, when is classification pattern and when is reject pattern.

(Refer Slide Time: 04:03)

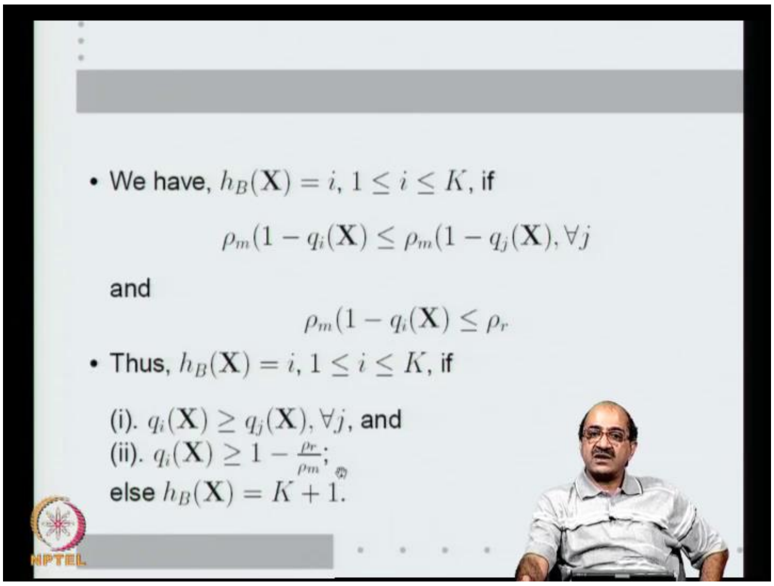


When the classifier wants to take actions 1 to K that means, classify into one of the classes then the loss is rho m, if i misclassify, otherwise it is 0. So for i i between 1 to K the risk of i given x is rho m, q j z not equal to I, because when i correctly classify there is nonono loss, otherwise all other misclassifications have the same loss rho m.

So, now rho m comes for this examination and it becomes rho m into 1 minus q i X on the other hand, if classifier takes the action K plus 1, then for all class labels j the loss is rho r, so the risk will be summation over j rho or q j, rho r comes with a summation and summed over q j is equal to 1, so this becomes rho r. Here because, this summation over j is not equal to i, summation q j will become 1 minus q i, here this is for all j, so summation q j becomes 1. So, this becomes rho r.

So, what does it mean, when can I call a particular class or i given x, should be better than r j given x, for a all other class labels j and it is also should be better than r K plus 1 given x right. So, my my new my best classifier can say i for one of the class labels, if the risk associated with i, that is rho m into 1 minus q i should be less than or equal to risk associated with any other class label rho m into 1 minus q j, also risk associated with action i should be less than the risk associated with the action K plus 1, which is rho r. So, both these have to be satisfied for me to call class i.

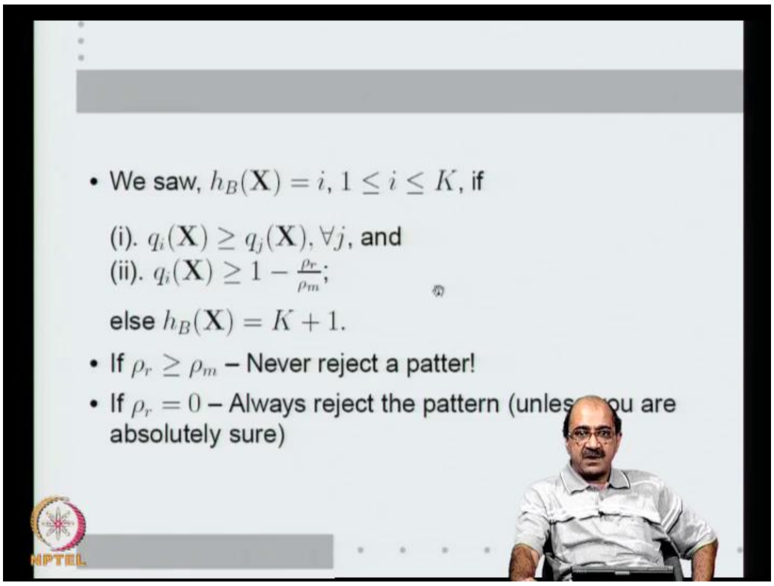
(Refer Slide Time: 05:51)



So, let us simplify this, so the Bayes classifier will, now say a class label i, if rho m into 1 minus q i X is less than rho m into 1 minus q j X for all j as well as rho m into 1 minus q i X is less than rho r. The first inequality, I can cancel rho m from both sides and first inequality becomes same as q i greater q j. So, let us simplify the second also, so I can call i, the first inequality says q i is greater than or equal to q j and the second inequality says I will i will bring rho m this side this is 1 minus q i less than rho r by rho m or if you bring q i the other way q i is greater than 1 minus rho r by rho m.

So, if both these conditions are satisfied then i can call a class i, obviously there will be some i for, which q i is greater than or equal to q j, which ever is the highest posterior probability class. But, earlier when I did not have reject the Bayes classifier simply puts it in the class corresponding with the highest posterior probability, but no that is not enough for me to call a class because, I am allowed reject not only, I should be the highest posterior probability class. But, the probability of the highest posterior probability class itself should be greater than some threshold, if this is not true then it is better to call reject. Otherwise, h B X will be K plus 1, because this inequality are not satisfied, which means the least risk will be for the action K plus 1.

(Refer Slide Time: 07:11)

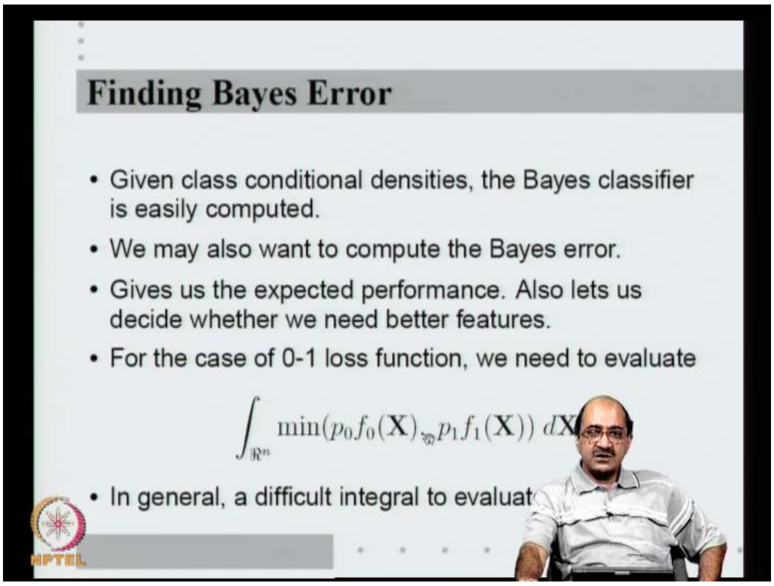


So, let us try and understand it again. So, in the reject case my new Bayes classifier will say a class label i, if this is true where, the highest posterior probability is greater than 1 minus rho r by rho m and then, I will call the highest posterior probability class. Otherwise I will call K plus 1 to understand this let us look at few special cases. Suppose, cost of rejection is greater than cost of misclassification, what does that mean, if I misclassify a pattern, I suffer less loss than, if I reject a pattern thenthen, it should not be there should be no condition under, which reject is good. Now that is what this will tell me, if rho r is greater than rho m then rho r by rho m will be greater than 1. So, 1 minus rho r by rho m, will be negative and hence q i being a probability will always be greater than this.

So, then it boils down to whole thing for some i is of that q i X is greater than q j, I will, call that i, I will never ever call K plus 1 right, so I never reject a pattern. Now consider the other extreme case, suppose cost of rejection is 0, cost of rejection is same as cost of correct classification. So, what should I do, I might just reject everything what does my derivation say if rho r is 0, I will call one of the class labels only if q i X is greater than or equal to 1.

So, what does that mean, if cost of rejection is 0, I always reject a pattern unless of course, I am absolutely sure if q i X is equals to 1, then I can call I, unless I am absolutely sure it is better to reject a pattern because, rejection cost me nothing. So, these are just extreme cases for us to understand that the senice, so this is another example of how I may derive a Bayes classifier, where classifiers actions may be different from class labels. Here, we have one extra action namely the reject option, so with this example, we will stop discussing examples of Bayes classifiers.

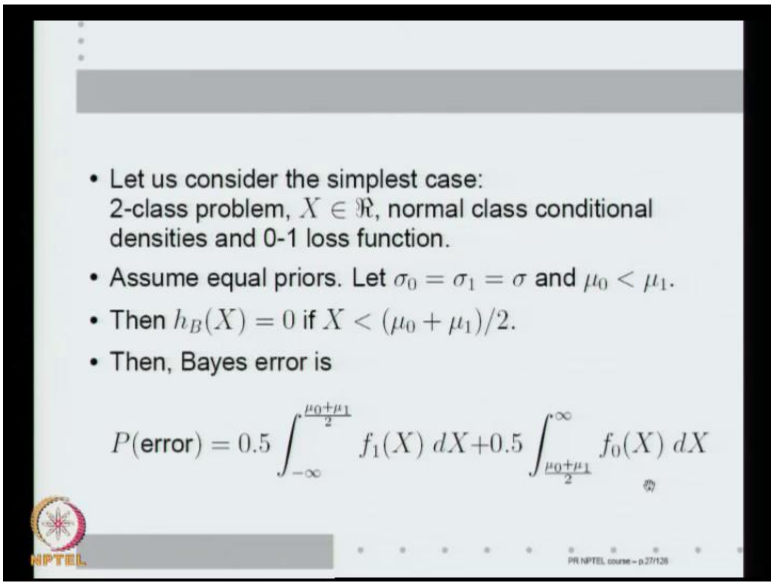
(Refer Slide Time: 09:12)



Let us move to one other important issue with Bayes classifiers, how do I find the Bayes error from all the examples, we have considered, so far given class conditional densities the Bayes classifier is easily computed right. For various loss functions, we know how to compute it, now we may also want to compute the Bayes error, because that tells me, what is the expected error of the classifier, what is the expected risk of the classifier.

Now for example, if the expected risk is not within acceptable limits accept accept expected error rate is not within acceptable limits, then I may have to rethink my whole problem in the sense, I may want to get better features because, with these features this is the best performance I can get. So, estimating or finding Bayes error is useful for me to know whether my classifier will meet the specification requirements. So for the 0, 1 loss function, we have already seen when, we derived the optimality of Bayes classifier, that this is the error. So, we have to integrate minimum of p 0, f 0, p 1, f 1 over the entire feature space to find the error rate of the classifier, this is the probability of misclassification by the Bayes classifier. In general it is a very difficult integral to evaluate, because it is a mean inside the integrand.

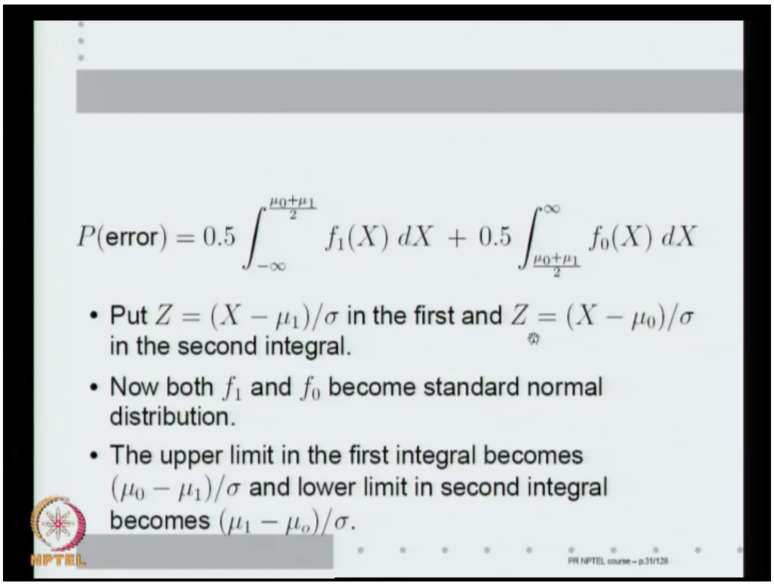
(Refer Slide Time: 10:27)



First let us look at a very simple case, let us consider one dimensional feature space, 2 class problem 0, 1 loss function, assume equal priors assume normal class conditional densities with equal variance. This is about the simplest special case, you can consider and for notational convenience, let us assume the mean of class 0 is less than mean of class 1.In this case, we have already known, because the both variances are same the threshold for the Bayes classifier is midway between the 2 means. So, if X is less than mu 0 plus mu 1 by 2, I will call class 0, it it is greater than that, I will call class 1, this is the Bayes classifier. So, what will be the error of the Bayes classifier.

So, for X less than mu 0 plus mu 1 by 2, I will always call 0. So, the probability of class 1 patterns coming with X less than this is 1 half of the error. So, probability that X belongs to class 1 and X is less than mu 0 plus mu 1 by 2 is this probability this integral, integral of the f 1 the class 1 class conditional density over this range. Similarly, the other error occurs, when a class 0 pattern comes with X greater than mu 0 plus mu 1 by 2. So, this is the Bayes error in for normal density, it is easy to evaluate, so let us evaluate it.

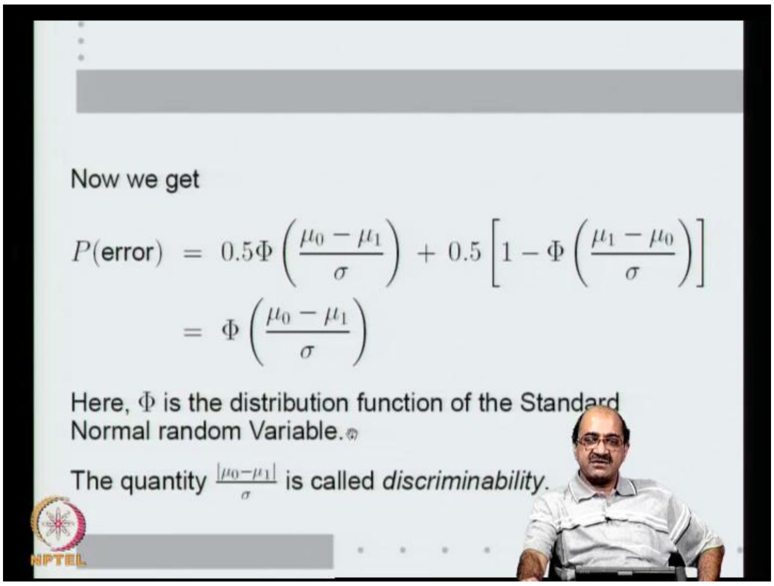
(Refer Slide Time: 11:49)



So, this is the error integral f 1 is normal with mean mu 1 and variance sigma f 0 is normal with mean mu 0 on variance sigma. So, using the standard substitution by change the variable X in this integral to Z, where Z is X minus mu 1 by sigma, then this density assume the form of the standard normal density. Similarly, for this integral because, f 0 is normal with mean mu 1 on variance sigma, if I use the substitution Z is equal to X minus mu 0 by sigma.

This becomes a standard normal density integral, now what happens to the limits, when a put Z this, when X goes up to mu 0 plus mu 1 by 2, Z goes to mu 0 minus mu 1 by sigma and similarly, here. So, this becomes an integral of the standard normal density over minus infinity to mu 0 plus mu 1, mu 0 minus mu 1 by sigma and this becomes from mu 1 minus mu 0 by sigma to infinity of another standard normal.

(Refer Slide Time: 12:52)



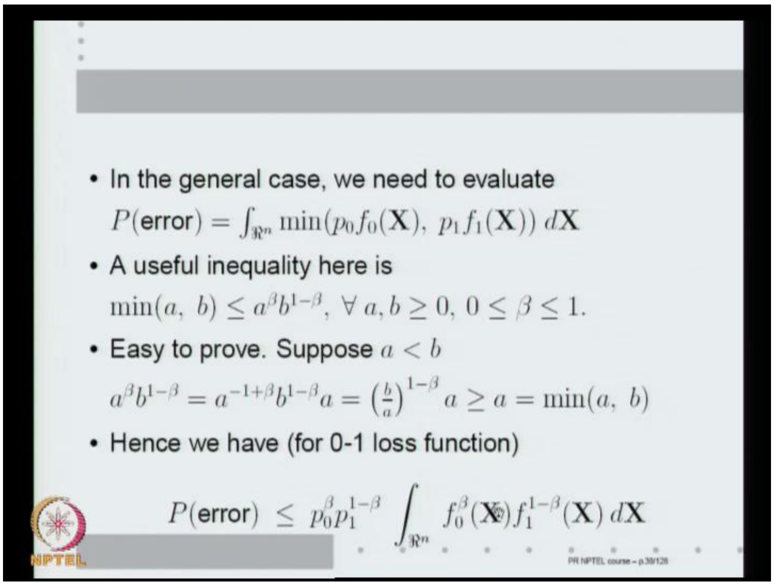
The error density is this, where phi is the distribution of the standard normal density. So, the first integral is from minus infinity to mu 0 minus mu 1 by sigma. So, that is phi of mu 0 minus mu 1 by sigma, second integral is from mu 1 minus mu 0 by sigma to infinity. So, this is 1 minus phi f mu 1 minus mu 0 by sigma, Because, the distributional function standard normal is symmetric phi of 1 minus phi x is equal to phi of minus x right, what is in this big brackets here is same as this. So, the 0.5 goes away, so the error becomes this. So, for equal variants both class conditional densities being normal, this is the Bayes error.

So, essentially a Bayes error depends on mu 0 minus mu 1 by sigma, as you would expect, when sigma is same, I am just putting the point midway. So, how much error I make depends on how much the means are separated, if means are separated by a large amount then, I will make less error, if means are separated by a small amount, I will make more error and small and large amount is relative to the variants of the distribution right.

So, that is what this expression denotes, the quantity mu 0 minus mu 1 by sigma, where mu 0 minus absolute value of mu 0 minus mu 1 by sigma is called the discriminability this, when this quantity is large right, note that, we are assuming mu 0 less than mu 1. So, when discriminability is large this should be phi of some large negative quantity, so close to 0.

So, if mu 0 and mu 1 are separated by a large amount relative to the variance, then I make very small error, if they are separated by a small amount, then I will make a large error the Bayes error. So, that is why, this quantity is called the discriminability, now this is for a very special one dimensional case, you know normal densities with equal variance.

(Refer Slide Time: 14:49)

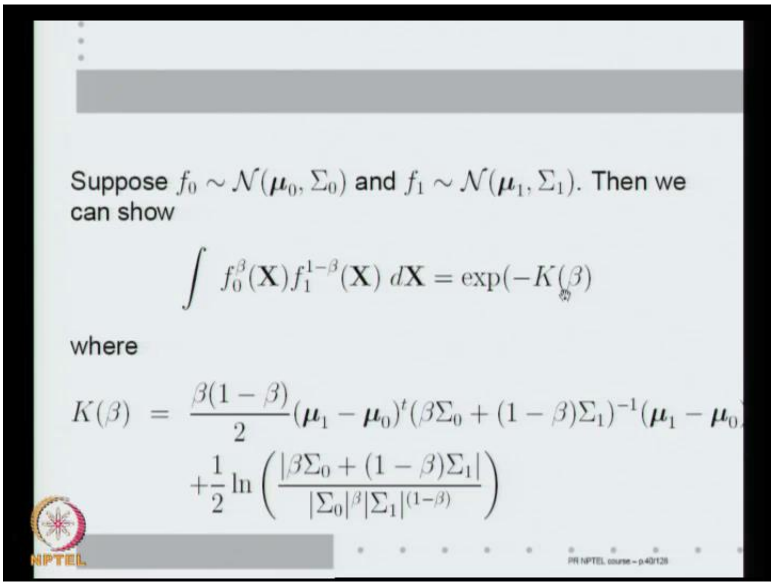


What about the general case, in general, we have to evaluate this integral as the already said mean inside the integrand is often difficult to evaluate. But, 1 can use, 1 very standard and useful inequality, for any 2 real numbers a and b mean of a comma b can always be bounded above by a to the power beta into b to the power 1 minus beta, for any beta between 0 and 1. I here, we are assuming that both a and b are positive numbers, this bond is not difficult to prove.

So, let us prove this, let us suppose a is less than b, now what is a power beta into B power 1 minus beta, I can write it as a power minus 1 plus beta, that is divided by a and then multiply by a. So, this becomes b by a, whole to the power 1 minus beta into a, now I am assuming b greater than a. So, b by a is greater than 1, so b by a to the power 1 minus beta some quantity greater than 1. So, when you multiply that with a, I get some quantity greater than 1 and a is ofcourse mean of a comma b.

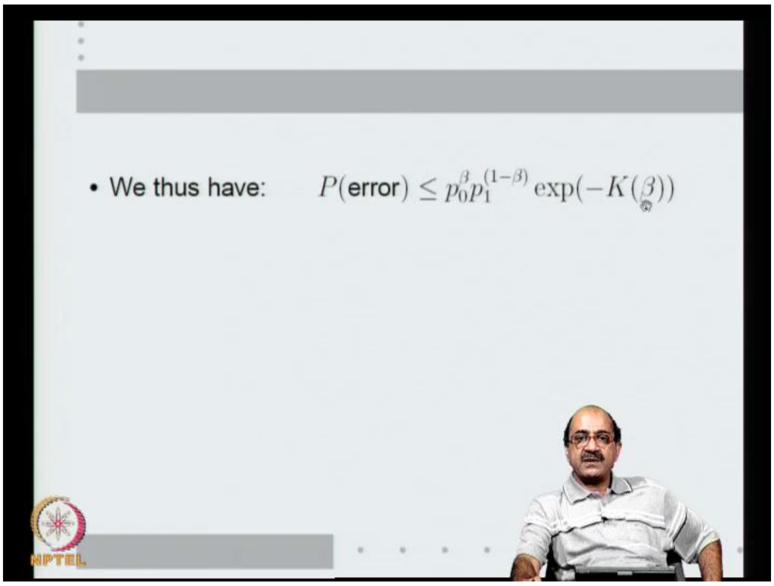
So, this shows that mean of a comma b is bounded above by a power beta into b power 1 minus beta. So, I can reduce this integral to, I can write this mean bound this mean by this to the power beta and this to power 1 minus beta, that becomes p 0 to the power beta p 1 to the power 1 minus beta integral of f 0 to the power beta and f 1 to the power 1 minus beta ah. This is slightly easier integral to evaluate than, this because, I do not have mean inside, it is the normal density raised to some fractional power.

(Refer Slide Time: 16:21)



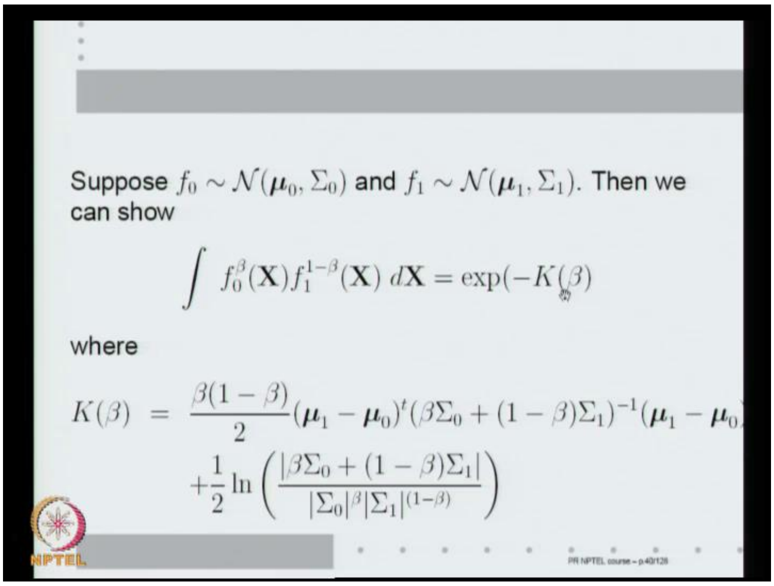
Even then it is a difficult integral to evaluate, but suppose, if I assume that f 0 and f 1 are some multidimensional Gaussian densities f 0 has mean mu 0 and covariance matrix sigma 0. F 1 has mean mu 1 and covariance matrix sigma 1, then one can show that this integral can be bounded above by e power minus K beta, where that K beta is some involved expression like this essentially some kind of a quadratic form, involving mu us and sigma 0. But, anyway this can be shown it is say just say algebraically difficult, but otherwise, this derivation is straight forward. So, 1 can show this.

(Refer Slide Time: 17:03)



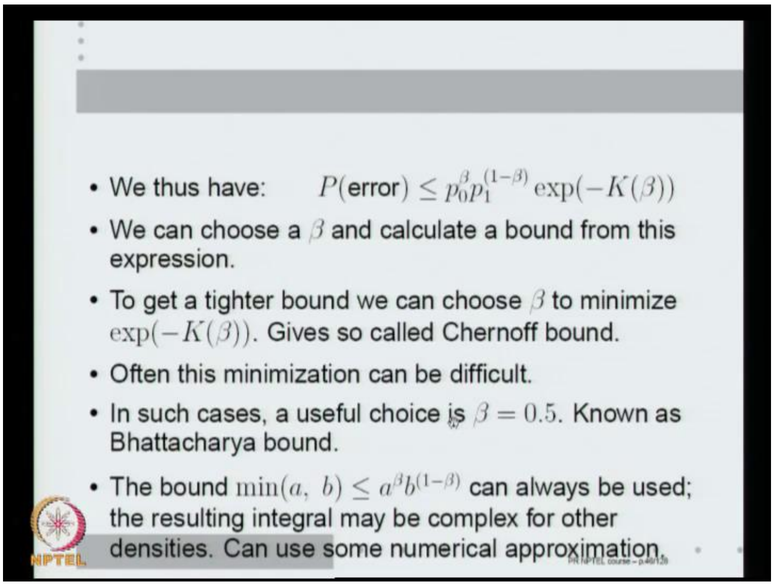
So, what does this mean, that P error is less than or equal to p 0 power of beta p 1 to the power 1 minus beta exponential minus K beta.

(Refer Slide Time: 17:09)



Where, the K beta term is given by this, which I can calculate, if I know mu 1 mu 0, sigma 0 sigma 1.

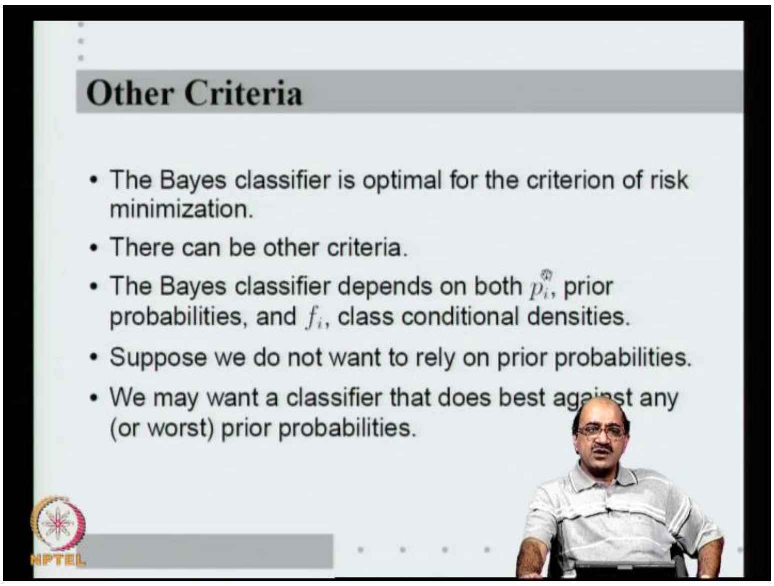
(Refer Slide Time: 17:17)



Now, how can I use this bound, this bound is true for all beta between 0 and 1. So, for example, I can ask, which beta will give me the tightest bond. So, we can choose a beta and calculate a bound from this expression. A bond calculated like that is often called Chernoff bound, we can get a tighter bound by choosing beta, that minimizes this expression.

Such a bound is called a Chernoff bound, if you do not want to do all that work, in practice a beta that often works is beta is equal to 0.5 and the bound and the Bayes error obtained through this expression, where I choose beta to be 0.5 is known as the Bhattacharya bound. There is another bound on the Bayes error. Of course, in general this bound can always be used though, for general class conditional densities, I would not have this exponential minus K beta, I will have that actual integral and we need to know how to evaluate it. But, even for normal class conditional density as you can see, we can only bound the Bayes error, it is not easy, to actually compute the Bayes error. But, this this is a this is one way in, which I can estimate the Bayes error, either using Chernoff bounds or the Bhattacharya bound all right.

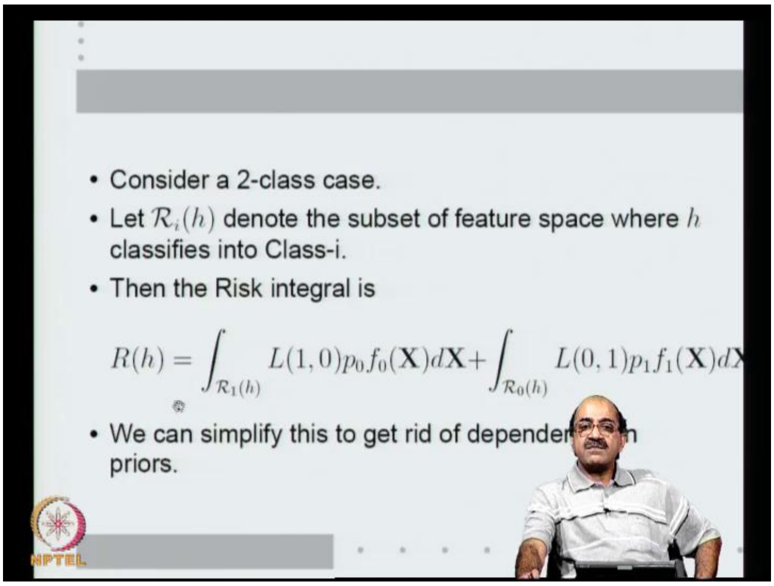
(Refer Slide Time: 18:40)



Let us move on from Bayes classifier, Bayes classifier is optimal for the criteria of risk minimization. But, risk minimization is only 1 criteria right, we can have many other criteria, what does risk there is one thing about Bayes classifiers, it depends both on the prior probabilities and class conditional densities. Now very often, I may not know prior probabilities, out there in the field, which patterns will come I may not know though, I may be able to estimate the class conditional densities, were sometimes, we may want a classifier that does well against any worst kind of prior probabilities. So, without knowing what is the prior probabilities is I do not want my Bayes classifiers to depend on priors.

Because, one day I might have to work with predominantly 1 class patterns, another day I may have to do work with predominantly another class patterns. So, I can ask minimizing risk is not what I want, I want a classifier that has the best risk against the worst possible prior probabilities. Now, this ofcourse, would not be the Bayes classifiers, because I do not know the priors, let us just intuitively see, what this will mean.

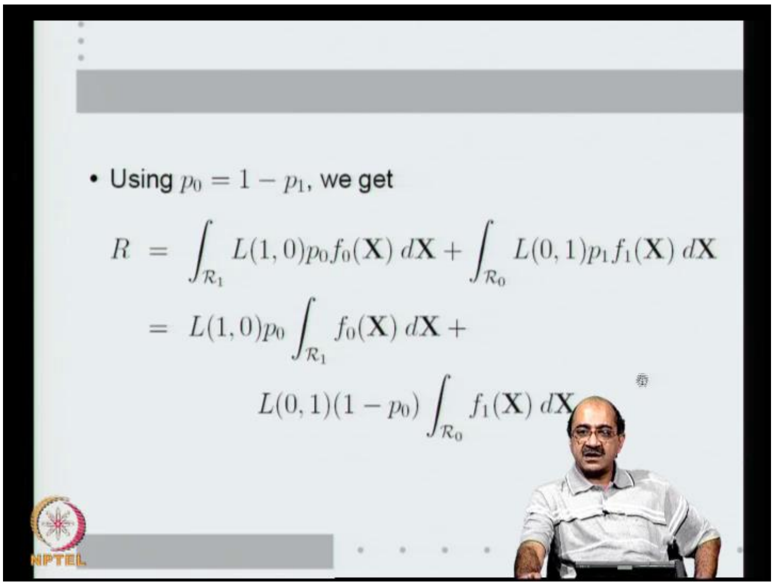
(Refer Slide Time: 19:56)



So, let us say, we will take a 2 class case as earlier let us say R i h denotes the subset of feature space, where the classifier h will put things in class I, that is R 0 is the region of class 0, R 1 is the region of class 1 and by region, I mean not the actual region of class 1 or class 0 feature vectors. But, that subset of the feature space, where the classifier h will put the patterns in that particular class.

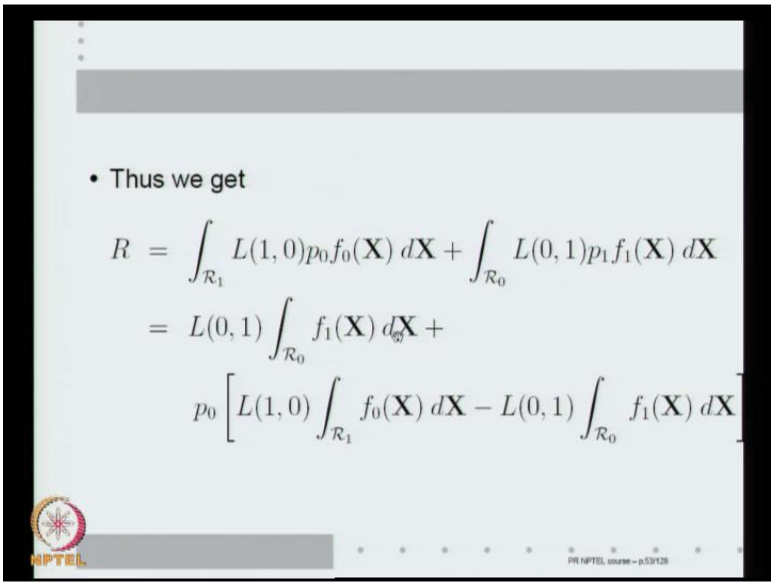
Then, if from what we derived earlier the risk integral is the probability that a class 0 pattern comes into a region where, h will put in class 1 and the probability that a class 1 pattern will come into a region that h will put in class 0, this is the same integral of that we got earlier. Now, what we want to do is we want to manipulate this expression. So, that it becomes independent of the prior probabilities.

(Refer Slide Time: 20:53)



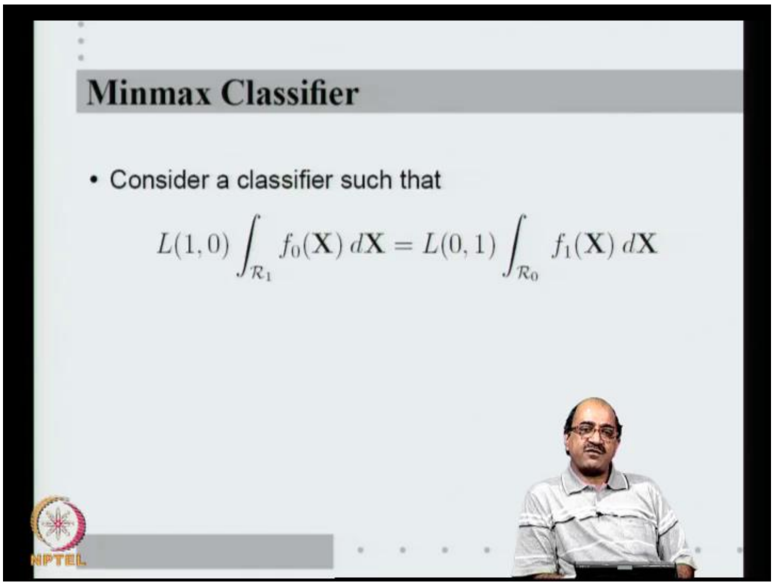
We can do that because, we know that p 0 is 1 minus p 1 or p 1 is 1 minus p 0, so we can eliminate one of them. So, this is my risk integral over R 1, it is L 1 0, p 0, f 0, over R 0, it is L 0 1, p 1, f 1, now I can eliminate one of p 0 and p 1 let us say I will substitute p 1 is equal 1 minus p 0, so that is my risk integral. Now, this integral has one term, which is constant L 0 1 into integral of f 1 over R 0 and another term that depends on p 0, p 0 into this the this first integral minus the second integral.

(Refer Slide Time: 21:34)



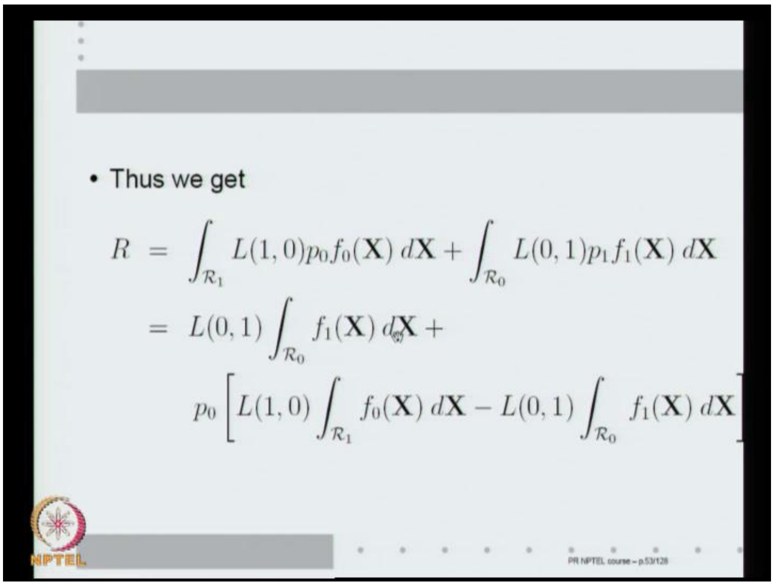
So, I can write the risk as 1 constant term plus p 0 times difference of 2 integrals. Now, for a classifier, we chooses the regions R 0 and R 1 in such a way that, this expression is these big brackets goes to 0, that classifier is risk is independent of priors right. The way the risk is written, if there is a classifier, which chooses class 1 and class 0 decision regions R 1 and R 0 in such a way that this expression becomes 0 right. The the the the second term in this expression has becomes 0, for that classifier the risk will be independent of the priors right.

(Refer Slide Time: 22:23)



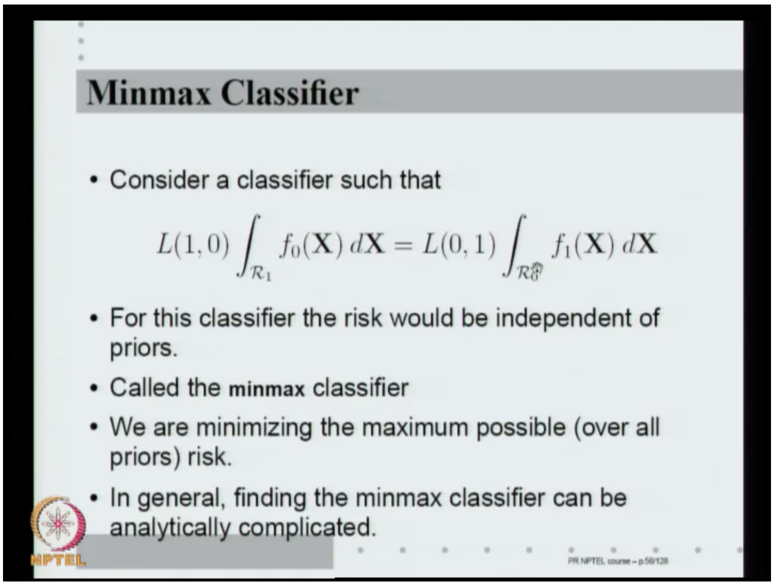
So, consider a classifier, which chooses any classifier, what do you mean by design of a classifier, once we design a classifier, I have designed a function from the feature space to the set 0 1, because we are considering a 2 class problems. So, which means each classifier, simply assigns some subset of the feature space where, it will if a feature vector falls in that subset, I will call class 0, similarly, the remaining substrate, I will call it class 1.So, if design of every classifier is simply choosing a region R 1 where, I will call class 1 and choosing a region R 0 where, I will call class 0. So, a classifier designed in such a way that the regions R 0 and R 1 are so chosen. So, that this equation is satisfied.

(Refer Slide Time: 23:07)



Where did I get this equation from this is nothing but, the term here right. I wanted to make this term 0.

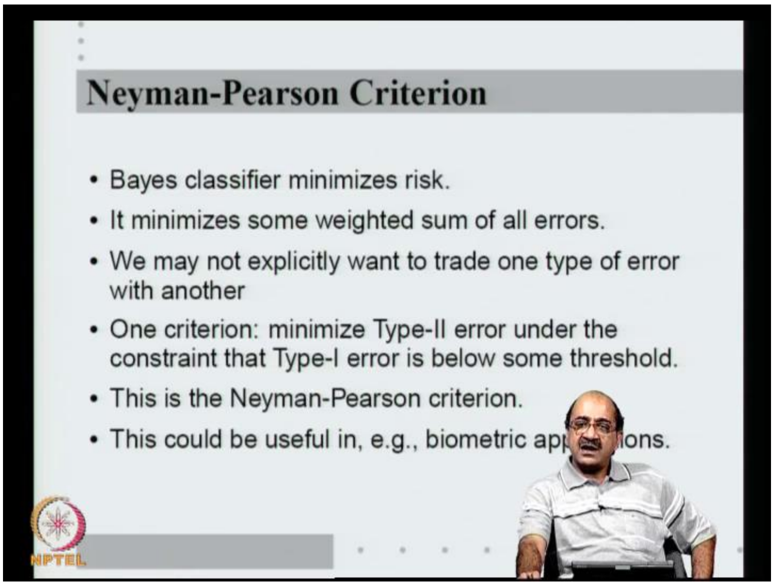
(Refer Slide Time: 23:13)



So, that term will become 0, if this equation is satisfied for such a classifier the risk would be independent of priors, this classifier is known as minmax classifier, because it can be shown it is enough to see that, we are minimizing the maximum possible risk or all possible priors, because we are canceling out the prior dependence and risk. So, we are budgeting for the maximum possible risk where, maximum is over all possible priors for the same class conditional densities.

Of course, finding analytically a classifier that satisfies such expression is not easy, in general finding minmax classifiers is an analytically complicated issue. But, the purpose of mentioning minmax classifier here is just to say that risk minimization is not necessarily the only criterion, we can have when, we are looking for classifiers, here is another example of a classifier. Which is different from Bayes, but it has it is own optimality criterion the minmax classifier, which minimizes the maximum possible risk, where maximum is over all possible priors.

(Refer Slide Time: 24:21)



Let us look at 1 more criterion, this is also a very famous criterion called Neyman pearson criterion. To understand this criterion, let us go back to Bayes classifiers again Bayes classifier minimizes risk, what is risk, risk is expectation of loss. So, each loss is what, I pay for an error and because, in expectation it is some weighted sum of the errors right, weighted sum of probability of errors weighted sum of losses. So, if i if i classify a class 0 pattern as class 1 pattern, there is some cause associated with it class 1 pattern as, class 0 pattern, there is some other cause associated with it I find weighted sum of all such cause and asking, which classifier minimizes it and that happens to be the Bayes classifiers.

But, the 2 kinds of errors may not always be tradable, this assumes that, if cost of 1 kind of error is 3 times cost of another kind of error, we are saying it is better to make 2 errors of 1 kind than, 1 error of this kind right. That is the trade of we are doing in minimizing risk. But, there would be situations where, we may not want to trade, 1 type of error with another type of error.

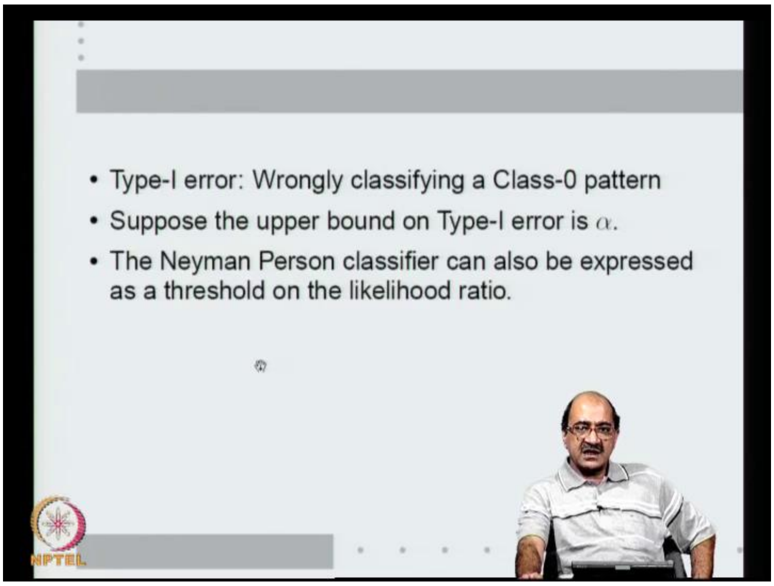
So for example, instead of trading errors, we can say for a fixed constraint and type 1 error minimized type 2 error, recall that type 1 error is wrongly classifying a class 0 pattern, why would this may be possible, let us suppose you are in a biometric application, suppose you are authenticating identity of somebody. So, there are 2 kinds of errors, when somebody is an imposter allowing him access is 1 kind of error, a authorized person not being allowed access is another kind of error.

Now, these 2 kinds of errors are qualitatively different and I may have do not want to find optimal by saying, so many time this error plus, so many time that error should be minimized. On the other hand I may say that I do not want more than 1 percent more than 0.1 percent of the time, an unauthorized person gaining access, while maintaining that can you give me, the best possible error rate for the other kind right.

So, I will put a particular threshold for type one error, that is I do not want more than 1 in 1000 more than once in 1000 times an unauthorized person should gain access. And among all classifiers, that meet this specification, I want a classifier, which minimizes the error of throwing away an authorized person, because, throwing away an authorized person is only an irritation.

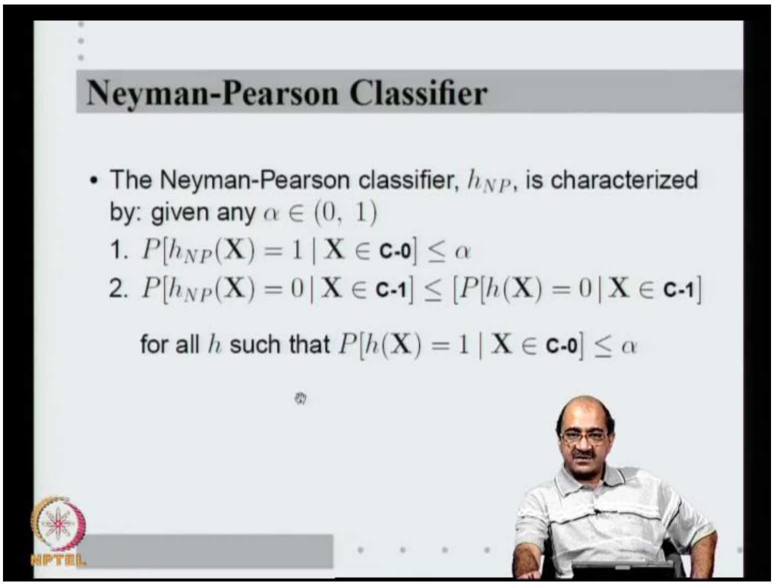
So, there are applications where, you do not want to trade errors of kind with another, I may want to put a threshold on the error of 1 kind, that is I want error of one kind should not exceed a probability of error of 1 kind should not exceed something and given that I want to thenthen minimize the error kind of error. So, it is as I said it is generally useful in biometric applications where, as I said I may want to put an absolute bound on how often, I may allow an unauthorized person to gain access and while, I am satisfying this specification. I want to minimize the number of times, I will throw away an authorized person.

(Refer Slide Time: 27:40)



So, type 1 error is let us say wrongly classifying class 0 pattern and let us say upper bound andprobability of type 1 error is alpha. So, the Neyman pearson classifier is a classifier, that achieves a bound of alpha and type 1 error, that is the probability of type 1 error by Neyman pearson classifier is less than or equal to alpha. And in addition, it minimizes the other kind of error. A matter of fact neyman pearson classifier can also be expressed as threshold on likelihood ratio as we have seen in Bayes case, it is simply a ratio on the posterior Probabilities are class conditional densities. We have put a threshold on this ratio Neyman pearson classifier can also be expressed and we will see how.

(Refer Slide Time: 28:27)

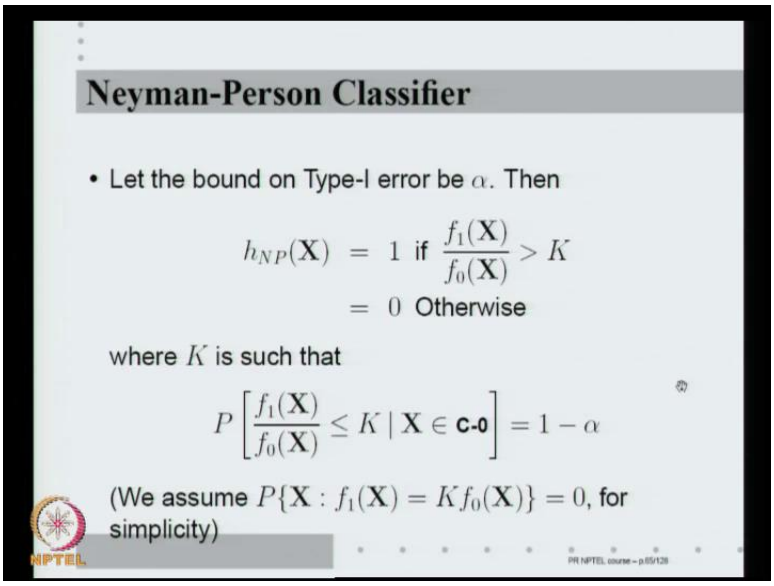


So, let us first define the Neyman pearson classifier, the Neyman pearson classifiers let us call it h N P, we were calling Bayes classifier as h b, so we will call a Nayman pearson classifier h N P. So, given any alpha in 0 1, this is the upper bound on the type 1 error what does h N P have to satisfy. Firstly, probability h N P X is 1 given X belongs to c 0, that is wrongly classifying a class 0 pattern is bounded above by alpha right.

This is 1 thing that h N P has to satisfy, then what is it have to satisfy, wrongly classifying a class 1 pattern, that is the other kind of error h N probability h N P is 0 given X belongs to c 1. That should be, less than the probability of wrongly classifying a class 1 pattern by any other classifier h. But, this is not for all h, but only those h, which also meet the bound on the type1 error because, N P minimizes the second kind of error while satisfying the bound on the first kind of error.

So, among all classifiers h, that satisfy the bound on type 1 error. So, if h is such that probability h X is equal to 1 given X belongs to c 0 is less than or equal to alpha, that means, this classifier h also satisfies the type 1 error bound. Then the probability of the type 2 error by h N P is less than or equal to probability of type 2 error by h right, I hope it is clear. So, the Neyman pearson classifier is characterized by firstly, it is type 1 error is bounded by alpha and it is type 2 error, that is wrongly classifying class 1 pattern is less than the type 2 error of any other classifier h, if h also satisfies the bound on the type 1 error.

(Refer Slide Time: 30:18)



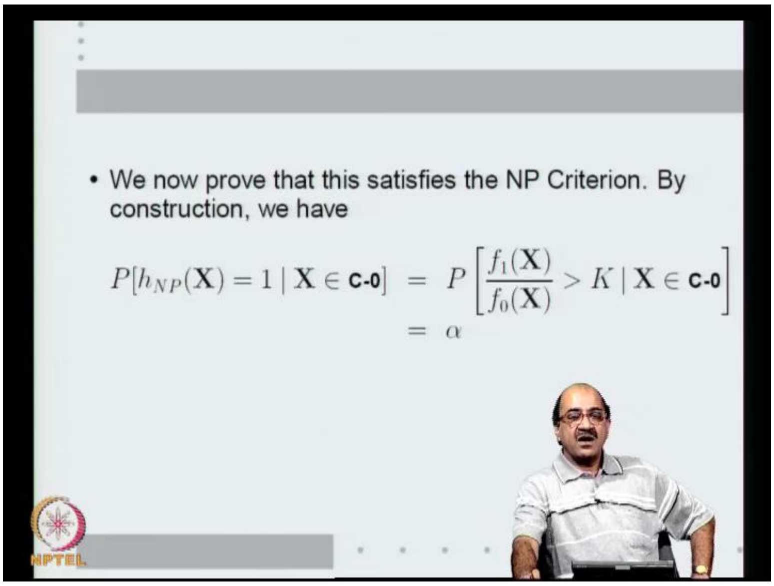
Now this is what the Neyman pearson classifier is let us suppose the bound is alpha as, we are saying then the h N P classifier is the Neyman pearson classifier is defined by it will assign class 1 to X. If f 1 X by f 0 X is greater than K, otherwise assign class 0 where, f 1 is the class conditional density for 1 and f 0 is class conditional density for class 0 where, the K itself is obtained by under the distribution of class 0, that is the another distribution f 0. The probability that, this ratio is less or equal to K is bound above by is equal to 1 minus alpha.

So, I choose a K see Bayes is some probability involving the random variable X under the distribution f 0, you think of this f 1 and f 0, simply as some functions. So, f 1 by f 0 is some other function g of X. So, this is probability g of X less than or equal to K. Under the condition that X is distributed as f 0, X belongs to f 0 means X is distributed as f 0 right. What does this ensure, this ensures that my type 0 error is equal to alpha, when will I wrongly classify a 0 pattern. I will call class 1, if this ratio is greater than K, when X belongs to c 0. This ratio will be greater than K the probability of the ratio is greater than K is equal to alpha because, K is chosen to satisfy this equation, the ratio is less than or equal to K is 1 minus alpha.

So, the probability ratio is greater than K is equal to alpha right. So, by construction the Neyman pearson classifier satisfies the bound on type 1 error ok. We will we will see the see it once more, just for completeness is sake the way, we stated this, we are assuming that f 1 and f 0 are true density functions. That is probability X belonging to any lower dimensional subspace, for example, a sub space the characters be f 1 X equals to K f 0 X this is 0. So, because, we are assuming that the ratio is either greater than K or less than equal to K.

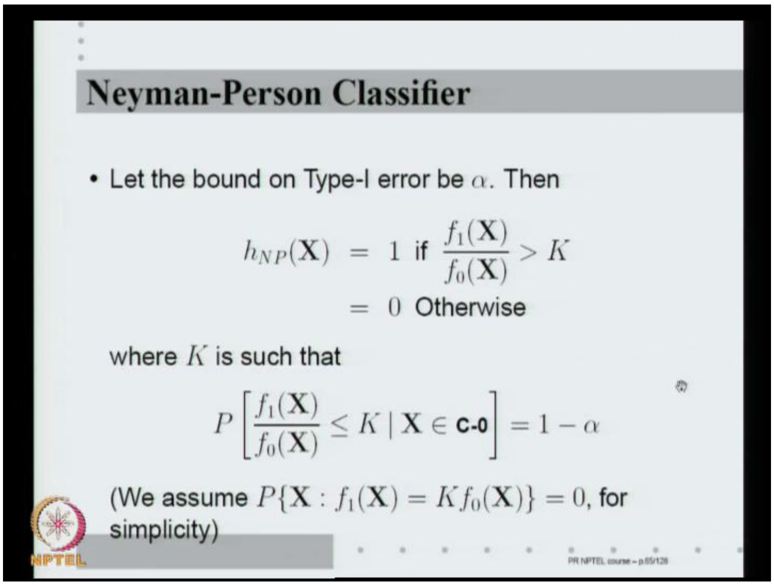
So we we we will we will not allow any any kind of derived delta part in the f f 1 and f 0, this is only a technical condition, those of you do not understand this, do not worry about it simply assume that f 1 and f 0 are nice smooth density functions then this will be all right. So, now let us prove that the classifiers that, we put here right. This is this is the specification this is the this is how Neyman pearson classifier will classify a new pattern X, this classifier is actually the Neyman pearson classifier that is it satisfies, the 2 condition that, we shut down for then a Neyman pearson classifier.

(Refer Slide Time: 33:17)



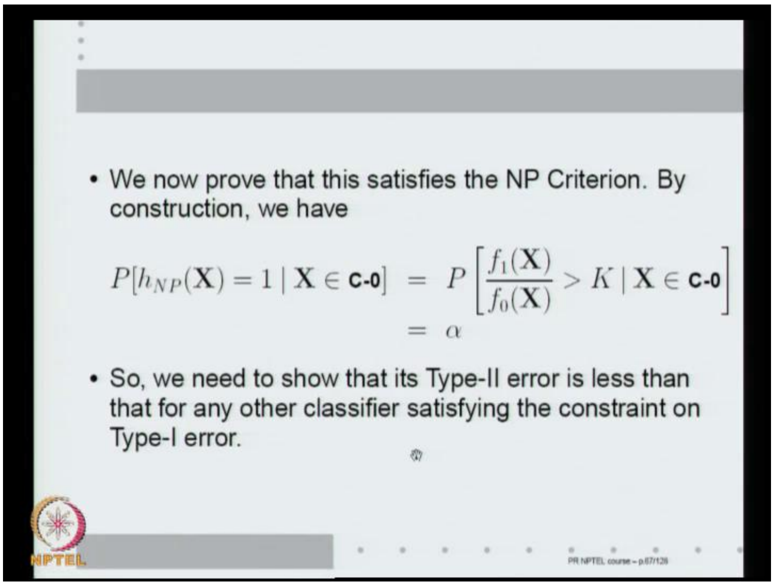
What are the 2 conditions, first is type 1 error should be less than or equal to alpha, for the type 1 error h N P X equal to 1 given X belongs to c 0.

(Refer Slide Time: 33:30)



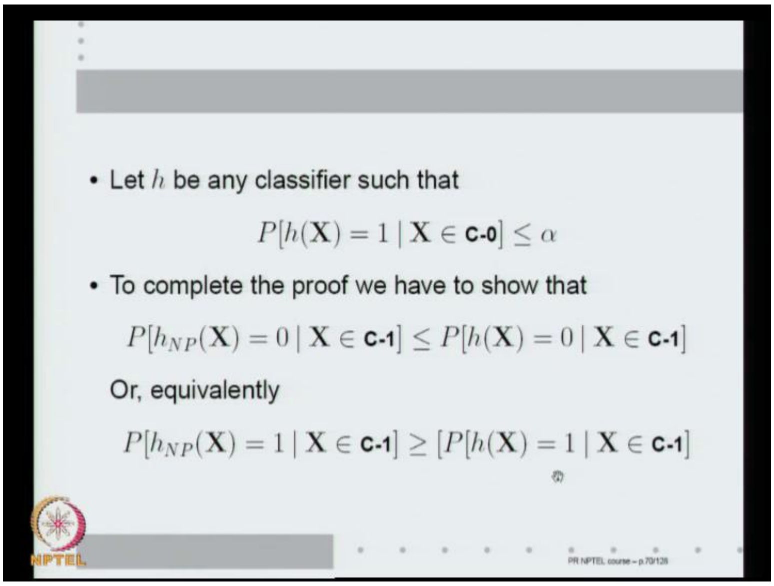
When will h N P X be 1 by by definition h N P X is 1, if f 1 by, f 0 is greater than 1 K. So, probability h N P X equal to 1 is probability f 1 by f 1 X by f 0 X is greater than K and when X belongs to c 0, this where the K in the Neyman pearson classifier is obtained by this equation. So, this equation ensures that probablity f 1 by f 0 greater than K is equal to alpha right.

(Refer Slide Time: 33:55)



So, the probability of type 0, type 1 error by Neyman pearson classifier is alpha. So, it satisfies the first crietrion. So, now we have to show there is type 2 error is less than that for any other classifier, which also satisfies the constraint on the type 1 error.

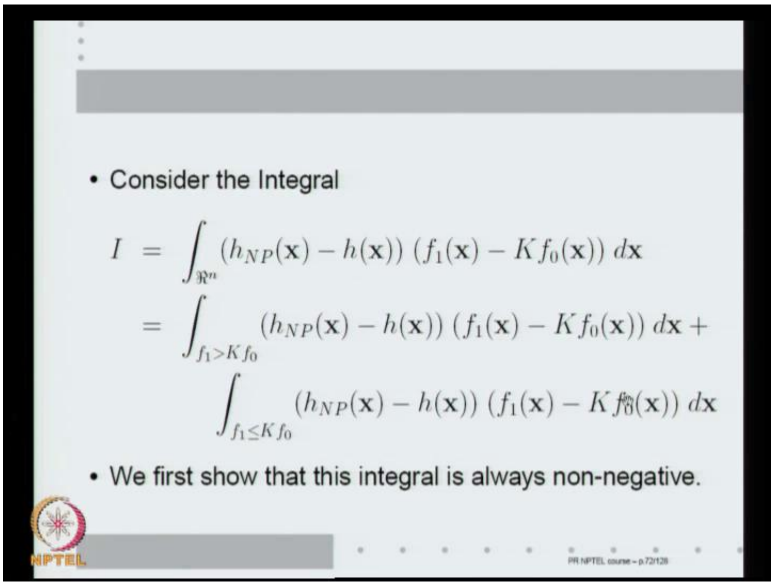
(Refer Slide Time: 34:16)



So let us prove this, so let h be any other classifier, which also satisfies the constraint on the type 1 error that is probability h X is equal to 1, when X belongs to c 0 is less than or equal to alpha. Then to complete the prove you have to show that, the type 2 error of h N P, that is probability h N P X equal to 0, when X belongs to c 1 is less than or equail to probability of h x equal to 0, for when x belongs to c 1.

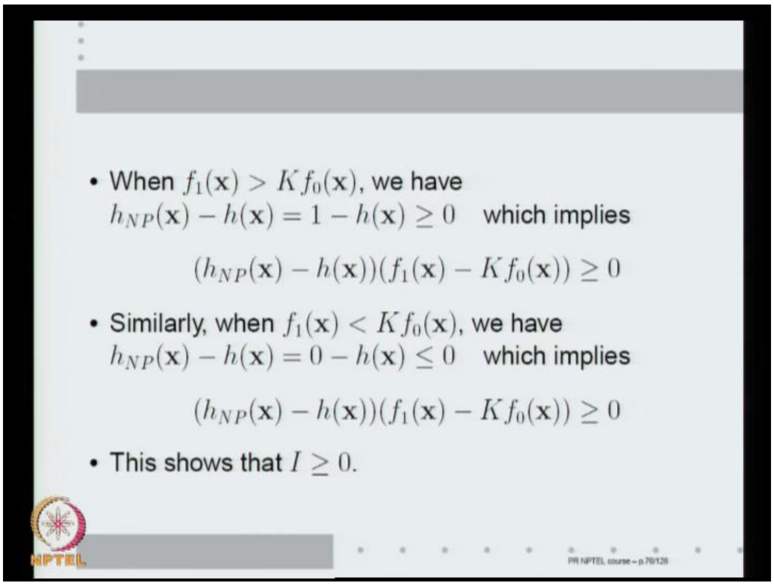
So, this is what we will next show, we actually wont show it in this form the way, we will show it is, we will show the compliment of this event. So, we will show that the probability h N P X is equal to 1, when X belongs to c 1 is greater than or equal to probability of h X is equal to 1, where X belongs to 1. So, instead of showing probability h N P X is equal to 0 is less than probability h X is equal to 0, when X belongs to c 1 instead of showing this is less than this, we are showing that the compliment event h N P X equal to 1 is greater than probability h X is equal to 1.

(Refer Slide Time: 35:18)



To show this let us consider the integral I, which is the integral over the entire feature space of the product of 2 terms, the first term is h N p X minus h X, second term is f 1 K f 0 X, where K is the threshold used in the n P classifier. Recall that h N P and h in this case as binary valued functions h N P takes value 0 or 1, h X also takes value 0 or 1. So, h N P minus h X is some real numbers as, a matter of fact it can be only either minus 1 0 or plus 1, what we are going to show first either this integral will always be positive and then, we show that, that completes the proof of the h N the classifier that, we gave is the Naymen pearson clasifier. Let us first know that I can split this integral into 2 parts, integral over all X. So, that f 1 x is greater than K f 0 X and integral over all X. So, that f 1 X is less than or equal to K f 0 X this spites R N into 2 parts and we are going to show that, for each half the integral is positive. Positive means greater than or equal to 0, but this is non negative.

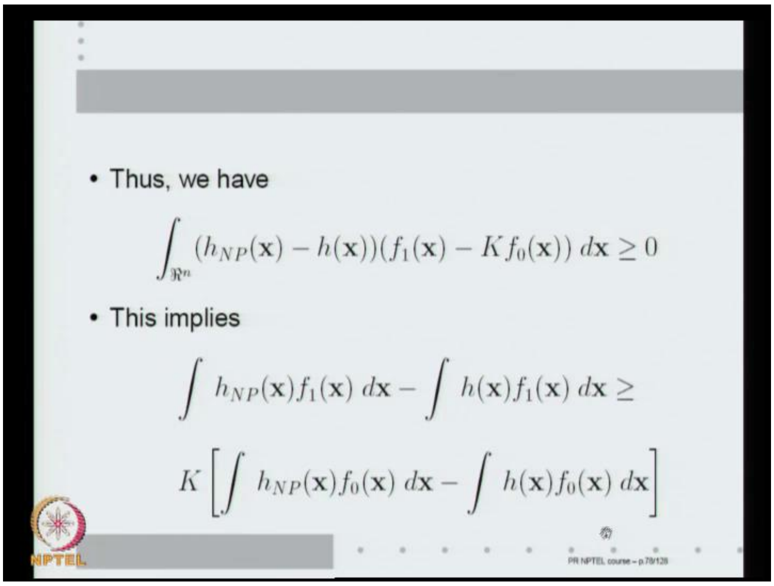
(Refer Slide Time: 36:24)



So, first let us consider all X. So, that f 1 X is greater than K f 0 X, recall that h N P X says h N P X equal to 1, if f 1 by f 0 is greater than K. So, for all X as that f 1 X greater than K f 0 X, we have h N P X is 1, which means h N P X minus h X will be 1 minus h x is always greater than or equal to 0, because h X can be either 1 or 0 right, no matter what classifier h is h X is either 1 or 0.

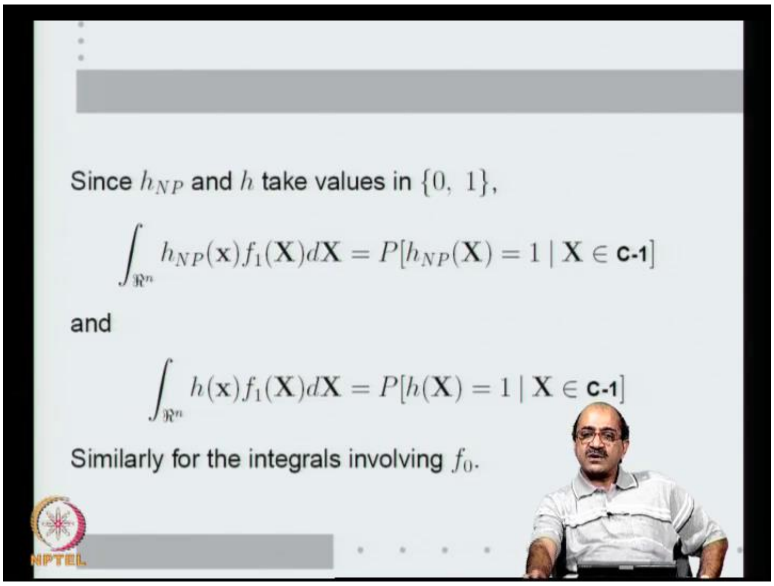
So, for all X either f 1 X minus K f 0 X, f 1 X greater than K f 0 X, h N P X minus h X will always be greater than or equal to 0, which means h N P minus minus h into f 1 minus f 0 is positive, because both terms here are positive. Now, let us look at the other way around let us look at al1 has the f 1 is less than K f 0 X, now h N P X will say 0 right, the those x are put in class 0 by Neyman pearson classifier. So, h N P X is 0. So, h N P X minus h X will be 0 minus h X, which for any classifier h X is less than or equal to 0, because h X can be either 0 or 1a. So, once agin the product h N P minus h into f 1 minus K f 0 is positive, because both factors here are negative, which means the integral, we started with is always positive.

(Refer Slide Time: 37:43)



So what we have shown, so far is that this integral is positive. So, now let us expand this integral by multiplying these 2 terms. So, multiply with with f 1 first. So, I get h N P into f 1 minus h into f 1 right, the those are the first 2 integrals, take the other terms on the other side. So, this is greater than or equal to K times, h N P into f minus h into f 0, because these, because I have taken them on the other side, now h N P term will become positive right. I just multiply this term and put 2 integrals on this side 2, integrals on this side.

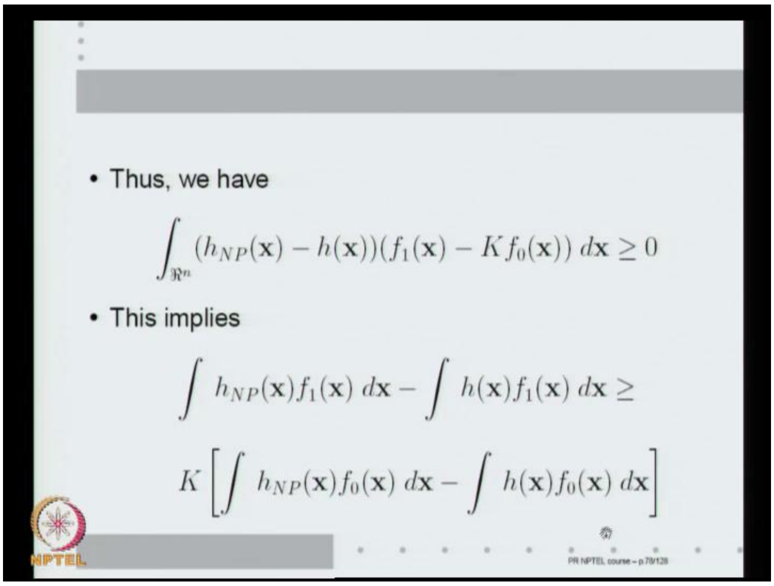
(Refer Slide Time: 38:29)



Since h N P and h are binary valued functions for any given X h N P X is either 0 or 1 similarly, h is a 0 or 1. So, if I take an integral or entire feature space R N h N P into f 1 X d, it is simply integral of f 1 over the region, over the set of all X., so that h N P X is 1 right. Because them integrating mode of 1, this integral is nothing integrating with f 1, here the integral is nothing but, conditioned on X belongs to c 1 probability that h N P is 1 right.

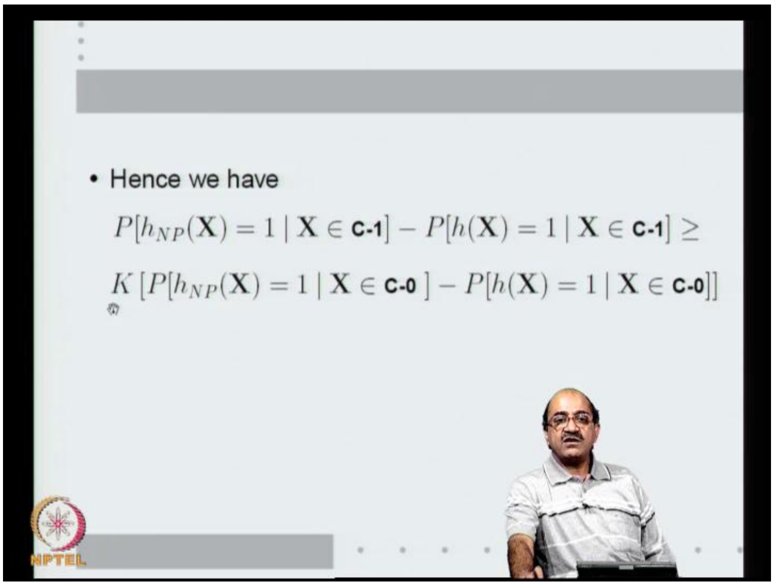
Because h N P is a 0, 1 valued function integral of h N P into f 1, over r N is same as integral f 1, over x at that h N P X is 1, which is same as because, I am integrating f 1, which is same as probability h N P X is equal to 1, conditioned on X belongs to c 1. Similarly, for h because h is also 0, 1 function integral h into f 1 is probability that h X is 1 conditioned X belongs to c 1. similarly, the integral with respect to f 0.

(Refer Slide Time: 39:37)



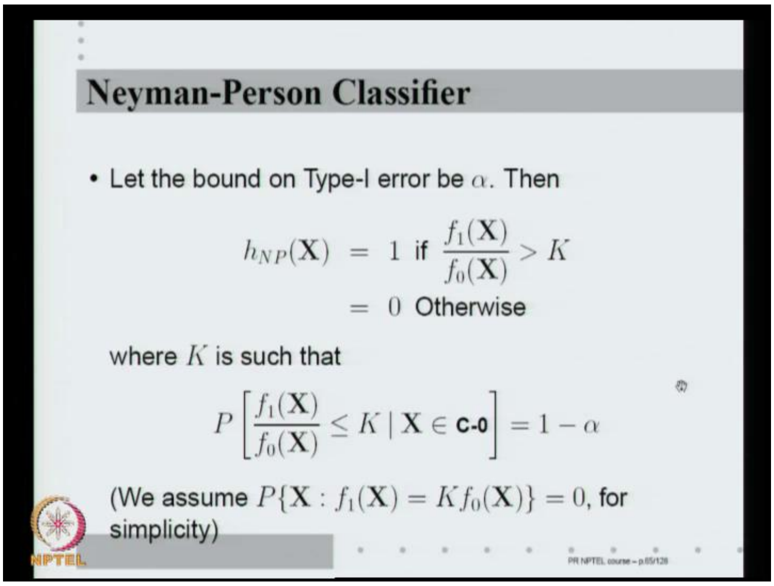
So, which means each of these integrals this can be retained as, probability h N P X is equal to 1 conditioned on X belongs to c 1. This is h X is equal to 1 conditioned on X belongs to c 1. Similarly, this is h N P X is 1 conditioned on X belongs to 0, h X is 1 conditioned on X belongs to 0 right.

(Refer Slide Time: 40:20)



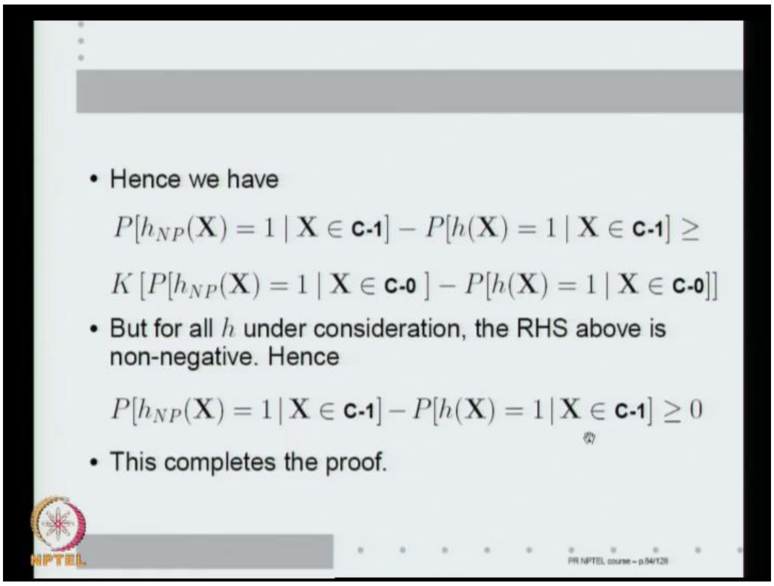
So this inequality can now be written as probability h N P X is equal to 1 given X belongs to c 1, minus probability h X is equal to 1 given X belongs to c 1 is greater than or equal to K times. Probability h N P X is equal to 1 conditioned on x belongs to c 0 probability h X is equals to 1, conditioned on x belongs to c 0, let us also remember that this factor K will always be positive right.

(Refer Slide Time: 40:52)



Because, two ways of looking at it, K is defined by, the reference of Neyman pearson classifier. This is the reference of Neyman pearson classifiers, both f 1 and f 0 are density functions, they are always positive. So, this ratio is always positive, so, if say K is negative is forever satisfied, because so, that is the 1 way looking at it, any case because, this is some positive function of and, we want it less than or equal to K has to have some positive probability K has to be a positive number right.

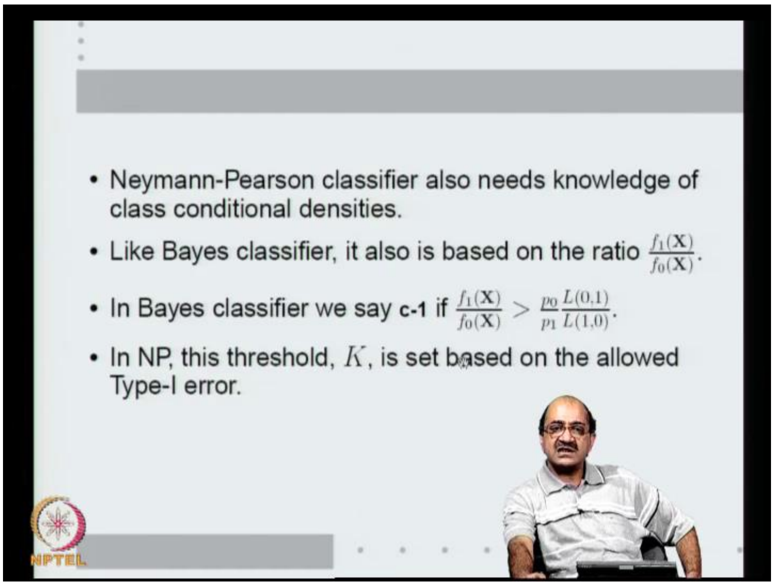
(Refer Slide Time: 41:30)



Given that case a positive number, let us look at what is there in the in the brackets here, this is h N P X equal to 1, conditioned X belongs to c 0 minus h X equal to 1 conditioned as X belongs to c 0. Now this is the type 1 error of h N P X of the Neyman pearson classifier, which by construction is alpha, this is the type 1 error of the classifier h and because, h is something that satisfies the conditions on type 1 error this is less than or equal to alpha. So, this factor is greater than or equal to 0 right.

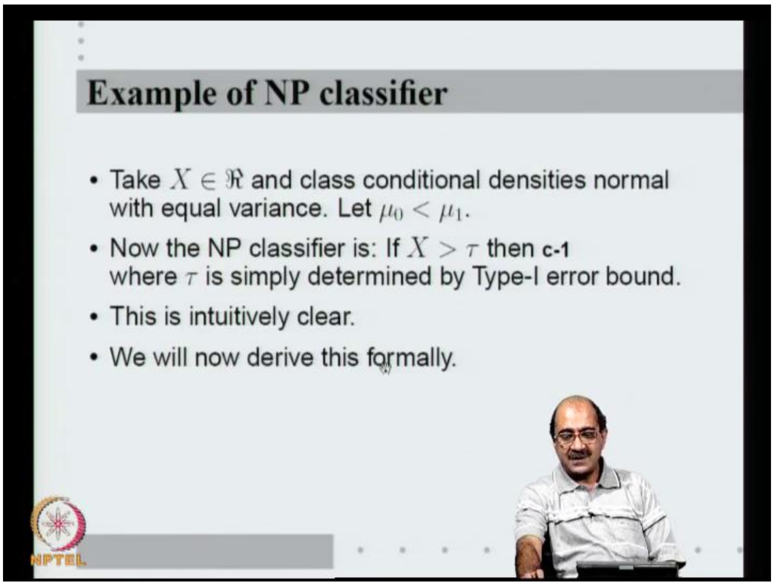
So, for all h under consideration all h, that satisfy the constraints on type 1 error right, the term on the R H S is always non negative, which means this is positive h N P X is equal to 1 conditioned X belongs to c 1 minus h X is equal to 1 conditioned on X belongs to c 1 is greater than or equal to 0. This shows that the Neyman pearson classifier has the smallest type 2 error compared to smaller type 2 error compared to any classifier, that also satisfies the type 1 error bound right. This shows that the classifiers that, we have actually put down as Neyman Pearson classifiers, satisfies the criteria for Neyman Pearson classifier.

(Refer Slide Time: 42:51)



So, Neyman pearson, classifiers also needs a knowledge of class conditional densities because, you have to calculate f 0 by f 1. Like Bayes classifier, it is also based on the ratio f 1 by f 0, in Bayes classifier, we say c 1, if f 1 by f 0 is greater than some threshold, which is which happens to be p 0, l 0 1 by p 1, l 1 0. In n p this is some other threshold K, which is set based on the allowed type 1 error, so both of them essentially threshold the ratio of the 2 class conditional densities.

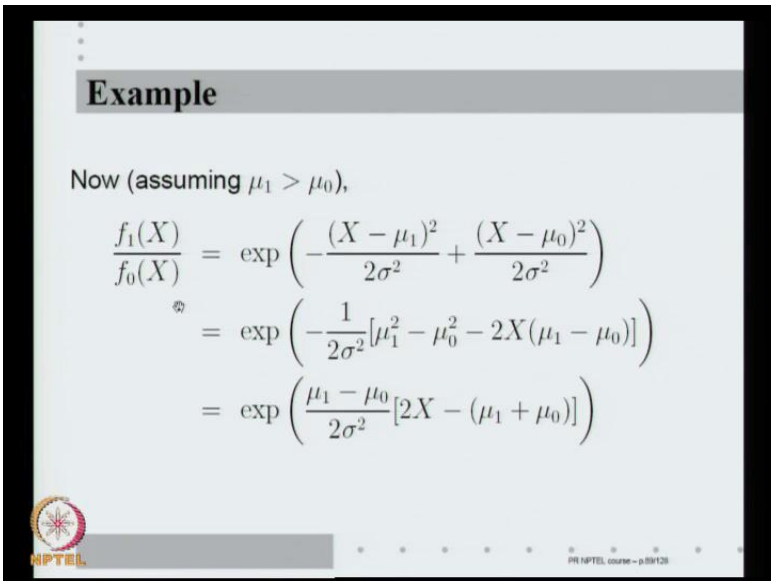
(Refer Slide Time: 43:20)



So, let us quickly look at a simple example, of Neyman pearson classifier, let us take a one dimensional feature space normal class conditional densities with equal varience, and suppose, by now you if you are, if you have been following all the lectures, you know that this is always a simplest case. A 2 class problem, one dimensional feature space, normal class conditional densities equal varience.

For let us assume that mu 0 is less than mu 1, mu 0 is the mean of the class 0 and mu 1 is the mean of class 1. So, what is the N p classifier, if X greater than tau then c 1, where how where how do, I choose tau, tau is simple taken by the type 1 error bound.

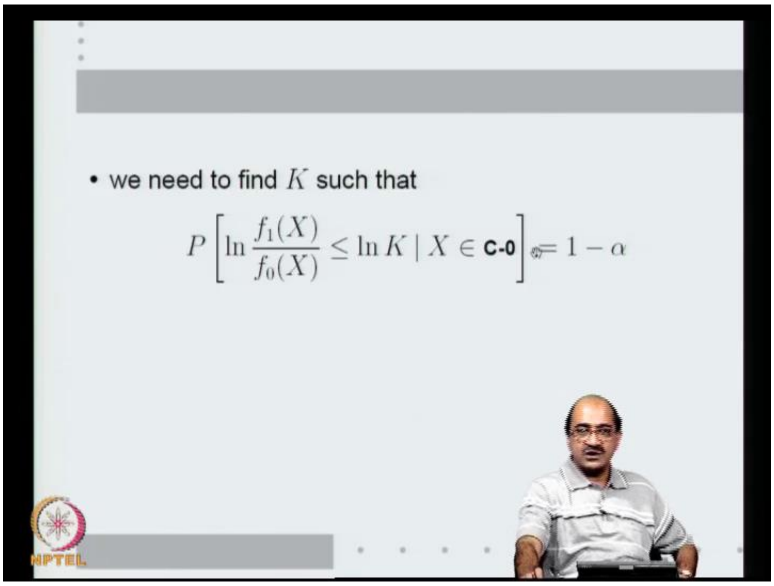
(Refer Slide Time: 44:07)



So let us derive this formally, how do how do I get this term, because both f 1 and f 0 are normal, it is easy to see f 1 by f 0 exponential minus X minus mu 1, whole square by 2 sigma square plus X minus m square by 2 sigma square, the other factors 1 by sigma root 2 pi will cancel right. Now, we can expand this the X square term will cancel.

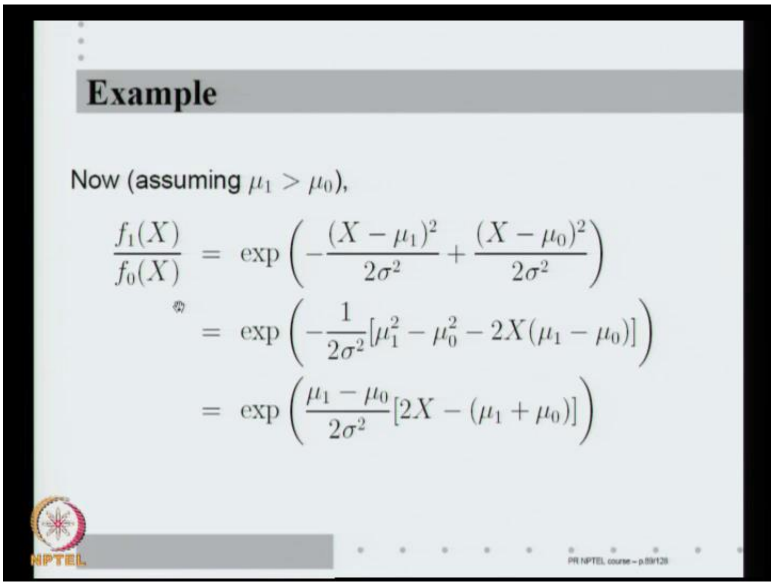
So, what will get minus 1 by 2 sigma square mu 1 square from here, sorry mu 1 square from here minus mu 0 square from here right and 2 X into mu 1 minus mu 0, I can absorb this minus sign. So, I can write it as mu 1 minus mu 0 by 2 sigma square into 2 X minus mu 1 plus mu 0. So, this is the ratio of f 1 by f 0 for the case of normal densities with equal variance.

(Refer Slide Time: 45:00)



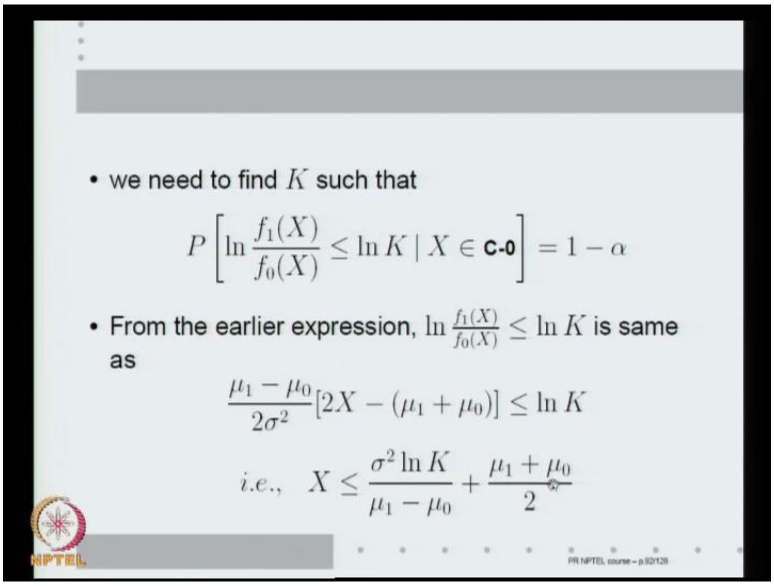
We need to find a K, such that probability f 1 by f 0 less than or equal to K is 1 minus alpha, because log is a monotone function, which is same as probability log of f 1 by f 0 less than or equal to log of conditioned on c 0 that is also good enough.

(Refer Slide Time: 45:20)



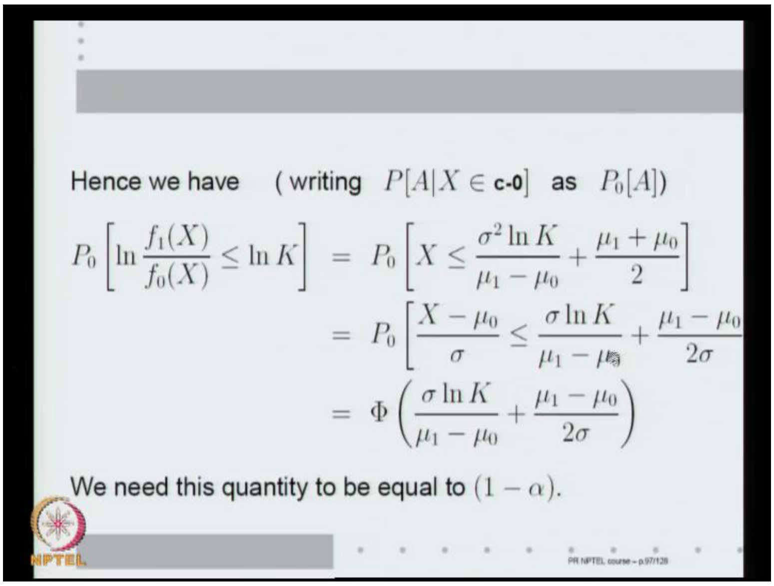
Now this is f 1 by f 0. So, log of f 1 by f 0 will be just simply, what is inside the exponent.

(Refer Slide Time: 45:26)



So, what we get is mu 1 minus mu 0 by 2 sigma square into 2 X minus mu 1 plus mu 0 should be less than or equal to probility K, so I can use this. So this is some expression involving random variable X right. I want the probability of this event to be equal to 1 minus alpha, I have to choose K like that. So we can do it like this, so what will this give me, I can first take 2 sigma square by mu 1 minus mu 0 this side right. Then bring mu 1 plus mu 0 on this side and then, I will get 2 X here divided by 2. So, this inequality is same as X less or equal to sigma square l N K by mu 1 minus mu 0 plus mu 1 plus mu 0 by 2.

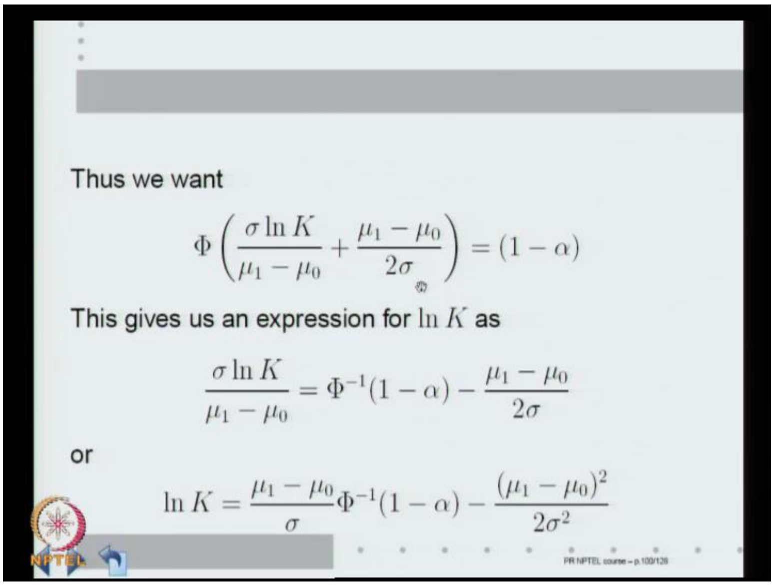
(Refer Slide Time: 46:10)



So, what we want is let us say P, subscript 0 is the probability conditioned on X belongs to class 0. So, P 0 l N f 1 X by l N of f 1 by f 0 less than or equal to l N P is same as p 0, of X less than or equal to that quantity. Now because, this is a P 0 probability means, I am taking this probability under the class 0 class conditional density that is mu 0 mean and sigma variance. So, I can write this as this X less than, this same as, X minus mu 0 by sigma less than or something else.

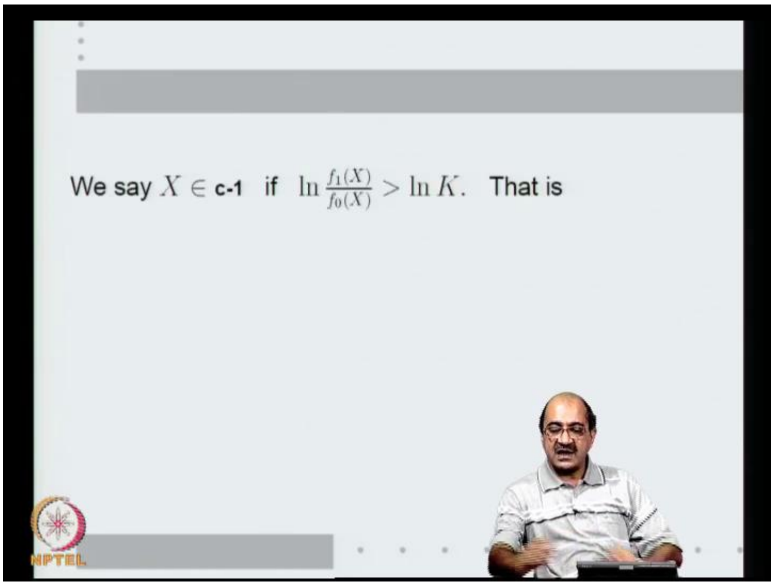
So, just subtract mu 0 and divide by sigma, I get this expression, why did I do that, because I know the distribution of X minus mu 0 by sigma. Because X is under this probability X belongs to c 0, c 0 is normal with mean mu 0, variance sigma X minus mu 0 by sigma is standard normal. So, this probability is given in terms of the standard normal function, phi of this quantity sigma l N K by mu 1 minus mu 0 plus mu 1 minus mu 0 by 2 sigma, where phi is the density of the standard normal. So, ultimately to get K, I have to equate this quantity to 1 minus alpha. Alpha is given in Neyman pearson criteria what I am given is alpha that is the allowed type 1 error bound. So, to get K, I have to equate this to 1 minus alpha.

(Refer Slide Time: 47:33)



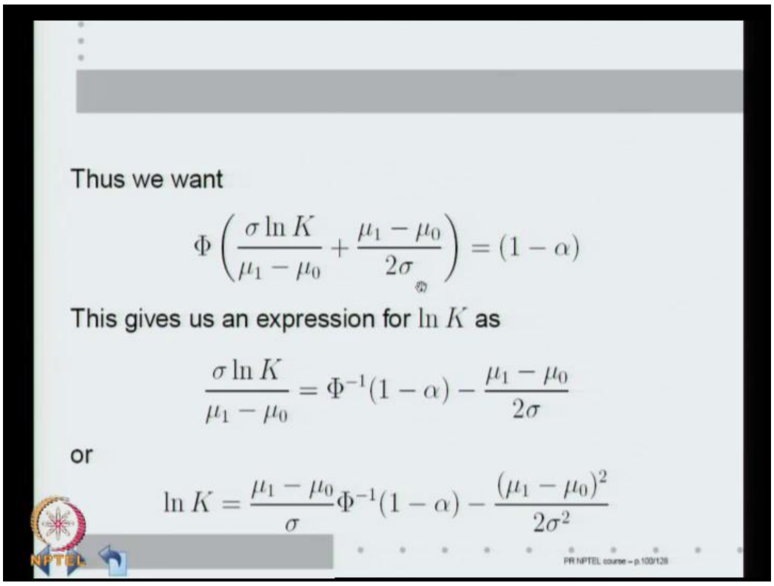
So, let us equate that phi of this is equal to 1 minus alpha, now I can solve this for K or l N K that gives me sigma l N K by mu 1 minus mu 0 is phi inverse of 1 minus alpha minus mu 1 minus mu 0 by 2 sigma. Now multiply by mu 1 minus mu 0 divided by sigma that gives me l N K is this much, from this, I can get K. So this is the threshold, I want for Neyman pearson classifier.

(Refer Slide Time: 48:06)



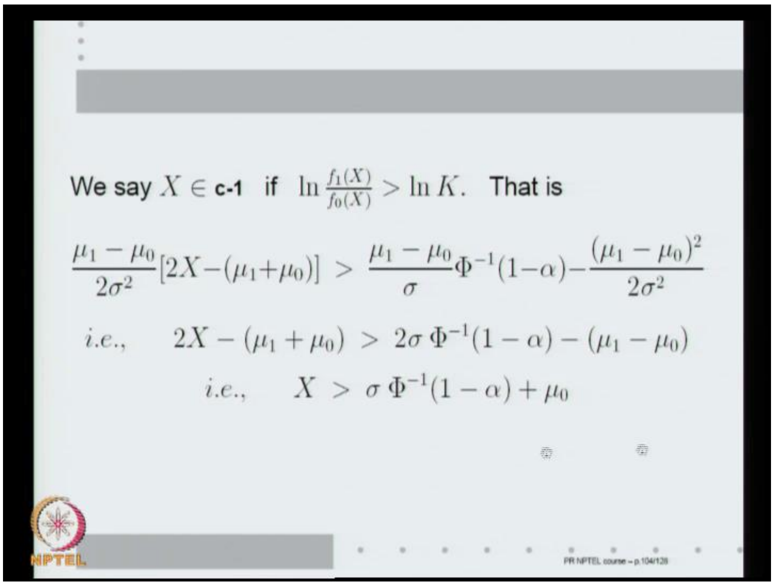
Let us ask what does this classifier mean so, in Neyman pearson classifier, we will put X in class 1, if f 1 by f 0 is greater than K, which is same as l n f 1 by f 0 greater than l n K.

(Refer Slide Time: 48:19)



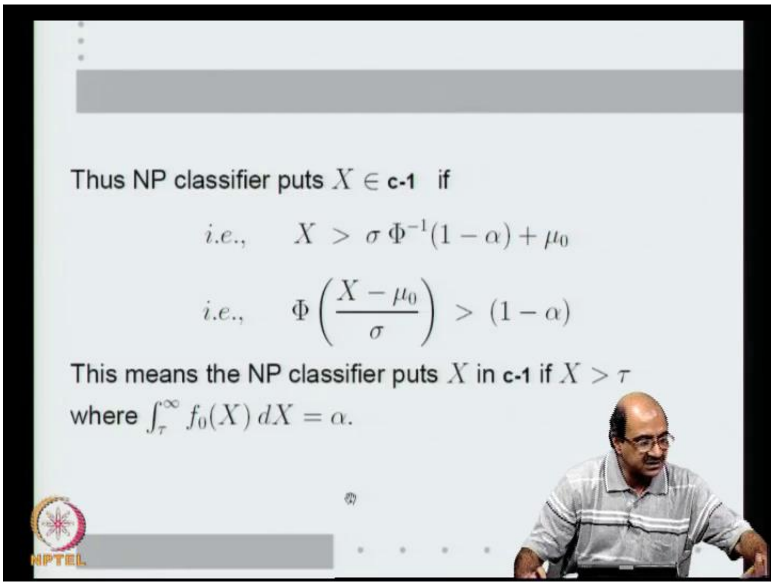
And the l n K is given by this expression that, we have just now derived.

(Refer Slide Time: 48:23)



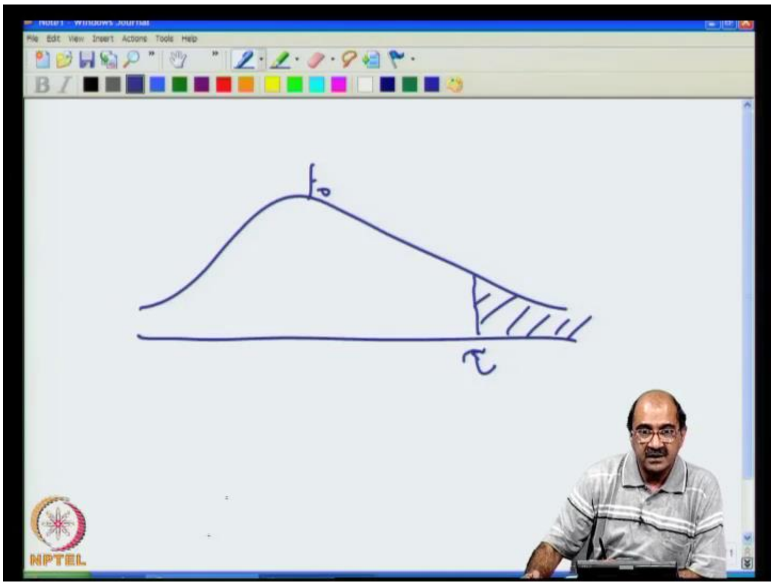
So, I will put X in c 1, if l n f 1 by f 0, that is this expression is greater than l n k, that is this expression. Now I can simplify this expression 2 X minus mu 1 plus mu 0 is multiplied by 2 sigma square and divided by mu 1 minus mu 0, I get this right. 2 sigma phi inverse 1 minus alpha minus mu 1 minus mu 0, remember, we are assuming mu 0 less than mu 1. So, mu 1 minus mu 0 is a positive quantity. So, when I divide by it the the inequality does not change. Now, if being mu 1 minus plus mu 0 on this side and divided by 2, this is same as X greater than this.

(Refer Slide Time: 49:09)



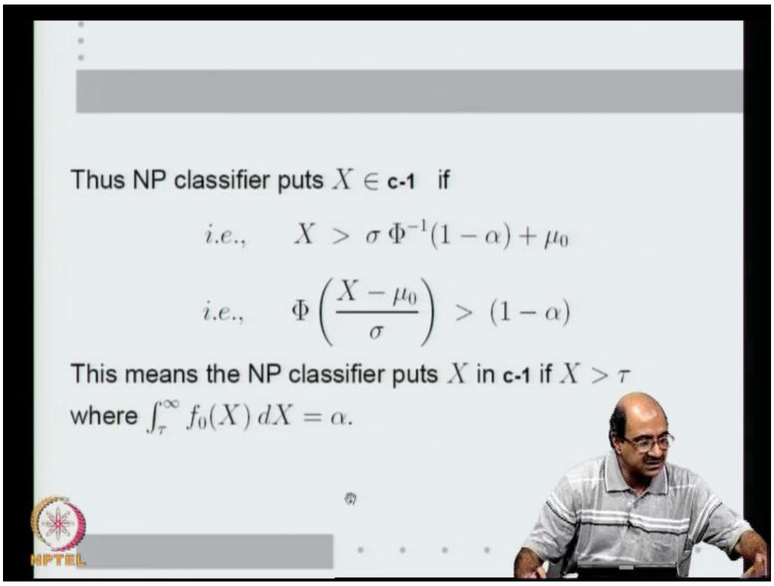
Which, so the Neyman pearson classifier, we will put X in class 1, if X is greater than this. Which is same as X minus mu 0 by sigma and bring phi inverse this side phi of X minus mu 0 by sigma is greater than 1 minus alpha right. What is phi of X minus mu 0 by sigma that is the the distribution of the standard normal density. So, what does this mean, this means this threshold X greater than tau, if I think of this as tau, the integral of the density function of class 0, starting from this tau to infinity will exactly be equal to alpha right. Because phi of this is greater than alpha, what is remaining in the integral will be equal to alpha.

(Refer Slide Time: 50:14)



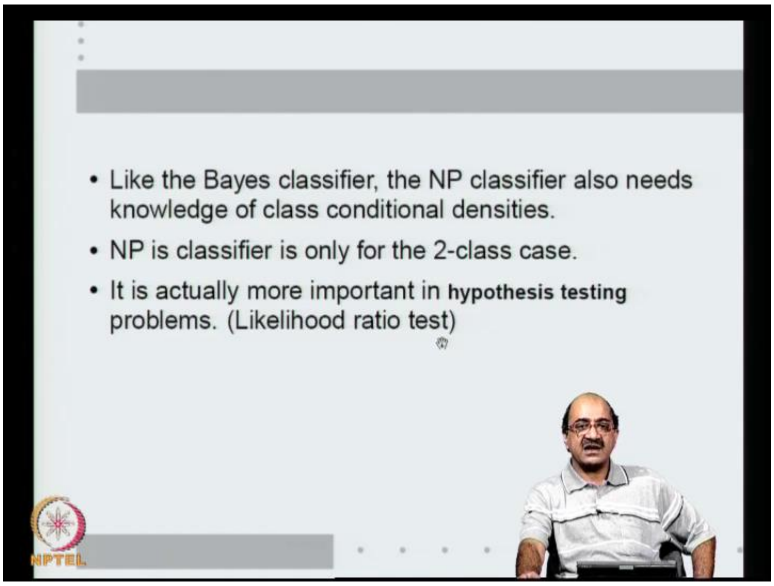
So let us let us look at this what it means, so if this is my f 0. I am choosing a tau as my threshold, such that this area. Area from tau to infinity is equal to alpha that is the allowed type 1 error.

(Refer Slide Time: 50:41)



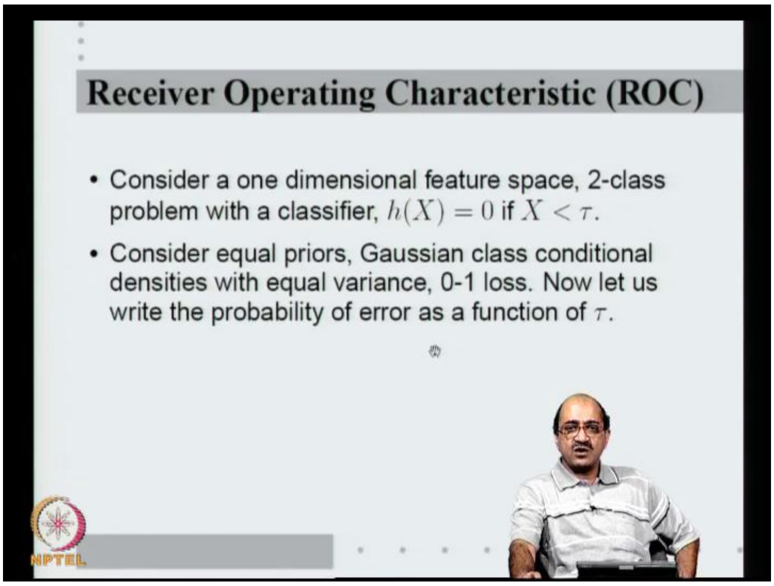
So, the Neyman pearson classifier ultimately that that ratio being some greater than K, turns out to be same as, X greater than tau, where tau is chosen, so that tau to infinity f 0 X d X to alpha, this is what, we want. Because, in in that normal and equal variance ultimately, the classifier is a threshold and type 1 error, because we are assuming mu 0 is less than mu 1, type 1 error is simply integral from tau to infinity of the class 0 density function.

(Refer Slide Time: 51:26)



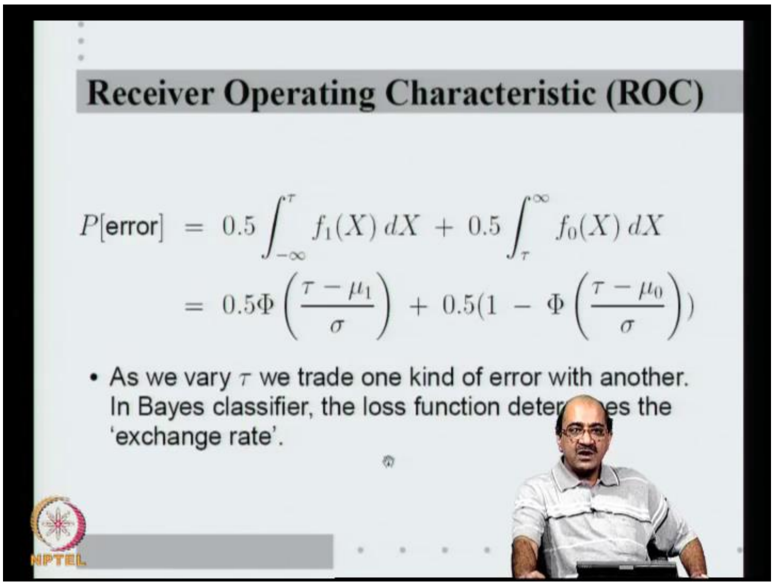
So, like the Bayes classifier, the Neyman pearson classifier also needs knowledge of class conditional densities and this classifier isis only for 2 class case, we did not defined for multiclass case and is often difficult to define it, for extinctive multiclass case. Just as general information, it is more important in certain statistics problems called hypothesis testing problems. More than classifier though for 2 class case is also used, especially when, you do not want to trade 1 kind of error with another rather than, that you want to put a bound onone kind of error and given that, bound is satisfied minimize the other kind of error.

(Refer Slide Time: 52:04)



Actually, as we have seen in the Neyman pearson classifier is for for trading one kind of error with another kind of error. So this kind of thing happens in many other ways a a good way of looking at it is what is called the receiver operating characteristic. If you consider a one dimensional feature space, 2 class problem with h x equal to with with a particular threshold tau. Now, if I if i think of class conditional densities once again as normal with equal variance 0 1 loss function.

(Refer Slide Time: 52:47)

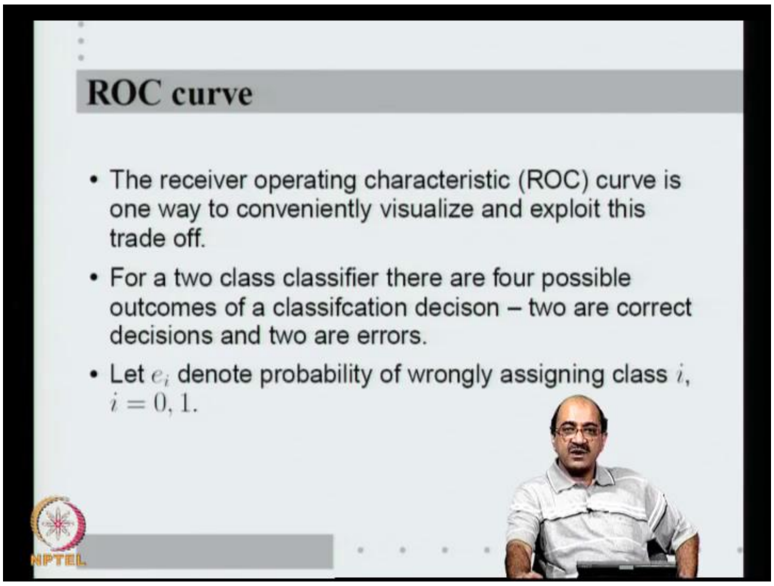


We can write the error as a function of tau, I am below tau, I am putting class 1 above tau, I am putting class 2. So, this is my error integral as, we have already written earlier right. So, my error integral is point, because priors are equal 0.5 times probability that class 1 pattern comes below tau and 0.5 times probability that class 0 pattern comes above tau right.

Once again I can if because, f 1 and f 0 are normal. I can write the standard normal as, we vary tau essentially, we are trading 1 kind of error with another, when I change tau 1 kind of error may increase and other kind of error will decrease. So, varying tau allows us to trade 1 error with another and hence atleast a threshold Bayes classifier, we can actually sit and decide, how we want to do the trade of right, risk function is 1 way of doing the trade of where the loss function gives me.

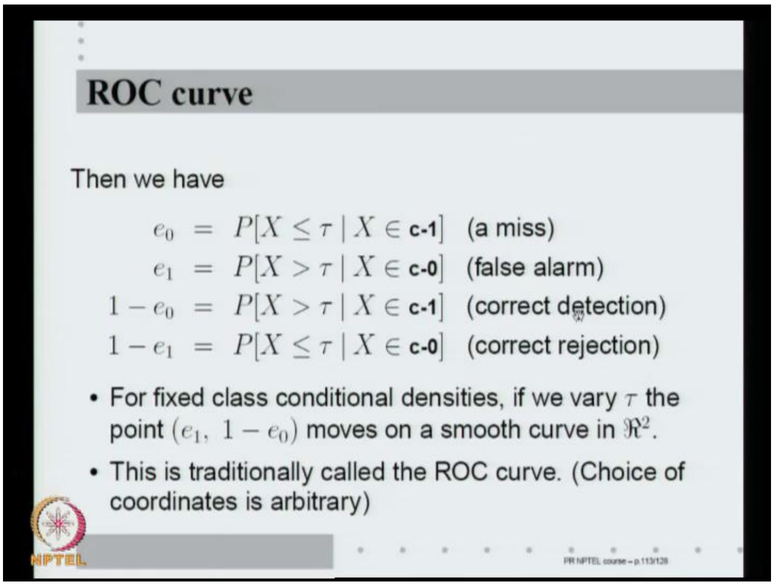
So, to say the exchange rate how much of 1 kind of error, I can trade for how much of other kind of error that these the relative values of the 2 losses. Neyman pearson criteria gives me another way of trading these errors right. I want this error below some alpha and and minimize the other error, but I can choose my own trade of right by using what is called a receiver operating characteristic.

(Refer Slide Time: 54:04)



The receiver operating characteristic curve is 1 way of conveniently visualizing, this trade off for a 2 class classifier. There are 4 possible errors right, 2 are correct 4 possible outcomes of a classification decision. 2 are correct and the other 2 are wrong, let e i denote the probability of wrongly assigning class I, that is e 0 is wrongly assigning class 0, at calling 0, when it is actually 1, let us say e 1 is wrongly assigning class 1, which means calling 1, when it is actually class 0.

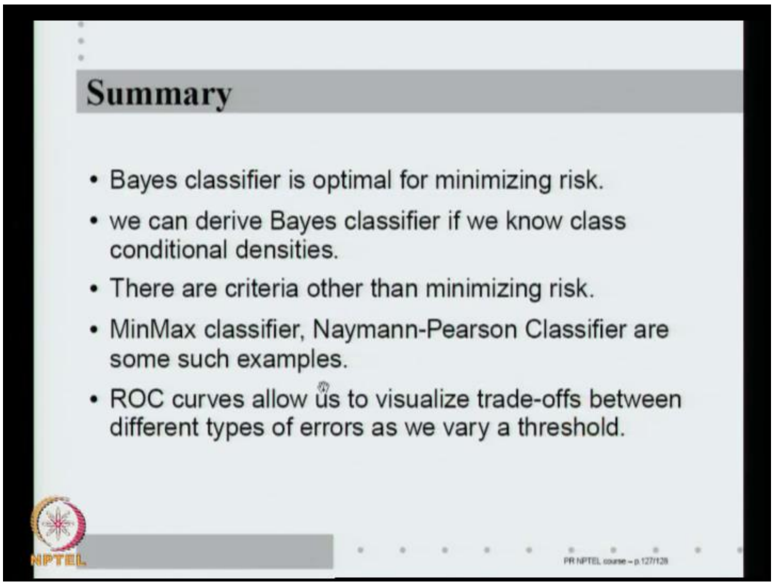
(Refer Slide Time: 54:38)



So, actually I can define these like this e 0 is the probability that, when X is actually in class 1, X is also less than tau. So, I call class 0, it is called a misse, e 1 is wrongly assigning class 1, that is even though, X is actually in for comes from c 0, because X is greater than tau. I call class 1 it is called a false alarm right, thethe 1 minus e 0 and 1 minus e 1 probabilities are called the correct detection and correct rejection.

For fixed class densities, If we vary tau, the point e 1 comma 1 minus e 0 moves on a smooth curve in r 2, ofcourse, I could have chosen any 2 of these numbers the choice of co ordinates is arbitrary. But, this curve, which plots the false alarm rate verses the the correct detection rate that is e 1 verses 1 minus e 0 right. For various values of tau right, that is on the e 1 1 minus e 0 plane, for each tau there is 1 value of e 1 1 value of 1 minus e 0. So, as I vary tau it becomes a smooth curve and such a curve is called the receiver operating characteristic. This is another way of trading 1 kind of error with the other kind of error ok.

(Refer Slide Time: 56:04)



So, let us close down today is class with a summary, the Bayes classifier is optimal for minimizing risk, we have seen that in last class and we have seen a couple of more examples, this class. We can derive Bayes classifier, if we know class conditional densities and for various kinds of loss functions, we can derive. There are criteria other than minimizing risk as we have seen minimizing risk is not the only way, we can we can run this problem.

For example, minmax classifier, Naymann pearson classifier are some examples of criteria other than minimizing risk. All of these are essentially trying to trade of errors in a way different from the trade of that to the Bayes classifiers does, Bayes classifier trades of errors, using the loss function of the exchange rate. Where as, there are other ways of trading of errors and receive a operating curve characteristic curves allow us to visualize the trade of between different types of errors, as we vary a threshold. We will once again briefly look at the receiver operating characteristics next class.

Thank you.