Applied Multivariate Analysis

Prof. Amit Mitra
Prof. Sharmishtha Mitra
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

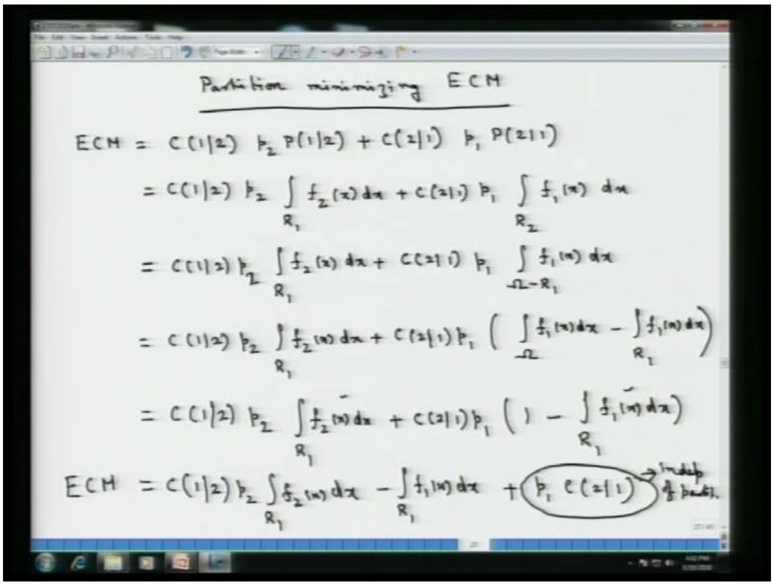
Lecture No. # 32
Discriminant Analysis and Classification

In the last lecture, we had started discussing about deriving optimum classification rule based on certain criterion. For example, we had introduced under a general classification framework. What we mean by a total probability of misclassification. We had also talked about another criterion which talks about expected cost of misclassification. In the last lecture, we had derived specifically the rule which corresponds to one that would minimize the total probability of misclassification.

And we had also shown that when we talk about a total probability of misclassification optimizing rule; that is a classification rule on the partition of the sample space which leads us to the rule which minimizes the total probability of misclassification; that also is same as that of the Bayes rule. Now, today what we are going to look at in this lecture is first we will look at what is the optimum rule that we are going to get when we talk about expected cost of misclassification. Now, expected cost of misclassification is a criterion that is attached with once we have got a classification rule.

Then there is ofcourse, as we have discussed that there is a possibility of an observation coming from one population getting misclassified into another population. Now, along with that we also put some cost constraints in the sense that suppose an observation coming from population pi 1 is misclassified to pi 2, then there is a cost attached to that and vice versa. And accordingly, if we have correctly classifying an observation coming from population number 1 into population number 2 1 itself, then there is no cost as such of misclassifying. And hence, we take C i C 1 1 or C 2 2; both of them to be equal to 0 in both the situations.

(Refer Slide Time: 02:12)



Now, let us first look at today that particular thing that I said, we are looking at the partition or the optimum partition, which minimizes the expected cost of misclassification. So, we will first look at this particular thing, and then look at some examples; some examples corresponding to a multivariate normal distribution. So, what is our expected cost of misclassification? Let us recall what that was... So, we have got a cost C 1 given 2; that is an observation coming from population number 2 is misclassified into population number 1. So, this is the cost which is attached with such an event that we are looking at a misclassification cost of misclassifying an observation coming from 2 into 1.

And then the corresponding probability of this would be given by p 2 P 1 given 2 this plus the cost that we incur in misclassifying an observation coming from one into the corresponding probability, which is given by this particular expression; where these we had defined earlier. Now, in terms of the partitions R 1, R 2; remember, we had said that we are talking about partition. So, partition of the sample space; so, it is going to divide the sample space into region R 1 and R 2; wherein if x belongs to R 1, then we classify it into pi 1 and if x belongs to R 2, then we classify it into the second population; that is the pi 2 population.

So, this can be written in terms of this is getting classified into 1 and hence this region is R 1 and then the population in that particular object is coming from population number R 2. So, it has got the density f 2 in our notation in our earlier notations f 2(x) dx this plus

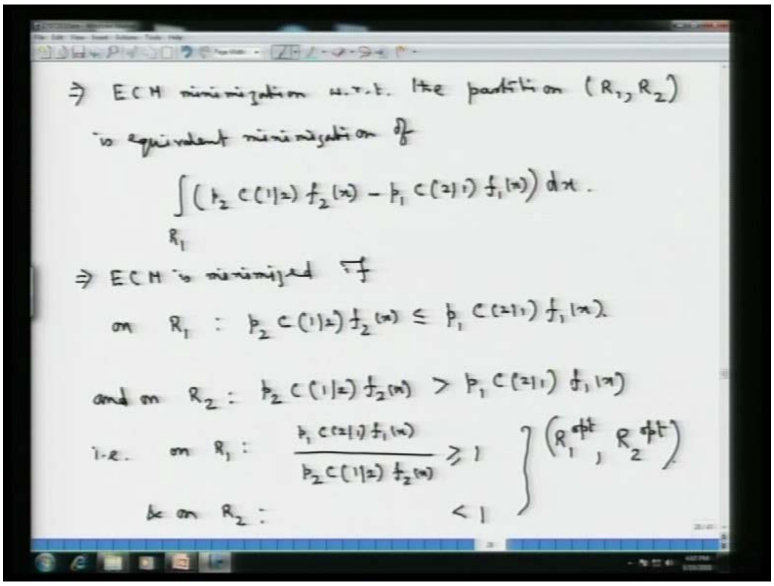
C 2 given 1 into p 1. These small p i's are the prior A priory probabilities of the corresponding populations; this integral over the region R 2; because we are classifying it into the second population and then it is coming from the first population. And hence, what we have is this particular term here.

Now, if we look at this particular term, we can write it in terms of the complementary region of the first partition R 1 segment. So, this can be written first term as it is and the second term, we can write as this into omega minus R 1; that is the region R 2 this over f 1(x) dx. So, what we can see from this expression is the following that C 1 given 2 into p 2 integral over R 1 f 2(x) dx this plus… Now the first integral, this now gets splitted into these particular two terms; this p 1 this integral the first integral over omega f 1 (x) dx this minus integral over R 1 f 1(x) dx. So, this is going to lead us to this particular expression; now what it is written here.

Now, this term here; this is integral over the entire space omega and hence this integral would be equal to 1 just and hence, we can write this expected cost of misclassification in a compact form in R 1 f 2(x) dx this plus this particular term. Let me write this term before the second term or let us just stick to whatever orientation we are having; this into 1 minus integral over R 1; we leave it as it is f(x) dx. Now, thus collecting this term and the term corresponding to this; we can write this as C 1 given 2 p 2 into integral of R 1 f 2(x) dx this minus integral over R 1 f 1(x) dx this plus p 1 times C 2 given 1.

I think to be noted here in this particular expression for expected cost of misclassification is the following that this is a term, which is independent of the partition. So, this is a term which is independent of partition and hence the optimum partition, when we are looking at we are looking at R 1 opt, R 2 opt. So, that is the optimum partition that is what we are looking at and this is not going to play any role, when we are trying to minimize this particular expected cost of misclassification with respect to the partition that is with respect to R 1 and its complementary region with respect to the sample space.

(Refer Slide Time: 07:22)



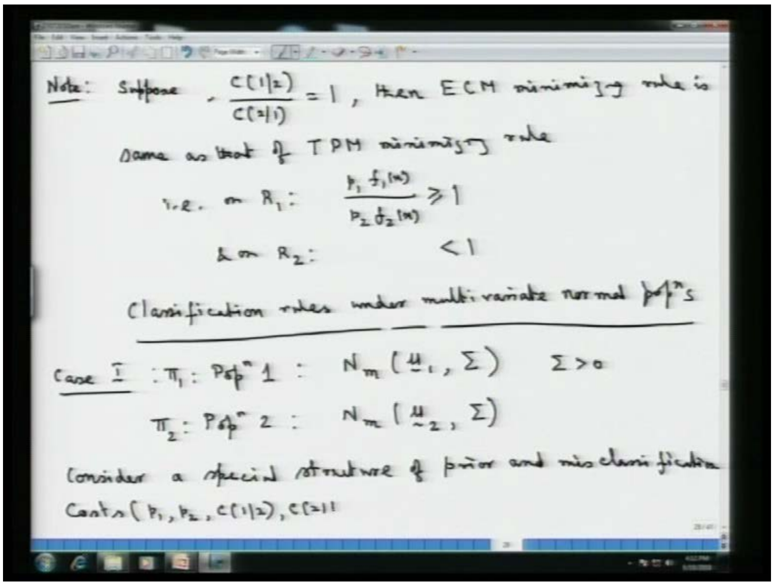
So, what we have here is that this would imply that expected cost of misclassification minimization is equivalent to minimization with respect to the partition the partition is our R 1, R 2 is equivalent to minimization of the following quantity of (Refer Slide Time: 02:12) this first expression here. I missed out something; this constant also comes in here. So, this term has got a constant multiplier out here, which is C 2 given 1 that times p 1. So, this term when multiplied with this term leads us to C 2 given 1 into p 1 into this particular term. So, when we are looking at finding the partition which would minimize the expected cost of misclassification, we can look at just this particular term and look at what is that partition with this, which is leading us to the minimum of the expected cost of misclassification.

So, this is minimization of the term; that is what we have integral over R 1; writing it in one expression, this is going to be p 2 into C 1 given 2 this multiplied by f 2(x) that is coming from (Refer Slide Time: 02:12) the first term here. So, C 1 given 2 p 2 into f(2) x that minus this term in to f 1(x); so, this minus p 1 C 2 given 1 into f 1(x) d x. So, this is where, we are trying to find out R 1's is this is minimized. And thus the E C M minimizing rule, this would imply this would imply that E C M is minimized if on R 1, we have this quantity to be less than or equal to this particular quantity and hence this is where the region comes in. So, this is p 2 into C 1 given 2 that times f 2(x) this is less than or equal to p 1 times C 2 given 1 f 1(x).

And on R 2, we will have the other way round that is if p 2 into C 1 given 2 into f 2(x); this is greater than p 1 times C 2 given 1 into f 1(x). So, this R 1, R 2 partition, so this is the set of all x's such that this quantity is less than or equal to the right hand side. So, that is the region of all x's for which x is going to be classified into pi 1 and this is the set of the complementary x's, for which this expression is strictly greater than this right hand side here. That is, in other words on R 1, we will have in terms of these quantities which is our p 1 C 2 given 1 to f 1(x) this divided by p 2 C 1 given 2 f 2(x). This is going to be greater than or equal to 1 and on R 2, the same quantity is less than 1.

Now, if we have this to be the optimum partition, we call that this say R 1 opt, R 2 opt. So, this is the optimum partition R 1 opt and R 2 opt. So, this is what is leading us to… Now note that in this particular situation, if one looks at this particular region this partition of the sample space; if we have the corresponding cost quantities to be equal; if the costs of misclassification C 2 given 1 and C 1 given 2 is equal to 1; say if they are same; there is if the ratio is equal to 1, then this expected cost of misclassification minimizing rule is the same rule as that which minimizes the total probability of misclassification.

(Refer Slide Time: 11:56)



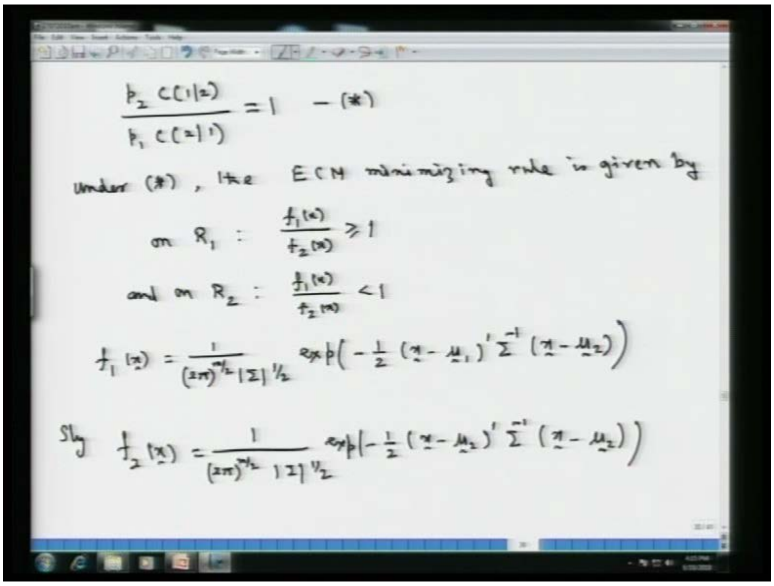
So, just note that, suppose we have got this ratio C 1 given 2 by C 2 given 1 this is equal to 1, then this ECM minimizing rule is same that of TPM minimizing rule then ECM minimizing rule minimizing rule is same as that of the TPM minimizing rule. That is, if this holds that is on R 1, we will have p 1 times f 1(x) that by p 2 times f 2(x); that is greater than equal to 1 and on R 2, we will be having this term to be less than 1. So, this is what we have. So, we have got two types of criterion two types of objectives before us under the general classification problem; that we can either go for a total probability of misclassification minimizing rule or we can look for a rule, which would minimize the expected cost of misclassification.

We have derived both these optimum rules under the two philosophies and say that if such a thing holds, then the two rules are basically equivalent. Now, let us derive the look at the following, say classification rule in case of multivariate normal populations till now up till this particular point, we have not assumed any particular form of the populations. We had just said that we have got two populations pi 1 and pi 2 with some prior probabilities with densities given by f 1(x), when a particular observation is belonging to pi 1 and it is f 2(x), if it is belonging to second population. It is interesting to look at, what is the form of these rules when we have now some specific population like that of a multivariate normal population?

So, classification rules under these optimum strategies under multivariate normal populations under multivariate normal populations. Now, what it looks like? We look at two different cases. In the first case, we look at the two populations as follows. Population 1 is a multivariate normal say m dimensional with a mean vector equal to mu one and a covariance matrix positive definite to be equal to sigma. So, sigma is assumed to be positive definite and the second population; so, this in our notation is pi 1 population. This is the second population; we had earlier denoted that by pi 2. So, this is the second population, what we have? Let us assume that it has got a multivariate normal m dimension also with a mean equal to mu 2 and a covariance matrix same as that of the covariance matrix of the first population.

So, the difference between these two multivariate populations is coming in their mean vector. So, for population number 1, it is mu 1; for population number 2, it is equal to mu 2. Now, under such a situation, if we are trying to look at ECM minimizing rule or TPM minimizing rule, then what is the form of the classification rules that we are going to get? Let us look at first a simple example or a simple setup that consider we consider a special structure of a prior and our cost structure. Suppose we consider a special structure, there is a point why actually we are looking at this particular special structure of prior and misclassification cost costs. That is, the priors are p 1 and p 2 and our cost costs of misclassification are C 1 given 2 and C 2 given 1.

(Refer Slide Time: 16:36)



So, we have this special structure, which is going to tell us the following that p 2 into C 1 given 2 this divided by p 1 into C 2 given 1. Suppose that is equal to 0; so, it is a special case definitely. We assume that these two are in particular; if we have the two prior probabilities to be equal and the two costs misclassification to be equal, then we will naturally be having this particular ratio to be equal to 1. Otherwise also, if we have different priors, prior probability is p 1 and p 2 and different cost structures C 1 given 2 and C 2 given 1, even then we can also have this particular special structure here.

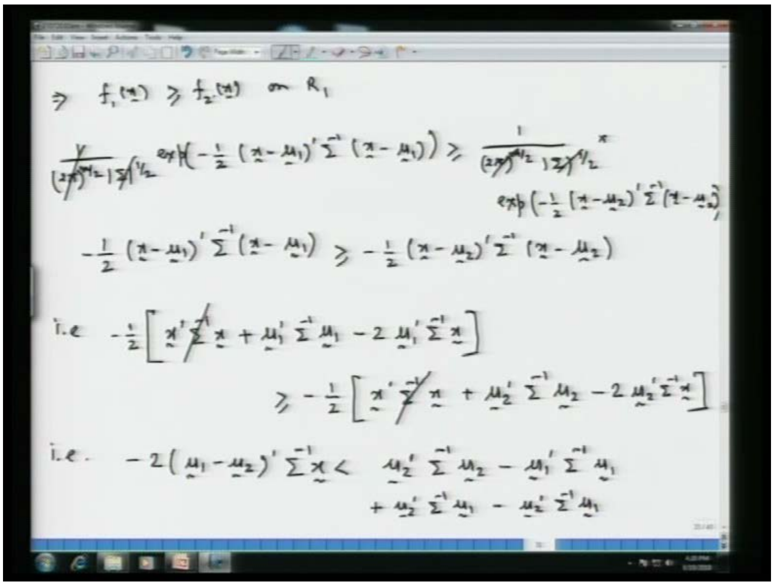
So, let us mark it as star. Now, under star under star the ECM minimizing rule for a general classification problem ECM minimizing rule is given by the simple form; that we have got on R 1, f 1(x) by f 2(x) greater than or equal to 1 and on R 2, the complementary region we will be having f 1(x) by f 2(x) this to be less than 1. From where does it come? It comes straight away (Refer Slide Time: 07:22) from this ECM minimizing rule that we have obtained out here. So, there we had just assumed a special structure in the priors and the costs.

We had assumed that this part here up to the cost part; that is equal to 1. And hence, the ECM minimizing rule is just f 1(x) by f 2(x) greater than or equal to 1 on R 1 and this on R 2, we will be having f 1(x) by f 2(x) to be less than 1. Now, what it is in terms of these populations now? (Refer Slide Time: 11:56) We have got the two populations pi 1 and pi 2 to be these two multivariate normal populations and hence we will be having this f

1(x). This is the density of the multivariate normal; this under that multivariate normal n mu 1 sigma.

So, this is 1 upon 2 pi to the power m by 2 determinant of sigma to the power half and then we have e to the power minus half x minus mu 1 transpose sigma inverse x minus mu 2. And similarly, we have f 2(x) the density of x or the joint density of the elements of that x vector under the pi 2 population to be given by 2 pi to the power m by 2 determinant of sigma to the power half and then we have e to the power minus half x minus mu 2 transpose sigma inverse x minus this mu 2. Then let us look at what this actually leads us to. This leads us to a simple form, a known known form actually; that is what we are going to show.

(Refer Slide Time: 19:54)



So, this f 1(x) greater than or equal to f 2(x); remember that is the region R 2. (Refer Slide Time: 16:36) R 2 is f 1(x) greater than or equal to 1 is that region. So, this is equivalent to f 1 greater than f 2 on R 1. This condition is equivalent to writing the two; 1 upon 2 pi to the power m by 2 determinant of sigma to the power half. Then we have e to the power minus half x minus mu 1 transpose sigma inverse x minus mu 1. This is greater than or equal to the term corresponding to f 2(x); that is 2 pi to the power m by 2 determinant of sigma to the power half. Then we have that product e to the power minus half x minus mu 2 transpose sigma inverse x minus this mu 2.

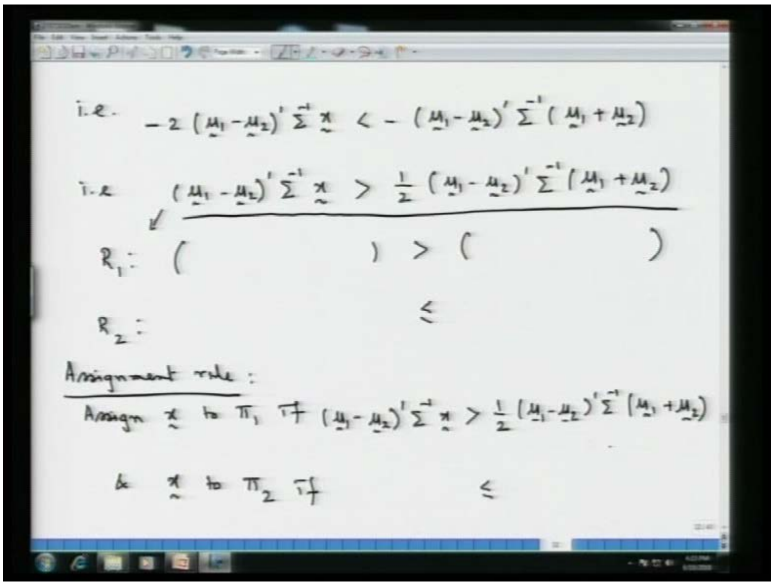
So, these two terms, this term and this term cancel out with these two terms and we simply have the following that it is let us open the two exponents. One can get rid of the exponents also; that is not a problem; because take a log on both the sides. Taking log on both the sides here; what we can see is that this is minus half on this side. So, this is x minus mu 1 transpose sigma inverse x minus mu 1. This is going to be greater than equal to log of this particular term here, which is minus x minus mu 2 transpose sigma inverse x minus mu 2. Let us write the terms out here. It is this minus half also can be observed; because it is common in both the sides.

So, what are the terms that we get here? We get the terms x transpose sigma inverse x. And then we have another positive term, which is mu 1 prime sigma inverse mu 1 and then the cross product term from here; that is minus 2 either in terms of mu prime sigma inverse x or in terms of minus 2 x transpose sigma inverse mu 1. Let us write that as 2 times mu 1 prime sigma inverse x quantity. So, this is what we have on the left hand side. This is greater than or equal to minus half of the similar terms x transpose sigma inverse x this plus mu 2 transpose sigma inverse mu 2 this minus twice mu 2 transpose sigma inverse x.

So, the terms that cancel out from both the sides is this term along with this particular term. So, we can write this expression in a compact way as minus 2 say mu 1 minus mu 2 whole transpose sigma inverse that to be less than with this sign; because this minus sign if we take it out, then we are changing the direction of the inequality. And hence, we can write it as the terms inside the bracket here; minus 2 times mu 1 prime sigma inverse then x this term this plus the term which comes from this side; which is plus twice mu 2 transpose sigma inverse x and this is now less than the term goes on the right hand side.

So, we will just be having this as mu 2 transpose sigma inverse mu 2. Then this term on the other side which is leading us to minus mu 1 transpose sigma inverse mu 1. Now, in this particular expression, let us introduce and subtract the following term. So, let me write it as mu 2 prime sigma inverse mu 1. So, we have this term extra; so, take it out here mu 2 transpose sigma inverse mu 1. The point in writing this in this particular form is that we are basically trying to show that this rule under a special structure in the prior and cost misclassification costs. This is going to lead us to a rule, which is known to us.

(Refer Slide Time: 24:51)

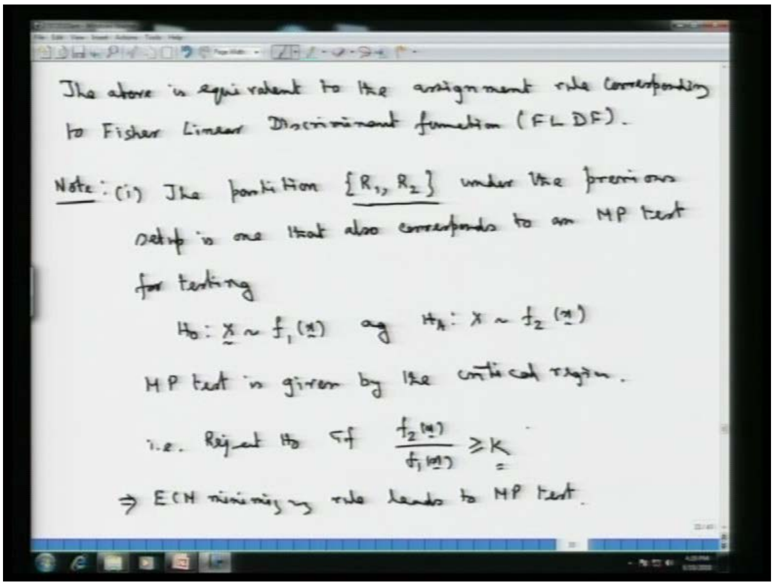


So, this can be written this expression can be written in the following way that we have got this term to be equal to minus 2 times; leave the left hand side as it is. So, this is mu 1 minus mu 2 transpose sigma inverse and x. This is less than minus mu 1 minus mu 2 prime a sigma inverse times mu 1 plus mu 2; that is the term which is coming (Refer Slide Time: 19:54) from this right hand side. So, what we have is this one; that is we have got this mu 1 minus mu 2 prime sigma inverse x. This is greater than half of this mu 1 minus mu 2 prime sigma inverse mu 1 plus mu 2.

So, when this ECM minimizing rule; so, remember that this is the region R 1. So, this is what is corresponding to R 1. So, R 1 is the region on which, we have got this particular term here to be greater than the term, which is on the right. And on R 2, we will be having this is to be less than or equal to… Now, identify that this particular term here that we are saying that on R 1, this is this and hence if x is such that this quantity is greater than this term out here. Then what will be having is x being classified into pi 1 population and if it is other way round, then x is getting classified into the second population; that is pi 2 population.

Now, this is nothing but it is the fisher linear discriminant function. So, the assignment rule assignment rule or allocation rule is the following. Assign x assign the random vector x to pi 1, if we have got this mu 1 minus mu 2 transpose sigma inverse x to be greater than half mu 1 minus mu 2 transpose sigma inverse mu 1 plus this mu 2 term and x to pi 2, if x belongs to R 2; that is, if this is less than or equal to this. Now, this is the same rule as what we had obtained earlier using the fisher linear discriminant function.

(Refer Slide Time: 27:49)



So, we make a note of that; that the above is equivalent to is equivalent to the assignment rule which is based on the fisher linear discriminant function is equivalent to the assignment rule corresponding to fisher linear discriminant function or FLDF in our earlier abbreviated form, FLDF. So, what we have shown for this particular example of the normal distribution is that as the special case of the general classification problem. If we are having two populations to be multivariate normal populations differing by their mean vector; the covariance matrix remaining the same, then the expected cost of misclassification minimizing rule with the special structure of prior probabilities and the costs of misclassification is same as that of the fisher linear discriminant function based classification rule.

Now, let us look at this rule itself and try to say something about some characteristics that emerge out of this particular rule. (Refer Slide Time: 24:51) Now, note that if we have got the rule the assignment rule as this one or (Refer Slide Time: 19:54) to start with we had got this to be our region on R 1. That is, we are going to classify x in to pi 1, if this happens. Now, how does this particular rule corresponds to a most powerful test critical region? It is natural; because we are looking at two different populations. And we are basically trying to build some rule, which is going to solve this particular problem of whether it is coming from pi 1 population or it is coming from the second pi 2 population.

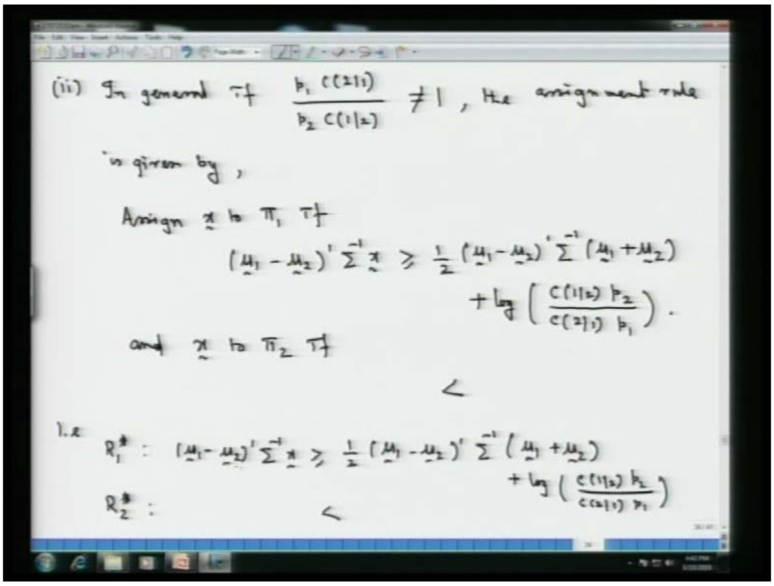
Now, we say that the partition that we have got the partition R 1, R 2 under the previous setup under the previous setup is one that also corresponds to that also corresponds to an MP test for testing the following null hypothesis, H naught; that x, the random vector; this follows a distribution, which has got f 1(x). This is to be tested against an alternate hypothesis say H A that x is following f 2(x). So, in terms of our classification problem, we are saying that x is in pi 1 population. That is, it has got this f 1(x) density and this has got the density, which is there corresponding to the second population.

And hence, if we are looking at the most powerful test for testing this null hypothesis against this alternative hypothesis, then what are we going to do? We are going to use the Nyman Pearson fundamental lemma. So, by the Nyman Pearson fundamental lemma, the most powerful test would be given by the ratio of the density under H A divided by the density under the null hypothesis being greater than or equal to k. So, we will be having the most powerful test is given by the critical region is given by the critical region. That is, reject H naught, if we have f 2(x) by f 1(x); this is greater than or equal to k say and we accepted, if it is otherwise.

Now, what this is going to lead us to this k is going to be such that the size condition is going to be satisfied. Now, what we have in our given problem is that the region R 2 is that f 2(x) by f 1(x) is greater than or equal to 1. So, the classification problem which partitions the sample space into R 1 and R 2, which is corresponding to the expected cost of misclassification minimizing rule is what is giving us a most powerful test at a fixed size. So, this would imply that ECM minimizing rule the ECM minimizing rule leads to most powerful test of a fixed size.

So, we are not having say freedom to choose that particular constant k as is chosen, when one is actually trying to find out the most powerful test at a particular level alpha. So, the alpha level ofcourse, here is not not say suppose is alpha is belonging to 0 1; any value between 0 1. So, we will be choosing k in such a way that the size condition is satisfied. However, the ECM minimizing rule is f 2(x) by f 1(x) greater than or equal to 1 in the region R 2. So, that is the partition corresponding to the ECM rule and which thus is a rule, which corresponds to the MP test only.

(Refer Slide Time: 33: 40)



Now, the second note that we put here is that; we have considered in the previous example a special structure of the prior and (Refer Slide Time: 16:36) this misclassification costs as in specified through this equation number star; that we have got this ratio to be equal to 1. So, it is not always true that we will be having this. So, what happens, if we have got a general costs structure and prior probability is such that the ratio is not equal to 1? So, in general in general, if we have got that this p 1 C 2 given 1 that divided by p 2 into C 1 given 2; if this is not equal to 1, then the previous derivation of the rules still holds (Refer Slide Time: 16:36) with a rider that this term here is going to get multiplied by this particular ratio.

So, if you look back further, that the ECM minimizing rule (Refer Slide Time: 07:22) under the general setup is this particular term here. So, we will be able to write this ratio f 1(x) by f 2(x) to be greater than or equal to p 2 into C 1 given 2 that divided by p 1 into C 2 given 1. But we have already computed (Refer Slide Time: 16:36) what is that particular ratio f 1(x) by f 2(x), which we have reduced (Refer Slide Time: 19:54) in terms of the fisher linear discriminant function in the form this and hence consequently (Refer Slide Time: 24:51) in terms of this. And hence, if we have got now a general situation wherein this ratio is not equal to 1, the assignment rule would be given by the derivation which just would differ from the previous derivation by this constant or log of this particular constant.

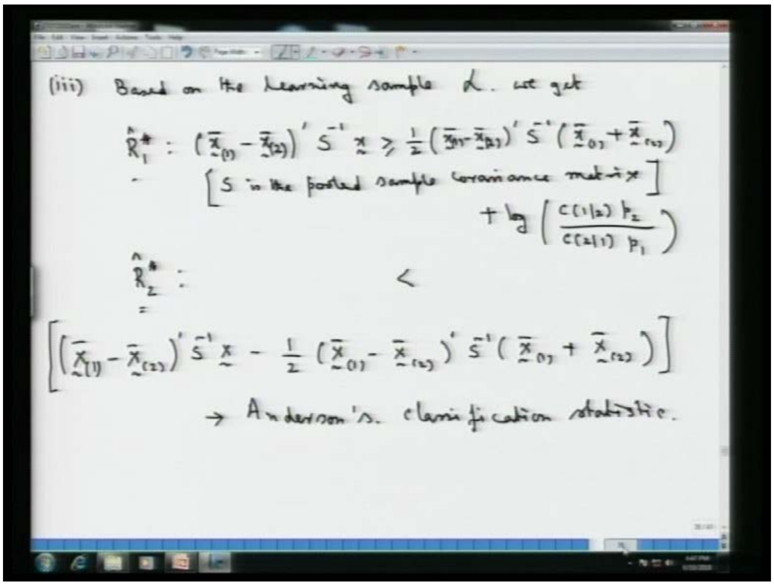
The assignment rule is given by assign x to assign x to pi 1, if we have got this mu 1 minus mu 2 prime; this is coming from the previous expression itself; sigma inverse times x this is greater than or equal to say half times mu 1 minus this mu 2 prime sigma inverse mu 1 plus mu 2 up to this particular term. We had in the previous example, where wherein this ratio was assumed to be equal to 1. Now, there is a constant term log of that term still remains in this particular expression. And hence in the general situation, just this term is added to the previous term C 1 given 2 times this p 2 C 1 given 2 times p 2, this term comes there; that divided by this p 1 or C 2 given 1 times this p 1.

And x to pi 2, if this quantity is less than the right hand side and thus, we have got this R 1 and R 2 region for this that R 1. Let us write that to be R 1 star to distinguish it from the previous R 1. So, this region now is mu 1 minus mu 2 prime a sigma inverse times x this is greater than or equal to... So, it is a region of all x's for which, this left hand side here is greater than or equal to the right hand side; mu 2 prime sigma inverse mu 1 plus mu 2 this plus log of this C 1 given 2 divided by C (( )) 1 times (( )). And R 2 star, the region for assignment to the second population would just be given by this less than this particular term.

So, if we have got the partition that we are we are looking at the assignment rule that; x to pi 1, if mu 1 minus mu 2 prime; this is greater than or equal to right hand side here and x to pi 2, if the left hand side is less than the right hand side here. That is, the partition that we get for a general setup, wherein we have this ratio here to be not equal to 1. That is R 1 star region, which is which is different from the region R 1 that we had got earlier. So, this region is the set of all x's for which this quantity mu 1 minus mu 2 prime sigma inverse x is greater than or equal to half times mu 1 minus mu 2 prime sigma inverse mu 1 plus mu 2 plus log of C 1 given 2 times p 2 divided by C 2 given 1 times p 1.

So, this becomes the region R 1 star; that is the set of all x's for which this happens. So, that is the region R 1 star; that is the assignment region for the pi 1 population and this is R 2 star, which is the left hand side less than the right hand side here. So, we have got this classification rule under the general setup also. Now, note that if we are looking at this region here; this involves quantities, which are usually unknown in the population. That is, mu 1, mu 2, sigma inverse all these quantities are unknown in the population. So, what is done?

(Refer Slide Time: 39:36)



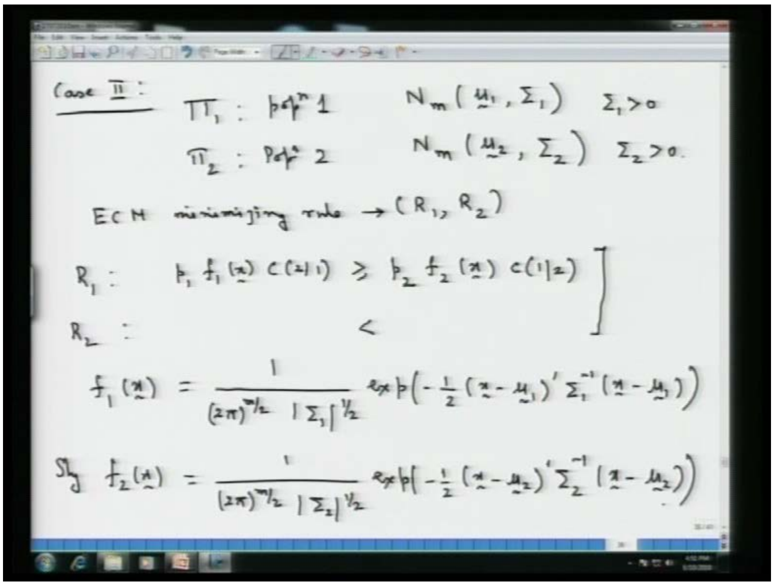
When we have got a learning sample, learning sample consists as we have discussed set of pre-classified examples. So, based on learning sample based on the learning sample, the sample of pre-classified cases say l, we get the estimate of this particular region which is going to be given by… Now the thing that we do here (Refer Slide Time: 33: 40) is to find the estimator of mu 1, mu 2 and sigma inverse. So, mu 1 can be estimated by the sample mean corresponding to the first population. So, let us denote that by x 1 this minus x 2 bar. This is the sample mean vector from based on the observation coming from the second population; so, this transpose times S inverse, where S is the pooled sample variance covariance matrix.

We write that S is the pooled sample variance covariance matrix. (No audio from 40:45 to 40:54) Now, that is going to be given by; suppose we have got n 1 observations from the first population, n 2 observations from the second population, then n 1 plus n 2 minus S is going to be equal to n 1 minus 1 times S 1; that is based on the n 1 observations plus n 2 minus 1 times S 2; the S 2 based on the second population samples of size n 2. So, this term is multiplied by x that is greater than or equal to half times the corresponding estimate (Refer Slide Time: 33: 40) that we get from the corresponding quantities here. So, this is x 1 bar minus x 2 bar; first and second populations; this transpose S inverse, pooled variance covariance matrix; once again this x 1 bar plus x 2 bar.

This is the first term and then the second term would remain as it is, because for any practical purposes these costs of costs of misclassification and the prior probabilities will be assumed to be known. And in R 2 star hat, which is the estimated region here. So, this is what is now is in a implementable form that given a particular x observation; new observations that has now come and we are trying to put it into either of the two populations pi 1 and pi 2. So, these are all quantities, which can be computed directly and hence we look at, what where a where that particular x is lying; whether it is lying on in this region or it is lying in this particular region.

Now, there is a special term that is usually used for the difference of this minus the first term on the right hand side. That is, this x 1 bar in terms of the random vectors x 2 bar this transpose S inverse x this minus half x 1 bar minus x 2 bar; that is the first term; that is there in the right hand side x 1 bar plus x 2 bar. This is called the Anderson statistic. So, this is called the Anderson's classification statistic. (No audio from 43:39 to 43:50)This is just a mean as such that this particular term here, which involves the random variables, is called this Anderson's classification statistic. Now, we had in the first case looked at the situation, where the two multivariate normal populations had got the same covariance matrix only differing by the mean vector in the two populations.

(Refer Slide Time: 44:14)

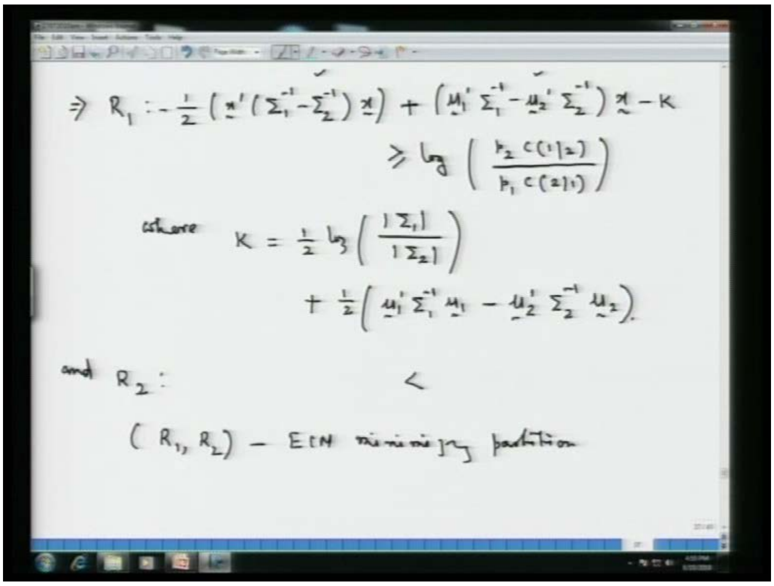


Now, let us look at a more general setup, wherein we look at the two populations of the following form that we have got this pi 1 population, which is the first population. Population number 1, which is a multivariate normal population with a mean vector as mu 1 vector and a covariance matrix as sigma 1; sigma 1 is assumed to be positive definite. And pi 2, the second population is also a multivariate normal population with a mean vector as mu 2 and a covariance matrix as sigma 2, where sigma 2 is also assumed to be positive definite matrix. So, suppose we have this particular setup, now the way that this case differs from the previous case is that we have got two different positive definite covariance matrices of the two corresponding populations.

Now, we are trying to see that what is our ECM minimizing rule. Say ECM minimizing rule under this setup ECM minimizing rule is given by the partition say R 1, R 2. Then we have got in the general setup without assuming anything on the cost structure. This R 1 is the region now of set of all x's such that we will be having p 1 times f 1(x) this into C 2 given 1. This is greater than or equal to the corresponding terms for the second population; that is, p 2 f 2(x) into C 1 given 2 and in R 2, we will have this to be less than this particular term. Now, since we have a continuous distribution, does not matter which side actually we look at this equality.

So, we have got this as the classification rule. Now for the given problem, we can say that this f 1(x) now; its term similar to what we had got earlier with only the difference that the sigma matrix is going to be different for the two populations. And hence, the density is going to be different here as well which was same earlier, because we had got the sigma matrix to be same in both terms there. So, x minus mu 1 transpose sigma 1 inverse x minus this mu 1 vector. And similarly, this f 2(x) is our 1 upon 2 pi to the power m by 2 determinant of sigma 2 to the power half e to the power minus half x minus mu 2 transpose sigma 2 inverse x minus mu 2. So, once again we look at what this region leads us to.

(Refer Slide Time: 47:20)



So, this would imply after simplification; that our R 1 region by plugging in here. (Refer Slide Time: 44:14) The value of f 1(x) as is given by this expression out here; f 2(x) as is given by this expression here. So, we will use those expressions and finally, what will be getting is the following term that; on R 1, we will be having the following expression that it is equal to minus half times x transpose sigma 1 inverse minus sigma 2 inverse. Now, this does not cancel out; because we have got two different sigmas at the moment in the previous example, when we had sigma same. So, this quadratic term was not present; because it was cancelling out.

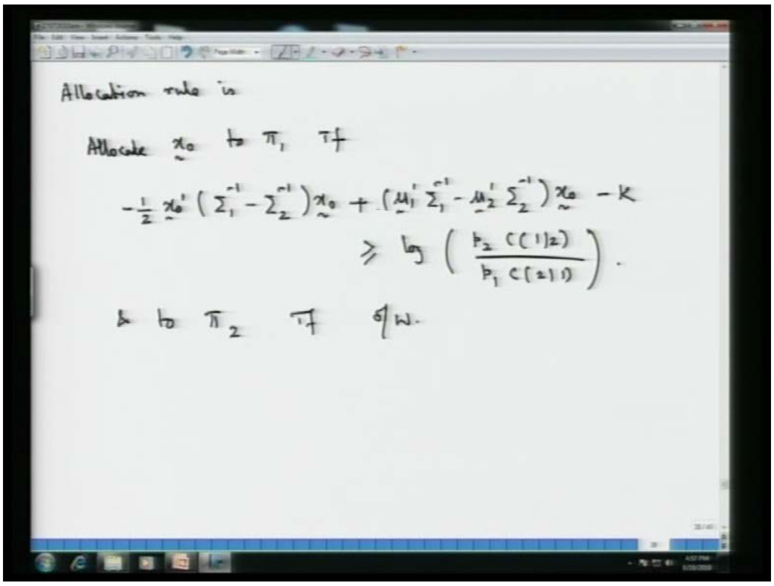
So, what we have is this term this plus after simplification this reduces to mu 1 prime sigma 1 inverse minus mu 2 prime sigma 2 inverse this times x. This is the linear term. In the previous example, we had sigma 1 to be equal to sigma 2 and hence, the term that we had there was mu 1 minus mu 2 prime sigma inverse times this vector x. Now, we are unable to do that type of simplification; because sigma 1 and sigma 2 here are different. So, we have got a quadratic term here; quadratic in x's, we have got a linear term here. Similar to the term we had previously, this minus I say a constant, k; I will say what it is.

This is greater than or equal to log of the term, which we also had earlier p 2 into C 1 given 2 that divided by p 1 into C 2 given 1, where this constant k this constant k now would be having terms, which are involving a mu 1, mu 2 and sigma 1 inverse and sigma 2 inverse from the two terms, which are coming from (Refer Slide Time: 44:14) these two density. That is, if you look at this mu 1 transpose sigma 1 inverse mu 1 and the term here mu 2 transpose sigma 2 inverse mu 2, those two are the terms that is going to come here.

So, this k term and also (Refer Slide Time: 44:14) we will be having the determinant terms here; log of that term to come, because they do not cancel out for this present setup. So, what will be having is this constant k is given by half log of determinant of sigma 1 that divided by determinant of sigma 2 this term plus the term which was there in the exponent. So, that is half of mu 1 transpose sigma 1 inverse mu 1 this minus mu 2 transpose sigma 2 inverse times mu 2. So, this term is these two terms are coming from the exponent of the density; these two terms coming from the denominator as in here.

And if we have got this to be the R 1 region, then R 2 region would just be given by that the left hand side here is going to be less than this term here on the right hand side. So, this is what is the ECM minimizing classification rule on the partition of the sample space into the two regions R 1 and R 2, wherein we have got two multivariate normal populations, which have got different mean vectors as well as different positive definite covariance matrices. Now, corresponding to this particular partition, what we have this R 1, R 2; ECM minimizing partition this ECM ECM minimizing partition.

(Refer Slide Time: 51:47)



We can say that the assignment rule is the following. Assignment or allocation rule, assign suppose we have got x naught to be a new observation, allocation rule is that allocate x naught a new observation to pi 1, if the corresponding quantity as what we have got there. That is, minus half x naught prime sigma 1 inverse minus sigma 2 inverse times x naught; this is the quadratic term here; this plus that mu 1 transpose sigma 1 inverse minus mu 2 transpose sigma 2 inverse this times x naught vector this minus that k constant is greater than or equal to the term with the prior probabilities and the costs of misclassification C 1 given 2 this divided by p 1 into C 2 given 1 and to pi 2, if it is otherwise.

Now, once again you see that this particular expression here; what we have got involves quantities like mu 1, mu 2, sigma 1, sigma 2. One would require to replace those by the corresponding sample (Refer Slide Time: 47:20) estimates and also just to tell or just to note that this term involves quadratic term, and hence such a discriminant function is called a quadratic discriminant function. So, this term here, since it involves a quadratic discriminant function here, it is called the quadratic discriminant function. So, we stop today's lecture at this particular point, then the next lecture, what we are going to look at is some criterion on which a classification rule can be based on or to look at criterion that would judge how good a particular classification rule is and also we look at multiclass problems. Thank you.