Statistical Methods for Scientists and Engineers
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology – Kharagpur

Lecture – 24
Multivariate Analysis – IX

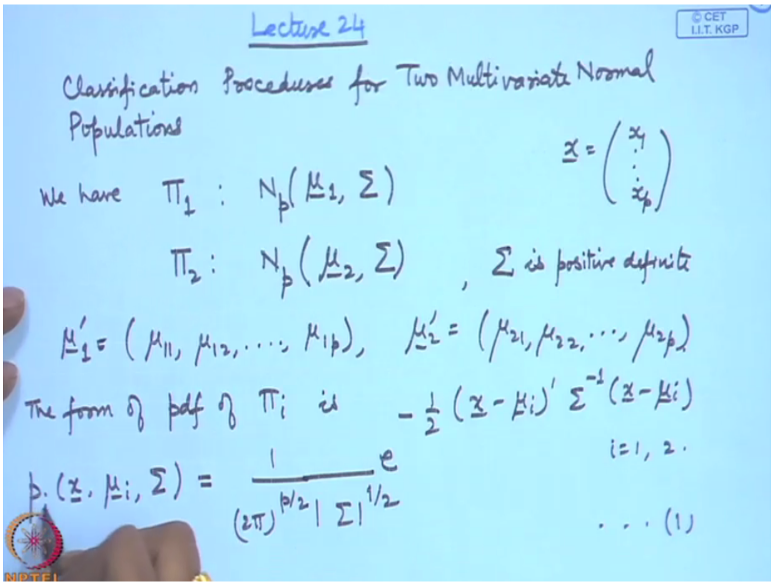
Friends, yesterday I have discussed in detail the procedures of classification into 2 populations. We have given a general framework. one is called the Bayesian framework and another I call as the Minimax method of procedures and what we have done is we have assumed the probability distributions say p1(x) and p2(x) the 2 populations and we have given that if we know in advance that the proportion of the 2 population.

That means how many observations actually belong to the first population and how many to the second. That means we can assign a priori probability say q1, q2 then we can develop a Bayes procedure. That means which will minimize the expected probability of misclassification. We also gave the concept of admissible procedure, minimax procedure, etc. and in particular, we proved that every Bayes procedure is admissible and every admissible procedure is Bayes.

And therefore the class of all the Bayes procedure is the minimal complete class. In particular, a member of the class of Bayes rules will be minimax procedure. Therefore, for all practical purposes we can restrict attention to the rules which are of the Bayesian form and the form is also of a very nice nature that we got that say p1(x)/p2(x) is greater than something or p1(x) / p2(x) is less than something.

So then, you classify in the population pi1 or pi2. So now, this gives the general framework for preparing classification rules for various problems. Now the problem of classification usually started on the discussion on the normal distributions. So we firstly discuss the procedures for that.

(Refer Slide Time: 02:31)



So classification procedures for 2 multivariate normal populations. Let me state the problem first. So we have say population pi1 it is specified by say p dimensional normal distribution with mean vector say mu1 and (()) (03:11) covariance matrix say sigma and pi2 where p dimensional normal distribution with mean vector say mu2, sigma. It is like this you can think of for example there is a patient.

A patient goes to a say medical practitioner and certain tests are conducted on the certain measurements. So it could be blood test and it could be certain other measurements on the patient and then it has to be decided. For example, the first population may correspond to a particular disease and second population parameters may correspond to another disease. So on the basis of the observations on the patient that is x.

We are having say x = x1 to xp we have to decide whether they are matching more with pi1 or more with pi2. So this is a classical example. We can think of in other areas also like land classification or classification on the basis of the economic characteristics of a country or individual or an organization. So these are the problems were we model according to the multivariate normal distribution.

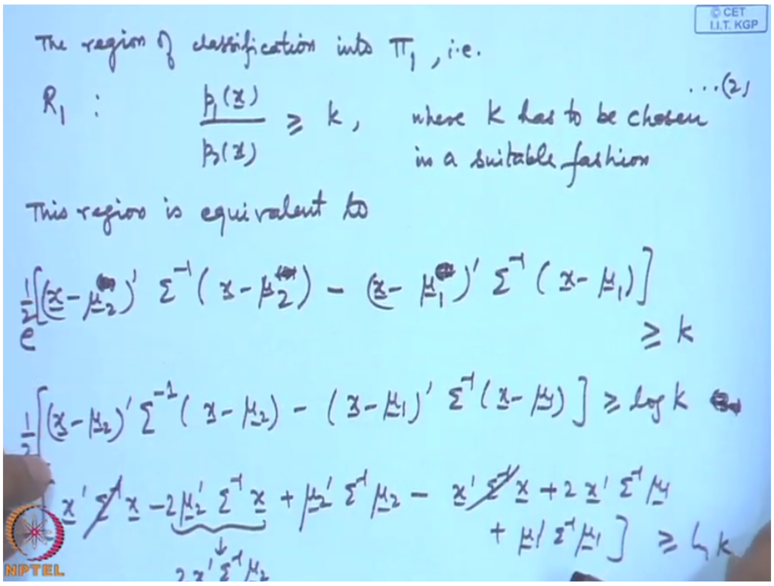
That means different characteristic, different components will individually normally distributed and at the same time, the correlated structure is giving you a multivariate normal distribution. Now in the first model I am starting with the covariance matrix to be the common. So here we specify like this the mu1 is actually your vector. So let me write it in the form of row vector mu1 1, mu1 2 and so on mu1 p and similarly your mu2 vector is mu2 1, mu2 2 and so on mu2 p.

So the form of the, we assume that sigma is positive definite. If we assume positive then the density functions can be written in the form of the pdf of pi is: so let us say pi (x, mu i, sigma) that is 1/2 pi to the power p/2 determinant of sigma to the power 1/2. Then we have in the exponent – 1/2(x – mu i) prime sigma inverse x – mu I for i = 1, 2. Let me call it question number (1).

If we want to classify according to the rules that we discussed yesterday, then the Bayes rules or the Minimax rules. So the class is, class of all admissible rules is of the form that is p1(x)/p2(x) say > k or <= k says for classifying k into first population or in the second population. Of course, when the priori probabilities are known then directly the form of k is known to us. That is q1 / q2 kind of thing.

But even it is not known then it is by Bayes rule with respect to some prior, okay. So that means the desirables are of this form only. If the prior probabilities are given, we can choose the corresponding Bayes rule in that class. If that is not, there we can choose any rule or we can choose the Minimax choice. So but the framework is given that means all the consider the rules of this nature.

(Refer Slide Time: 07:47)



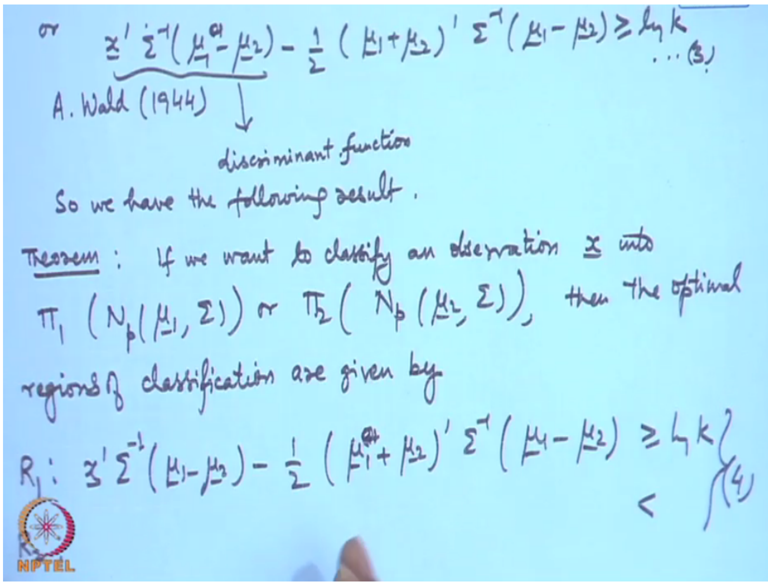
So let us consider the region of classification into pi1 that is we call it R1. That is the p1(x) / p2(x) is >= k where k has to be chosen in a suitable fashion. So if we substitute here. p1(x) and p2(x) here then this term will be cancelled out and here we will be left with e to the power. Let me write down the expression. See in numerator we are writing p 1 so here it will become mu1 and in the denominator, we have mu2.

So it will come in the mu2 here. So if I write the ratio here. So it will become something like this. (x – mu2) prime sigma inverse (x – mu2) – (x – mu1) prime sigma inverse (x – mu). This I have noted here.and you are getting e to the power 1/2 of this >= k. So if I take logarithmic here and I arrange the terms. So this will become 1/2 (x – mu2) prime sigma inverse (x – mu2) – (x – mu1) prime sigma inverse (x – mu1).

This is >= log of k. Let us call this as (2), this as (3) or let us further simplify this. So if we consider the expansion of these terms. I will get x sigma inverse x – mu2 prime sigma inverse x. It is 2 times here which you can also write this term as twice x prime sigma inverse mu2 and then you have + mu2 prime sigma inverse mu2 – x prime sigma inverse x + 2x prime sigma inverse mu1 + mu1 prime sigma inverse mu1.

So after expansion these are the terms I will be getting. You can see that this gets cancelled out and you can adjust the remaining terms as x prime sigma inverse (mu1 – mu2) and you are left with this term here I can adjust the terms for actually getting mu2 prime sigma inverse mu2 and this is actually becoming - here, - mu1 prime sigma inverse mu1.

(Refer Slide Time: 11:40)



So I can add and subtract the terms corresponding to mu2 prime sigma inverse mu1. If I do that then I can factorize and write this term as – 1/2 (mu1 + mu2) sigma inverse (mu1 –

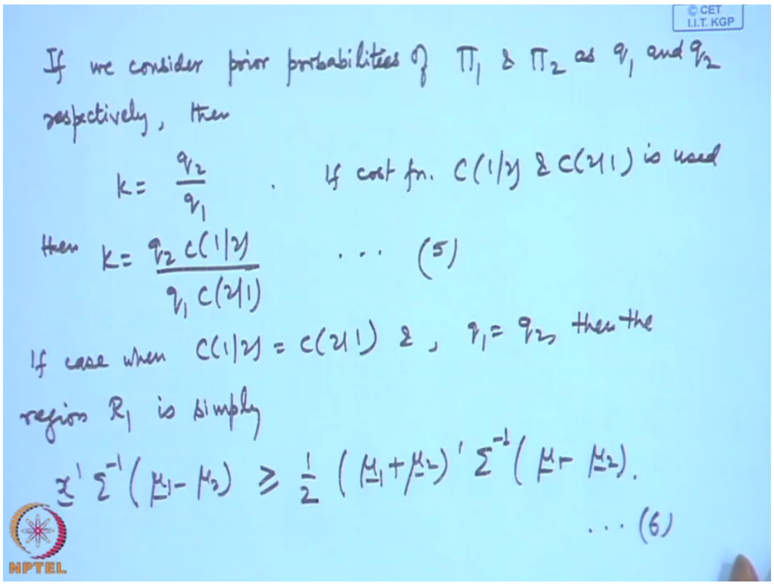
mu2). This is >= log of k. So this result was actually obtained by Abraham Wald in 1944. This is actually called the discriminant function because this term will be same for all observations.

When you are taking the observation x, which you want to classify then this part, is actually used for discriminating between the populations. So we call this as the discriminant function. So we have the following result. Then we are considering the density function of the form (1) that is multivariate normal distribution the best regions of classification are given by.

If we want to classify an observation x into pi1 (Np (mu1, sigma)) or pi2 that is (N p(mu2, sigma)), then the optimal regions of classification they are given by say R1 where we write x prime sigma inverse(mu1 – mu2) – 1/2(mu1 + mu2) prime sigma inverse(mu1- mu2). This is >= log of k. This is for the classification into pi1 and for pi2, this will become simply less here.

Why am I saying these are the optimal regions of classification? Because in the previous lecture I have proved that, the minimal complete class is exactly the class of Bayes rules. That means for good rules we do not look beyond the Bayes rules and the Bayes rules are of the form p1(x) / p2(x) >= k where k would be suitably chosen. It is a number between basically it is a ratio q1 – q2 particularly depend upon what values of q1 and q2 we are chosen. That is why I have written where k has to be suitably chosen.

(Refer Slide Time: 16:04)

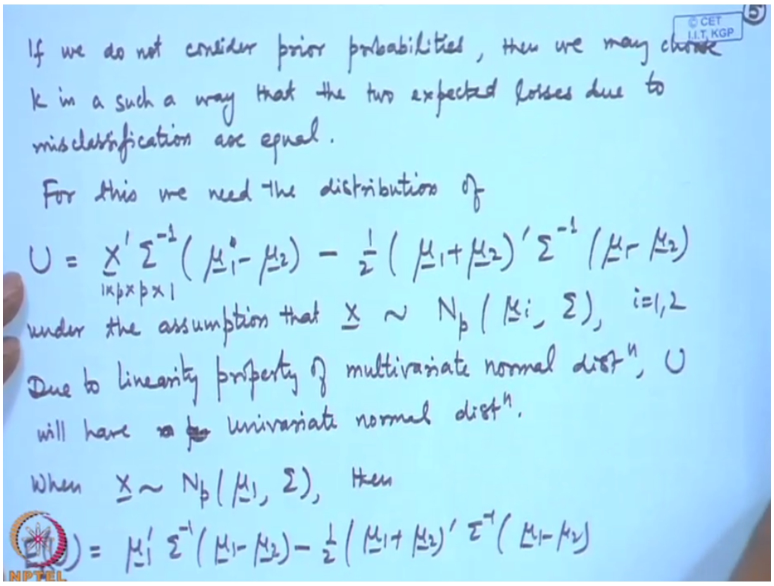


In particular, if we consider prior probabilities of pi1 and pi2 as q1 and q2 respectively. Then k is nothing but q2 / q1. If cost function c (1 | 2) and c (2 | 1) is used, then k = q2 c (1 | 2) and q1 c (2 | 1). So this can be chosen. In general, k can be anything but all of the choices will give you a rule in the minimal complete class. That means it is an admissible rule and it is a Bayes rule.

Now you were taken by extreme case when q1 and q2 are same, that means we do not discriminate between the 2 populations. In case when c (1 | 2) = c (2 | 1) and q1 = q2 then the region R 1 will simply because k will become I and therefore log of k will become 0. So this will become x prime sigma inverse (mu1 – mu1) is >= 1/2 times (mu1 + mu2) prime sigma inverse (mu1 – mu2). Now there can be a question where priori probabilities are either not assumed simply we have no information.

That means we cannot discriminate between 2 populations on the basis of prior probabilities. In that case, 1 can look at the, that means we can look at that we make the expected losses due to misclassification as the same. So let me just that point here.

(Refer Slide Time: 18:52)



If we do not consider prior probabilities, then we may choose k in such a way that the 2 expected losses due to misclassification are equal. That means I will need the probability of this. That is classifying into pi1 when it is belonging to pi2. That is under the assumption that x is having mu2 sigma and the other one will be less than this that must be classified to but we assume x to be in pi1.

That means x is following N p (mu1, sigma). So that means we need the probability of this statement greater than or equal to or less than. So in order to see that we actually need the distribution of. For this we need the distribution of, so this quantity I denote by U = x prime sigma inverse (mu1 – mu2) – 1/2 (mu1 + mu2) prime sigma inverse (mu1 – mu2) under the assumption that x follows N p (mu i, sigma), i = 1, 2.

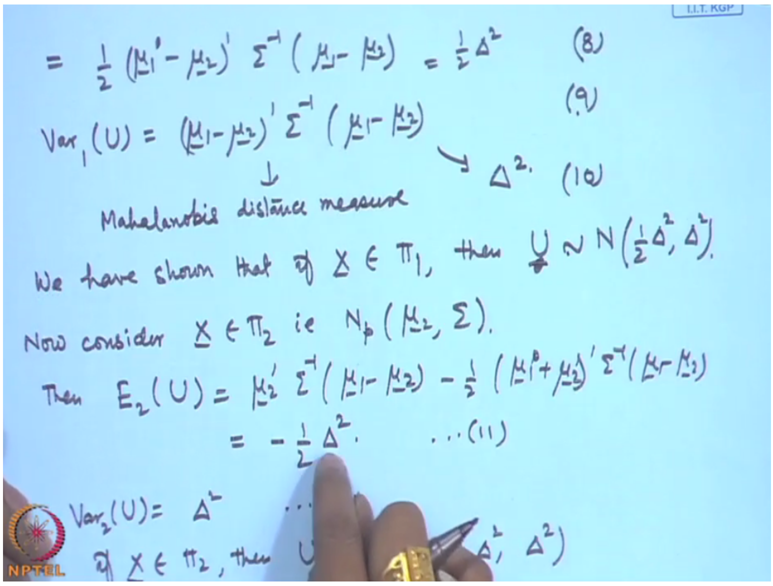
Now you can use the linearity property of the multivariate normal distribution here. See if x is following, since x is following N p (mu i, sigma) therefore x prime sigma inverse (mu1 – mu2) we can actually obtain. Due to linearity property of multivariate normal distribution, U will have p. So basically, the dimension will remain the same because this is also, so this will become univariate normal basically.

Because what is happening is x is a multivariate normal now what term you are writing is becoming a scalar quantity because this is 1 by p the you are having p by p and then you are having p by 1. So this is becoming a scalar quantity. So U will have a univariate normal distribution. Now let us calculate it separately when x follows N p (mu1, sigma) then it will be expectation of say U, we will call it expectation 1.

Then here it is becoming equal to mu1 prime sigma inverse (mu1 – mu2) – 1/2 (mu1 + mu2) prime sigma inverse (mu1 – mu2). Now this term we can simplify. Here you look at this is actually becoming mu1 prime sigma inverse (mu1 – mu2) and here I get 1/2 mu1 prime sigma inverse mu1. So this is the – sign it will become +. Similarly, you look at the cross product term that is mu1 prime.

In fact, if you go back to the original term where I derived this from there it will be clear how this term is coming. Initially I have written here. This term was mu2 prime sigma inverse mu2 - -mu1 prime sigma inverse mu1 which is written like this particular term. So this term is in the + and this is in the – here. So this is – and this is being cancelled out here – this one. So 1/2 is there so it will become + 1/2.

(Refer Slide Time: 24:28)



So then this term can be simplified to which is = 1/2 (mu1 prime – mu2) prime sigma inverse (mu1 – mu2). Let us put some question number here. So this is the definition of U let me put as (7) and this I put as (8) and then in this cases what will be the variance of U. For the variance you have the formula that because for x it is sigma. So it will become mu1 – mu2 prime sigma inverse then sigma and then sigma inverse this term will be coming.

So sigma sigma inverse will become identity. This term will be remaining. That means you will get it as (mu1 – mu2) prime sigma inverse (mu1 – mu2). Let me call it equation number (9). We also write here see this particular term which is written here actually this is basically a major, a distance major whi9ch was given in 1930 by P C Mahalanobis and it is called Mahalanobis distance major and it is called Mahalanobis d square.

This is the Mahalanobis distance. So let us call it as delta square. We give this term name as delta square, okay. So what we have, actually we can write in terms of this here. Expectation of mu1 is basically 1/2 of this. So this is 1/2 delta square. So what we have proved is that we have shown that if x belongs to pi1 then this U is following normal distribution with mean 1/2 delta square and variance delta square.

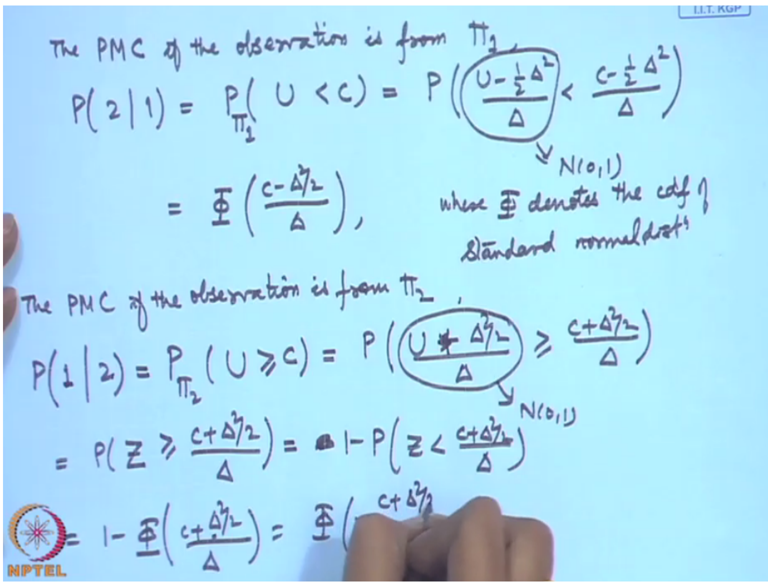
Now consider the other case that is N p (mu2, sigma). Then expectation of U in this case what will happen it will become mu2 prime here. So this will become mu2 prime sigma inverse (mu1 – mu2) – 1/2 (mu1 + mu2) prime sigma inverse (mu1 – mu2). Once again this can be simplified here. See this is – mu2 prime sigma inverse mu2 and here I will get + mu2 prime sigma inverse mu2.

So this will get cancelled out and you will get – and this term is -. So basically, you will get – 1/2 of delta square and similarly if you look at variance that will be same because in the variance that term does not change. That is, we are saying that if x belongs to pi2 then the distribution of U is normal with – 1/2 delta square and delta square. Now you can see this result is very interesting.

We have used U for basically discriminating between that populations pi1 and pi2 and here you can see the clear demarcation the average values of U under pi1 and pi2. They are showing opposite behaviour like here it is 1/2 delta square and here it is – 1/2 delta square and delta square I am giving a name Mahalanobis distance measure.

So if that 2 populations distance is given in terms of delta square then clear cut demarcation between the populations pi1 and pi2 is coming that means if it is actually belongs to pi1 then the mean of that is 1/2 delta square and in other case it is becoming - 1/2 delta square. So it is exactly on the opposite side. So this is quite interesting and you can think that heuristically it is actually a good classification rule. Now let us look at that, we want to make the 2 expected probabilities of misclassification to be the same then let us consider this.

(Refer Slide Time: 30:31)



So the probability of misclassification if the observation is from pi1. So that is p (2|1) that is now you classify into 2 if you are getting x prime sigma inverse (mu1 – mu2) less than this quantity. Basically you are saying U is < c. So our original classification rule that I have described here it is terms of U only. So basically this term was U. So U greater than some quantity or U less than some quantity. So we will use exactly that thing.

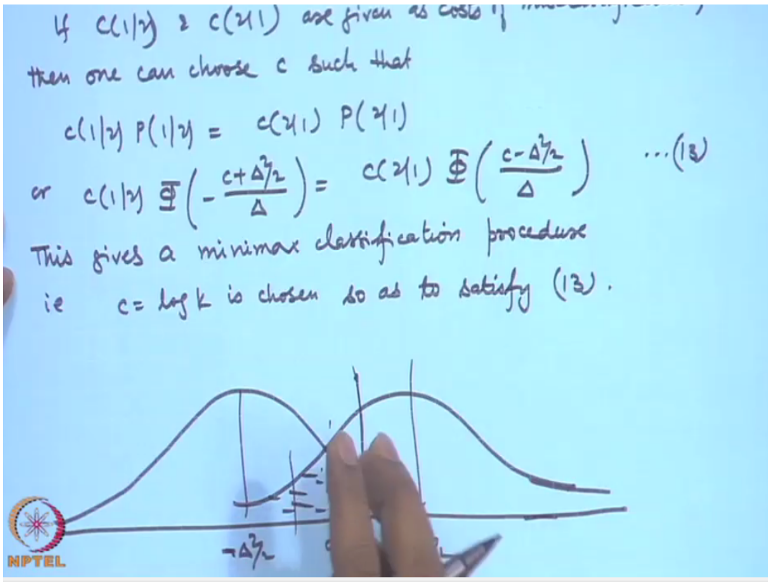
So it is probability of U < c when the true population is say pi1. Now under pi1 we have just now derived that U follows a normal distribution with mean 1/2 delta square and variance delta square. So we can consider probability of (U – 1/2 delta square / delta < c – 1/2 delta square / delta). So this is becoming standard normal random variable. So this can be written in terms of the cumulative distribution function of standard normal (c – delta square/ 2) / delta phi denotes the cdf of standard normal distribution.

Similarly, let us calculate the probability of misclassification if the observation is from pi2. If the observation is pi2 then the probability of misclassification is P (1|2) that is the observation is from 2 but I put it in 1 that is P pi2 of U >= c. So that is equal to probability of U when U is from, x is from pi2 then U has normal – 1/2 delta square , delta square.

So this will become U greater than well we will put it as (U + delta square / 2) / delta >= (c + delta square/2) / delta and again this is having a standard normal distribution. So this is equal to probability of Z where Z is a standard normal random variable (c + delta square / 2) / delta. So this is nothing but 1 – P (Z < (c + delta square / 2) / delta which is nothing but actually (c + delta square / 2) / delta which we can also write as phi (-c + delta square / 2) / delta.

So you can see here we are able to evaluate that 2 probabilities of misclassification. So we can choose c such that these 2 are same. If there is a cost function, then we can include that also.

(Refer Slide Time: 34:37)



If c (1|2) and c (2|1) are given as costs of misclassification, then one can choose c such that c (1|2) p (1|2) = c (2|1) p (2|1) or c (1|2) phi (- c+ delta square /2) / delta = c (2|1) phi (c – delta square/2)/delta. Let me call it as question number (13) here. Now this is quite interesting we are actually able to restrict our attention to a rule for which the expected probability of misclassification is same.

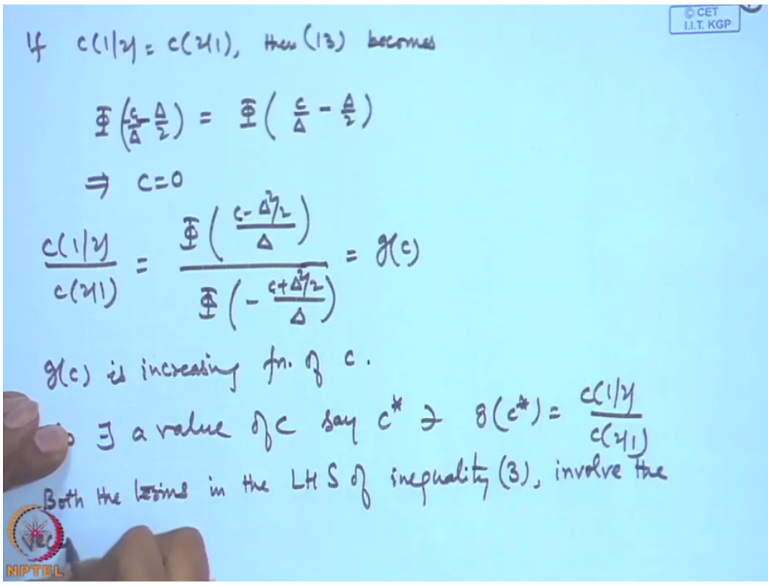
Since all the terms in this equation will be known because delta is based on the mu1, mu2 and sigma, which is the parameters of the 2 populations. c (1|2) and c (2|1) are the costs of this classification which will be some numbers. So all the terms here are known that means from the tables of the normal distribution that is the tables of the cumulative function of the normal distribution we can actually find c for which these 2 values will be are these 2 equations will be satisfied.

So this is actually Minimax classification procedure that is c = log k is chosen so as to satisfy equation number (13). You can look at this theorem. This is the point you have – delta square / 2 and this is the point + delta square / 2. So of course there will be so this is some 0 here say. So this is the (()) (37:27) area here and c will be somewhere here. May be c is here or it could be here etc. So we cut of like this actually if the point is here or here then there is no problem.

But in this portion, we have to decide whether we have to put in the population 1 or in the 2. So depending upon the value of c, which is here, or here etc. depending on the nature of mu1 and mu2. Because mu1 and mu2 will affect the value of delta square. If there is a large difference or if there is a small difference and also the magnitude of sigma all of this will affect the value of c.

So it could be that this intersection part is very small and in that case the classification will be good. If the intersection is more then suddenly classification rule will be slightly worse. That means the determination power of the rule will be much less. If we consider say c(2|1) = c(1|2) then it becomes much simpler problem. Because if you have c(2|1) = c(1|2) then this equation is reducing to. Let me write it here.

(Refer Slide Time: 38:56)



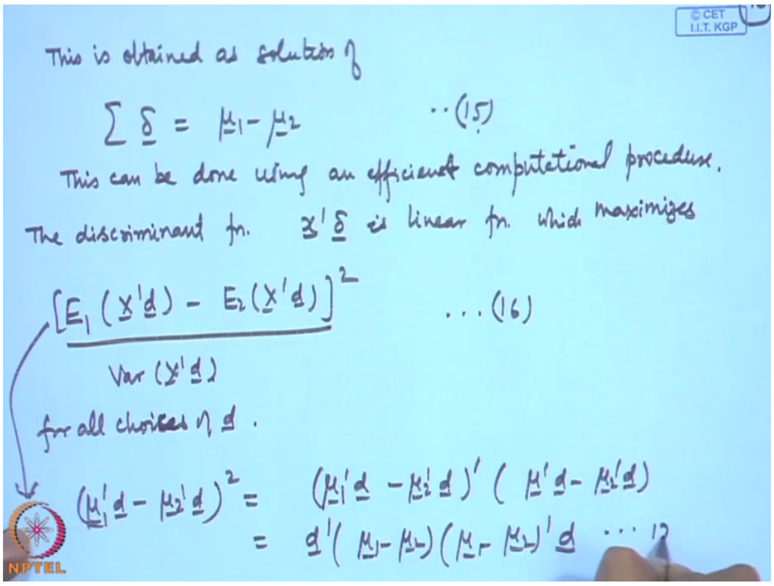
If the cost terms are, also same then (13) becomes. The first part will become – delta / 2 and the right hand part will become, sorry, that is (– c /delta) – (delta/2) = (c/delta)-(delta/2). So you can see here c = 0. So the rule which is written for q1 = q2 that means when we equate the 2 that is actually coming as the Minimax classification procedure. So if the costs are the same and if q1 = q2 then whatever rule is obtained that is actually become the Minimax rule.

But if c (1|2) is != c (2|1) then certainly you will find from the tables of normal distribution using the cdf of the standard normal distribution. We can also notice some further fact. We can also see that the ratio c (1|2) and c (2|1) then what do we get.c (1|2)/c(2|1) = phi ((c – delta square / 2)/delta) / phi ((- c + delta square / 2)/ delta). Let us call it say some g(c). This is, this g(c) is because if I consider increasing c then this will decrease.

But then it is in the denominator and it is the non- negative function. So this will increase when this is increasing. So this is an increasing function of c. So if it is an increasing function then certainly there exist a value of c for which equality will be attained. So there exists a value of c say c * such that g(c*) = c(1|2)/c(2|1). That means a solution will always exist. See both the terms in the expression (3) that means go back to the expression (3) here.

That is the original discriminant is here x prime sigma inverse (mu1 – mu2) – 1/2 (mu1 + mu2) prime sigma inverse (mu1 – mu2). So look at this. This part and this part is common. If we call it say delta, then basically we are looking at the solution of the equation of the form sigma delta = mu1 – mu2. So we can consider the involved vector delta is equal to sigma inverse mu1 – mu2.

(Refer Slide Time: 43:19)



This is obtained as solution of sigma delta = mu1 – mu2. This is regarding the computational part of x. because we suddenly need to calculate and this involves the inverse here. Actually, this term is not difficult x prime and then mu1 + mu2 prime. Only difficulty is to evaluate sigma inverse and then of course multiplication with something. So if we look at the solution of this type that means delta equal to this.

So if I consider some efficient computation procedure, numerical computation procedure then we can actually obtain the solution. For example, (()) (44:16) or any other method this is basically becoming a system here because I need the solution delta = mu1 – mu2. So this can be done by using an efficient computational procedure. We have further interpretation of this that the discriminant function x prime delta is linear function which is maximizing [E 1 (x prime d) – E 2 (x prime d)] whole square / var (x prime d) for all choices of d.

In fact, if you look at the numerator, the numerator here is (mu1 prime d – mu2 prime d) whole square. But this we can also write as (mu1 prime d – mu2 prime d) and prime of that into (mu1 prime d – mu2 prime d). If we write like this, then it is becoming d prime(mu1-mu2) (mu1 – mu2) prime d. So we can express this in a different way here and in a similar way if you look at the denominator here this var (x prime d) = d prime E(x – E(x)) (x-E(x)) prime d that is actually d prime sigma.

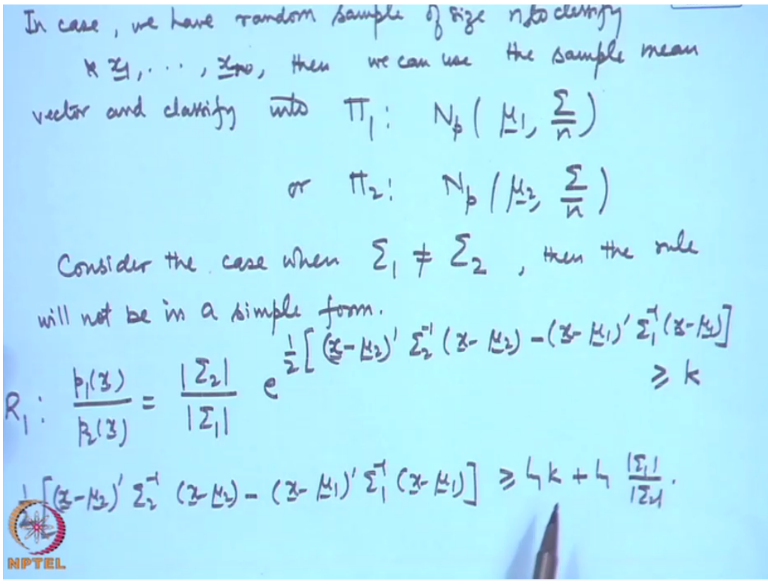
So basically we consider that we want to maximize (17) with respect to d such that 18 is a constant. So basically we are considering the term d prime (mu1- mu2) (mu1 – mu2) prime d – lambda (d prime sigma d – 1). So this is called Lagrange multiplication. Lambda is the Lagrange's multiplier. So you can consider derivative of this with respect to d equating to 0 what we will get? (mu1- mu2) (mu1– mu2 prime) d = 2 lambda sigma d.

There will be 2 here also so this 2 is actually cancelled out and this is actually a scalar. So we can actually write it as some nu, mu1 – mu2 = lambda/nu, this quantity I am calling nu, into sigma d. So you consider the solution is proportional to delta. Because delta you see here was sigma delta = mu1 – mu2 that is the solution. Here you can see that the solution is proportional to delta.

Now here it is a classification of a single observation x into 2 populations. But the general problem of classification is that in place of 1 observation we may have a sample of observations. If we have a sample of observations in that case, we can consider the distribution because of the sufficiency in multivariate normal situation x bar and x are the sufficiency statistics.

We can actually consider the distribution of x bar. So x bar will have N p mu sigma1/n and sigma 2/ n. Well, mu1 sigma/n and mu2 sigma /n. So the problem is just shifted in place of sigma we are considering sigma/n and the entire procedure will be the same.

(Refer Slide Time: 50:28)



So let me just mention this thing here. In case, we have random samples of size n to classify. That is the samples say x1, x 2, ..., xN then we can use the sample mean vector and classify into pi1 that is Np (mu1, sigma/n) or pi2, … Np (mu2, sigma/n). This is for the purpose because we are this from the first population then the sample mean will have this distribution and if it is from the second one then the sample mean will have this distribution.

So the entire problem is just modified and the procedure will remain the same. So this particular problem that I have discussed now it is for the classification when the parameters of the population are known and therefore the procedure that I described in the previous lecture is completely applicable here. That means I am able to derive a procedure which can be Bayes considering the choices here because the density functions are completely known.

In case the prior probabilities are not assumed, we can find out the Minimax choice. I have shown in the particular case that the choice can be explicitly found from the tables of the normal distribution.