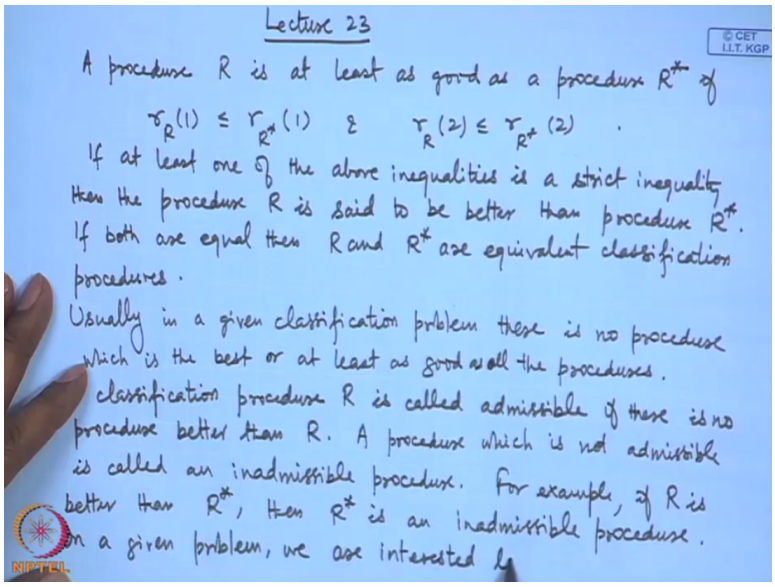
Statistical Methods for Scientists and Engineers
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology – Kharagpur

Lecture – 23
Multivariate Analysis – VIII

In the last class, I have introduced the problem of classification of the observations. In the problem of classification of the observations the classical problem is that we are given 2 populations pi1 and pi2 and it is required for us to find out whether a new observation that is given to us whether it belong to the first population or to the second population. To derive a classification procedure, we need say training samples from both the populations and we use that for constructing the proper classification procedures.

In the last lecture, I have introduced the expected loss of observation, expected loss if the observation is wrongly classified. Suppose it is belonging to the first population and it is classified into the second or it is in the second and it is classified into the first. Based on that we find what is the Bayes procedure and I had also described rR(1) and rR(2) that is the expected loss of observation if it is from pi1 and it is, if it is from pi2.

(Refer Slide Time: 01:40)



So we say that, a procedure R is at least as good as a procedure R* if rR(1) is <= rR*(1) and rR(2) is <= rR*(2). That means both the expected losses if the observation is from pi1 and if the observation is from pi2 should be smaller, then the procedure R is better than the

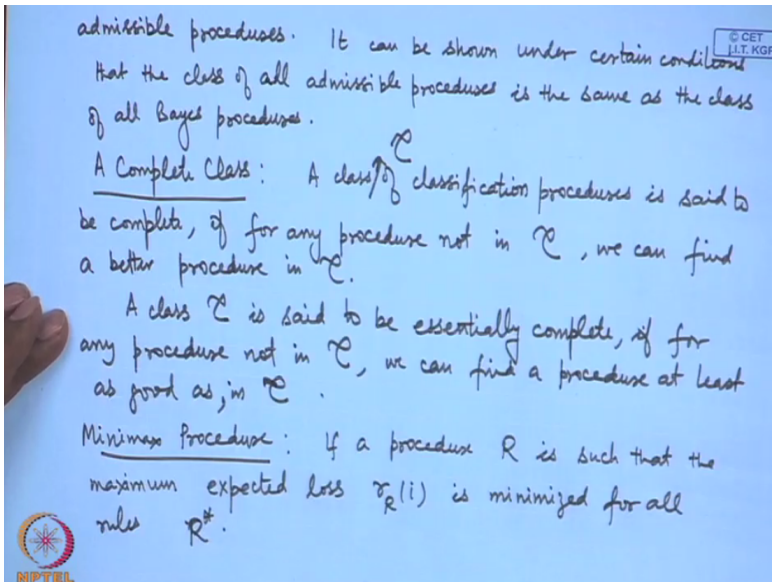
procedure R*. If at least one of the above inequalities is a strict inequality then the rule, the procedure R is said to be better than R*.

So when we say <= it includes the case of equality also. So when both of them are equal then R and R* are said to be equivalent. If both of are equal, then R and R* are equivalent classification procedures. So 1 may question that whether there is a procedure, which will be, better than all the given procedures that means you consider is it the best procedure.

The answer is that no. So usually in a given classification problem there is no procedure, which is the best, or at least as good as all the procedures. Now this can be explained like this. Suppose I make this procedure as the best that means there is no chance of error in procedure R, then this will become 0 here but at the same time this will become 1 and in that case this cannot be <= this 1. So that means there is a comparison, so sometimes this will be better sometimes the other one will be better.

So we say an admissible rule. A classification procedure R is called admissible if there is no procedure better than R. A procedure, which is not admissible, is called an inadmissible procedure. For example, in this case I have mentioned that if there is a strict inequality in this statement then the procedure R is better than R*. So in that case R* will become inadmissible procedure. For example, if R is better than R* then R* is an inadmissible procedure. So in a given problem we are interested to characterize all admissible procedures.

(Refer Slide Time: 07:10)



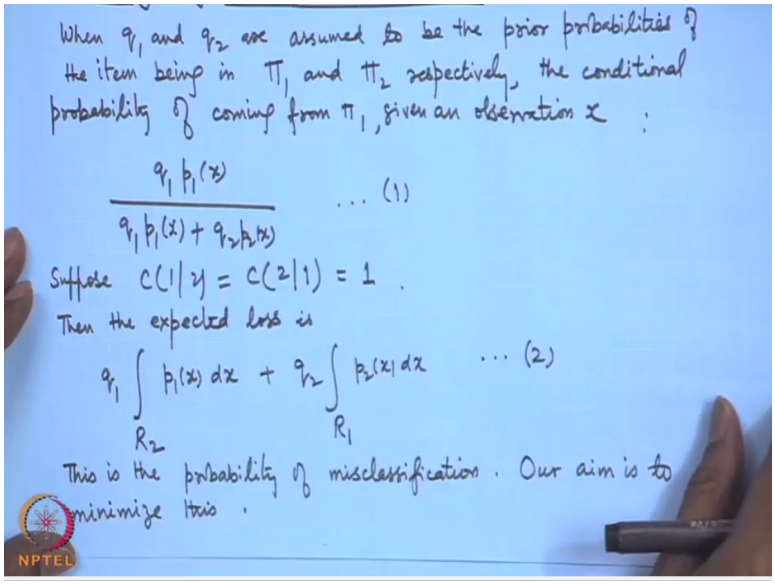
That means all the procedures which are admissible we should be able to characterize. So in that case we can restrict our attention to only the admissible procedures. Inadmissible procedures we can discard. It can be shown under certain conditions that the class of all admissible procedures is the same as the class of all Bayes procedures. Then we consider a complete class of procedures.

A class of classification procedures is said to be complete if for any procedure, so let us name this class as C, if any procedure not in C, we can find a better procedure in C. A class C is said to be essentially complete if for any procedure not in C, we can find a procedure at least as good as in, we can find a procedure, which is at least as good as in the class C. So that means in a given classification problem we would be interested to find out or to characterize what is the complete class of classification procedures.

So, another thing is that there may be some procedures, which are equivalent, but the procedures themselves may be same except on a set of major 0 or in a set of probability 0 then they will be certainly equivalent, so we can always consider that. Then we consider a Minimax procedure. So if a procedure R is such that the maximum expected loss that is rR(i) is minimized for all rules R*.

So basically what we are seeing that for any classification procedure let us consider the expected loss. So they will be like rR(1), rR(2), etc. so we look at the maximum of these values. Now that procedure for which this is actually the minimum that is consider to be the Minimax procedure. Now we discuss the methods of determining these Bayesian procedures and the Minimax procedure etc. in a given classification procedure how do we start with.

(Refer Slide Time: 12.05)



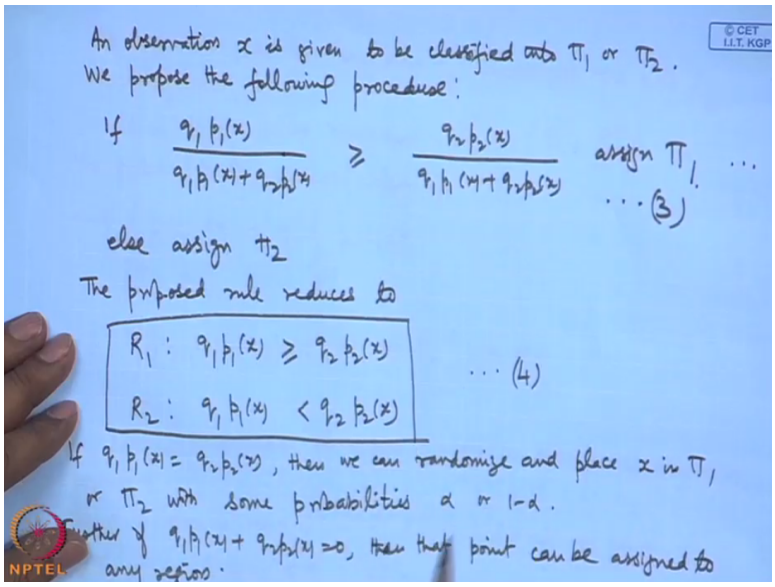
So we firstly consider finding Bayes classification procedures. So when q1 and q2 are assumed to be the prior probabilities of the item being in pi1 and pi2, then we consider the conditional probability of coming from pi1 given an observation x. Given an observation x what is the conditional probability that it came from pi1: So this will be equal to q1 p1 (x) divided by q1 p1(x) + q2 p2(x).

So let us look at the denominator denotes an observation. So then, an observation can from first one and then we write the density or the probability of that first one. Then it could be from the second one then the probability is q2 and this is the density of the second one. Now if we find the conditional probability that it actually came from the pi1 then we write in the numerator.

So this is actually just an application of the Bayes theorem. For the time being, let us consider the cost functions to be equal. So we can assume that c (1|2) that is misclassification 1 when it is from 2 and c (2|1) that is misclassification 2 when it is from 1. Let us assume it to be 1. Now then we consider the expected losses q1 p1(x) dx and R2 that means the observation is from q1 what we identify as into R2 + q2 p2(x) dx R1.

So this is actually the probability of misclassification. So, our aim is to minimize that. Our aim is to minimize this. Now if you look at this the conditional probability of coming from pi1 and similarly the conditional probability of coming from pi2 there it will become q2 p2. So if you compare that and take the higher one then this can be considered as a reasonable classification rule.

(Refer Slide Time: 15:54)

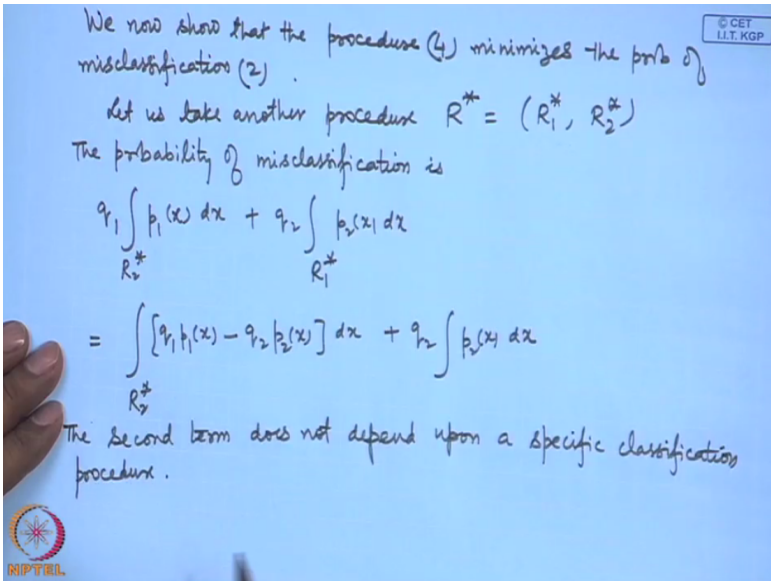


Okay, so let us start with an observation x is given and we want to classify this into pi1 or pi2. We propose the following procedure. If q1 p1(x) / q1 p1(x) + q2 p2(x) is >= q2 p2(x) / q1 p1(x) + q2 p2(x) then assign pi1 else assign pi2. Let me call it (3). So now you can easily see that this is simply equivalent to because the denominator is common. So therefore, the proposed classification rule reduces to.

That is R1 region is that region of classification into the population pi1 which is q1 p1(x) is >= q2 p2(x) and R2 region is the reverse of it, q1 p1(x) < q2 p2(x). Here one point to be mentioned here. I am taking >= here and < here. One can put here > and here <=. So in the continuous distribution models this will not create any problem. In case of discrete, there may be a situation where the probability of equality is positive.

In that case you can further randomize that means you can randomize the classification rule. That means when equality is there certain probability you assign to pi1 and certain probability you assign to pi2. So I am just mentioning this point here. If q1 p1(x) = q2 p2(x) then we can randomize and place x in pi1 or pi2 with some probabilities alpha oR1 – alpha and of course there is another possibility q1 p1(x) + q2 p2(x) = 0, then that point can be assigned to any region.

(Refer Slide Time: 19:32)

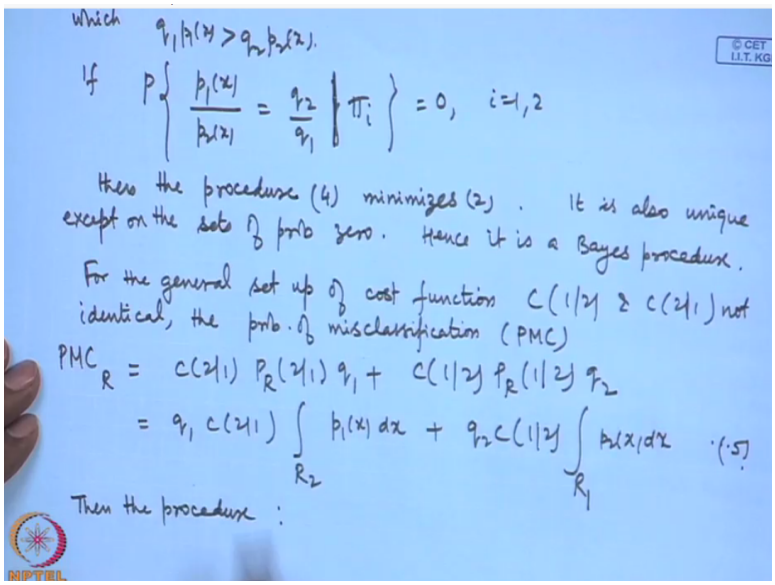


Now the question comes that we actually define the probability of misclassification as (()) (19:39) and our aim is to actually minimize this. Now we have proposed a rule here. We will show that this rule is the rule, which we have written here. This rule will actually minimize this probability of misclassification. We now show that the procedure (4) minimizes the probability of misclassification (2).

So let us take another procedure say R*. So that is (R1 *, R2 *) that means R1* is the region where the observation is classified into pi1 and R2 * is the region in which the observation is classified into R2 in pi2. So for this rule the probability of misclassification is q1 p1(x) dx R2 * + q2 p2(x) dx R1 * which we can write as [q1 p1(x) – q2 p2(x)] dx R2 * + q2 p2(x) dx. Now if you look at the second one this term has no R1 R2 coming here that means whatever be the procedure this term will be the same.

That means we have to look at the first term only. The second term does not depend upon a specific classification procedure. Now the first term if you look at this will be minimized if R2 * includes all the points x for which q1 p1(x) < q2 p2(x) and excludes those points for which q1 p1(x) > q2 p2(x).

(Refer Slide Time: 23:14)

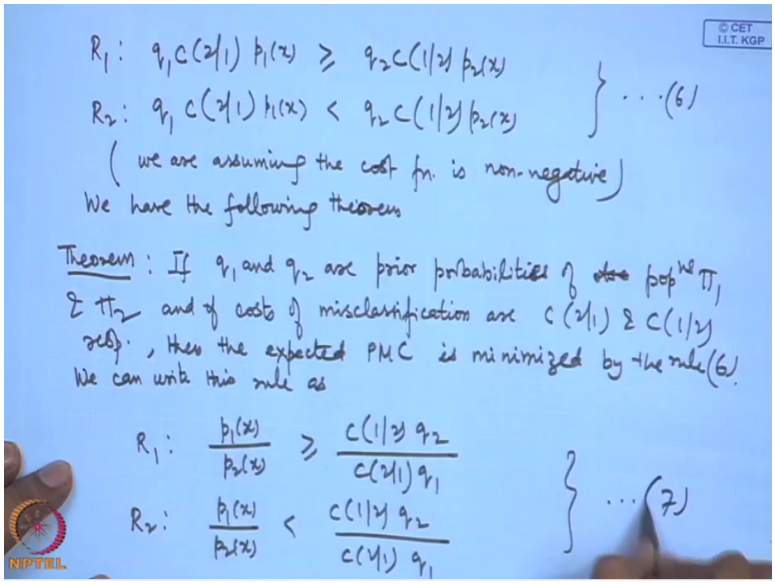


Now if you consider the probability of {p1(x) / p2(x) = q2 / q1 | pi i} = 0, for i = 1, 2 then the procedure (4). If you look at the procedure (4), in the procedure (4), we are considering exactly those regions where q1 p 1 >q2 p2 and we are excluding those which are having q1 p 1 < q2 p2. So whatever statement I gave here the first term is minimized of R2 * includes all the points x in which this is less and excludes those points for which it is greater, then the procedure (4) is exactly satisfying that condition.

Therefore, this procedure (4) minimizes (2). It is also unique except on the sets of probability 0. Hence, it is a Bayes procedure. So we can easily that we have been able to determine if the prior probabilities of the population pi1 and pi2 are given as q1 and q2 then the procedure that is given here, this actually a Bayes procedure. Of course, here we assumed the costs to be equal and that is why we put cancel on both the places. If the cost factor is given, then that will also be included. Let me just give an expression for that.

For the general set up of cost function that is when we have c (1 | 2) and c (2 | 1) not identical, the probability of misclassification we will write it as (PMC) probability of misclassification. That will become PMC R = c (2 | 1) PR (2 | 1) q1 + c(1 | 2) PR (1 | 2) q2 which is actually = q1 c(2 | 1) integral p1(x) dx R2 + q2 c(1 | 2) integral p2(x) dx R1. So in place of this procedure, now we will put cost factor also here and then the procedure that we will proposing can be written like this.

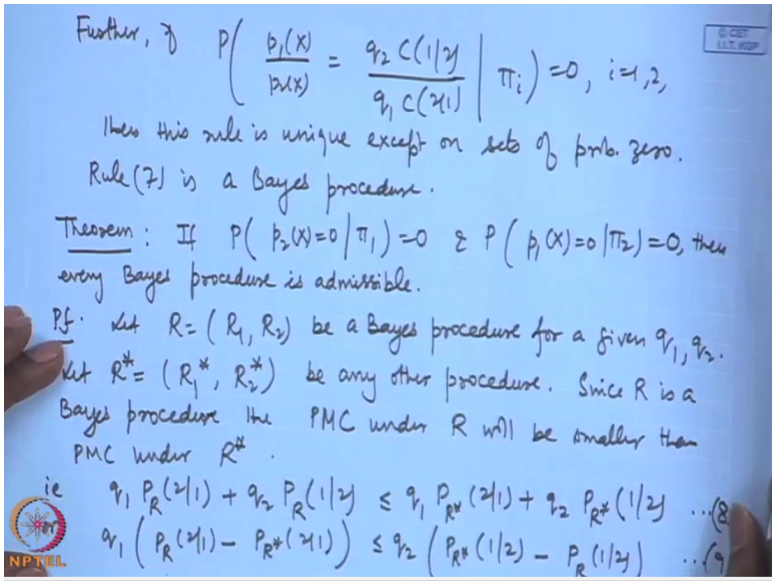
(Refer Slide Time: 27:13)



Then the procedure we consider R1: q1 c (2 | 1) p1(x) >= q2 c (1 | 2) p2(x) and R2: q1 c (2 | 1) p1(x) < q2 c (1 | 2) p2(x). When we write this one, assumption has to be made because we consider cost function. See in case of some gains this could be negative also. So we are assuming here that the cost function is non-negative. Otherwise, the inequalities will get modified. So we have the following theorem.

If q1 and q2 are prior probabilities of observations, of populations pi1 and pi2, that means observation is coming from whether pi1 or pi2 if q1 and q2 are the initial assigned probabilities and if costs of misclassification are c(2 | 1) and c(1 | 2) respectively, then the expected probability of misclassification is minimized by the rule (6). In fact, we can actually write this rule as, R1 you can write in terms of ratios p 1(x) / p2(x) is >= c (1 | 2) q2 / c (2 | 1) q1 and R2 is reverse of it. That is p1(x) / p2(x) < c (1 | 2) q2 / c (2 | 1) q1.

(Refer Slide Time: 30:26)



Further, if the probability of the equality p1(x) / p2(x) = q2 c (1 | 2) / q1 c (2 | 1) under both the populations if this is 0, then this rule is unique except on sets of probability 0. So this rule (7) is a Bayes procedure. So you can see we have actually solved a problem here. Now many times it happens in the classification situation usually the size of the populations will be known.

For example, if we want to classify say land, water we want to classify as a very good student or a mediocre student. So we may know the proportion of the size of the populations. If we know that, then basically we can assign q1 and q2 and in that case the best procedures you need is the Bayesian procedure and we are actually aware because it is minimizing the expected probability of the Bayes classification.

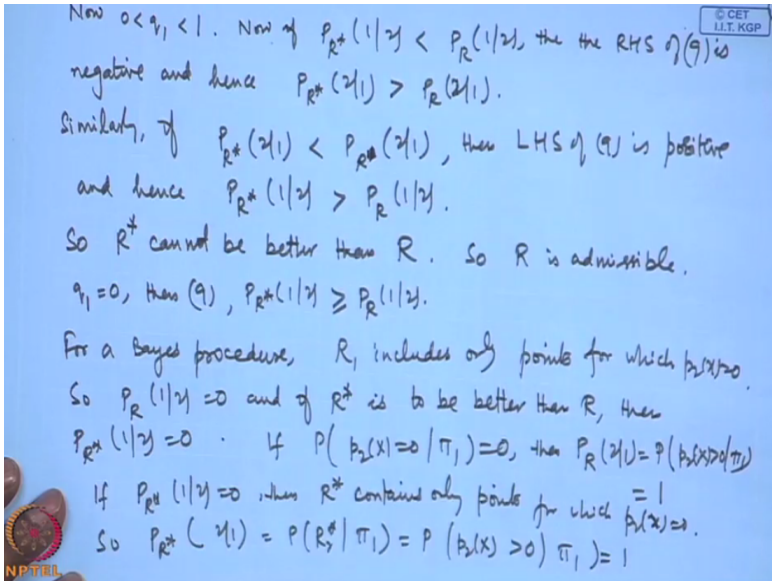
So you are actually having the best procedure. So this is actually you can see the first case and here we are assuming the probability distributions are completely known. That means p1(x) and p2(x) is known to us and in that cases we are actually able to get the best procedures. Now let us consider another thing. Now it could be that initial probabilities are not known. In that case, we can consider the general procedures.

So another thing could be that we may assign various probabilities that in place of q1 q2 suppose it is q1 * q2 *. That means it is not necessary that we have to strictly fix the prior probabilities and in that case the Bayes rule will change. Now, the question comes that whether if your initial assignment is not correct and you get another rule then is it alright or not?

The answer is that even then we are reasonably all right because all the Bayes rules will be admissible. This is proved in the following theorem. If the probability of p2(x) = 0 given pi1 =0 and probability of p1(x) = 0 given pi2 = 0 then every Bayes procedure is admissible. Let R = (R1, R2) be a Bayes procedure for a given q1, q2. Let R* = (R1 *, R2 *) be any other procedure.

Now we are assuming that R is a Bayes procedure. So using that we will have since R is a Bayes procedure the probability of misclassification under R will be smaller than PMC under R*. So you can consider say here q1 PR (2 | 1) + q2 PR (1 | 2) that is <= q1 PR* (2 | 1) + q2 PR* (1 | 2). This we can further simplify; we can write as q1(PR (2 | 1) – PR* (2 | 1)) is < q2(PR* (1 | 2) – PR (1 | 2)).

(Refer Slide Time: 36:18)



Now this q1 and q2 see these are the assigned probabilities so they are between 0 and 1. So we can make use of that. Now q1 is between 0 and 1. So if you are having this PR* (1 | 2) < PR (1 | 2). See what we want to prove that this R* cannot have both the components less. So suppose this is less, then this will become negative. If this is becoming negative, then what will happen that p R (2|1) will become < PR* (2|1).

That means if PR* (1|2) is less, then PR* (2|1) will become more. On the other hand, if you consider PR* (2|1) is less than this then this will become positive. If this is positive, then this is positive and then you will have PR (1|2) less than this. That means both the components

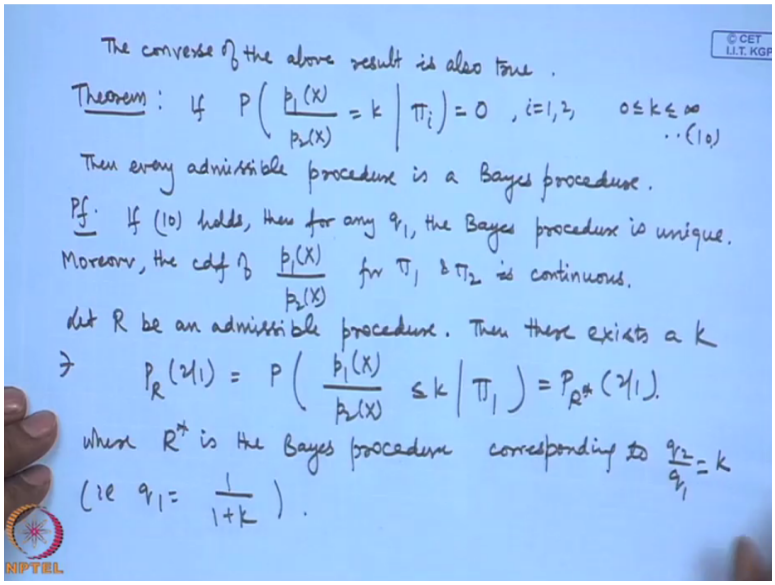
corresponding to R* of the probabilities of misclassification cannot be smaller than the corresponding components of probabilities of misclassification for the procedure R.

So let me just write it formally. Now if PR* (1 | 2) < PR (1 | 2) then the right hand side of (9) is negative and hence PR* (2 | 1) will become > PR (2 | 1). Similarly, if PR* (2 | 1) < PR* (2 | 1) then the left hand side of (9) is positive and hence PR* (1 | 2) will become > PR (1 | 2). So R* cannot be better than R. It cannot be better than R and so this proves that R is admissible.

Now you can consider the extreme case for example q1 = 0. If you take q1 = 0 then what it will give? That (9) will give you simply that PR* (1 | 2) is > PR (1 | 2). Now for a Bayes procedure what is happening? R1 includes only points for which p2(x) = 0. So PR (1 | 2) this will become 0 and if R* is to be better than R, then the only possibility is that PR* (1 | 2) = 0.

So if probability of p2(x) = 0 given pi1 = 0 then PR (2 | 1) = P (p2(x) > 0 | pi1) that is equal to 1. On the other hand, if PR* (1 | 2) = 0 then R1 * contains only points for which p2(x) = 0. So PR* (2 | 1) = P(R2 *| pi1) = P (p2(x) > 0) pi1) that is = 1.So we have proved that R* is not better than R. So here we took an arbitrary procedure R* and we are showing that the procedure R* cannot be better than R.

(Refer Slide Time: 41:53)

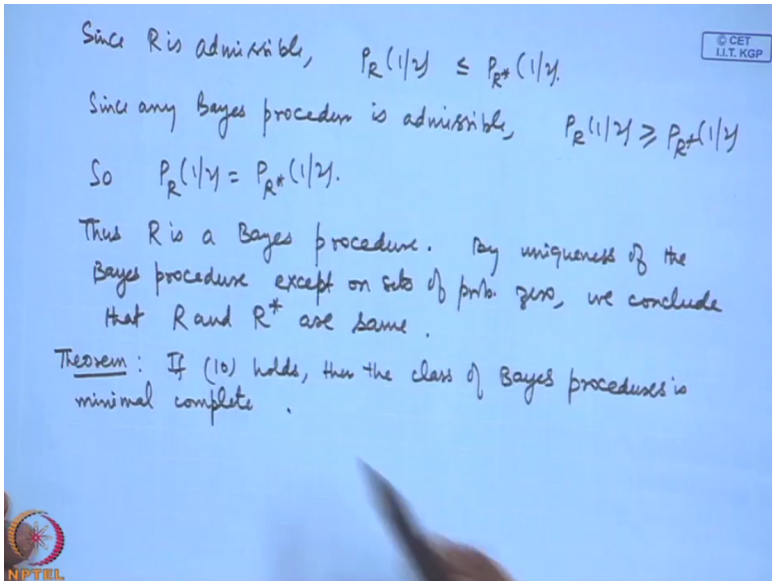


Now the reverse of this is also true. Under certain condition, every admissible procedure will be a Bayes procedure. We prove the following theorem. If probability of p1(x) / p2(x) = k under pi i = 0, for i = 1, 2 for any k between 0 to infinity then every admissible procedure is a Bayes procedure. I mentioned that characterization of the class of the admissible procedures

and I mentioned that it could be shown under certain condition that the class of Bayes procedure is same as that of the class of admissible procedures.

So the previous theorem and this theorem taken together they prove this statement. Let me give a proof of this. See if (10) holds then for any q1, the Bayes procedure is unique. Moreover, the cdf of p1(x) / p2(x) for pi1 and pi2, this is continuous. Let R be an admissible procedure. Then there exists a k such that PR (2 | 1) = the probability of p1(x) / p2(x) is <= k under pi1 = PR* (2 | 1) where R* is the Bayes procedure corresponding to q2 / q1 = k. That is actually you are saying that q1 = 1 / 1 + k and q2 = k / 1 + k.

(Refer Slide Time: 45:03)



Now, since R is admissible, therefore we must have PR (1 | 2) to be <= PR* (1 | 2). Since any Bayes procedure is admissible, we will have the reverse that means PR (1 | 2) has to be > PR* (1 | 2). So basically, it means that they are same. So what we have proved? We started with a procedure R, which is an admissible procedure, and R* is the Bayes procedure.

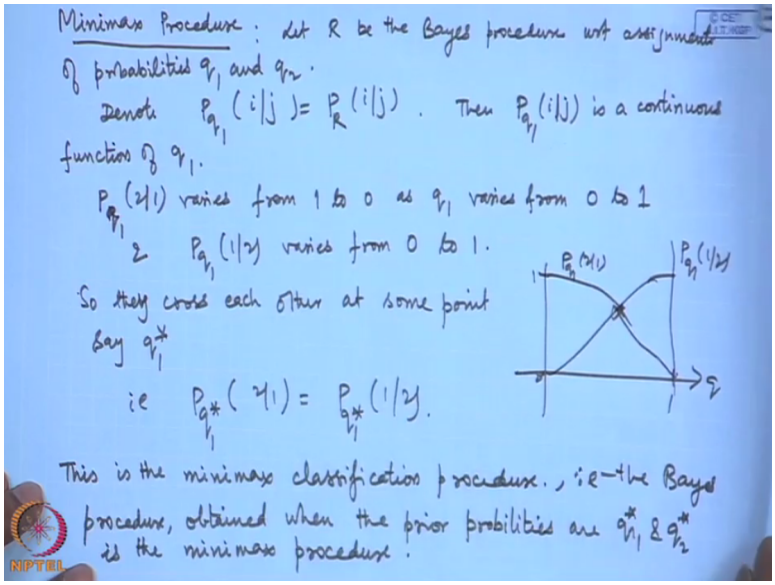
So what we did that we consider the Bayes procedure with respect to that for that this is the probability because p1(x) <=. So there is a k, which is appearing in the form of the Bayes procedure if you remember. We wrote it at here. The form of the Bayes procedure had this term here. Let me give it again. You can look at this q1 p1(x) >= q2 p2(x). So you are having p 1 / p2 >= q2 / q1 and p 1 / p2 < q2 / q1.

So if you combine these 2 then this is becoming a Bayes procedure. Thus, R is a Bayes procedure. By uniqueness of the Bayes procedure except on sets of probability 0, we

conclude that R and R* are same. So we have proved very significant result. That is all the admissible procedures are basically the class Bayes procedures.

The class of Bayes procedures is the class of admissible procedures. So in a given classification problem we can restrict attention to the class of Bayes procedures. If this statement number (10) holds, then the class of Bayes procedures is minimal complete. So this is a very powerful result because it allows you to restrict attention essentially to only Bayes procedures. Now let us also consider some discussion on the minimaxity.

(Refer Slide Time: 48:30)



Remember that the minimaxity criterion is based on a different philosophy. We are considering the worst possibility that means worst-case scenario that means the probability of misclassification is the worst and then among the worst we are choosing the best. In the Bayesian procedure, we are considering only the average loss or average probability of classification where as in the Minimax procedure we are considering the individual.

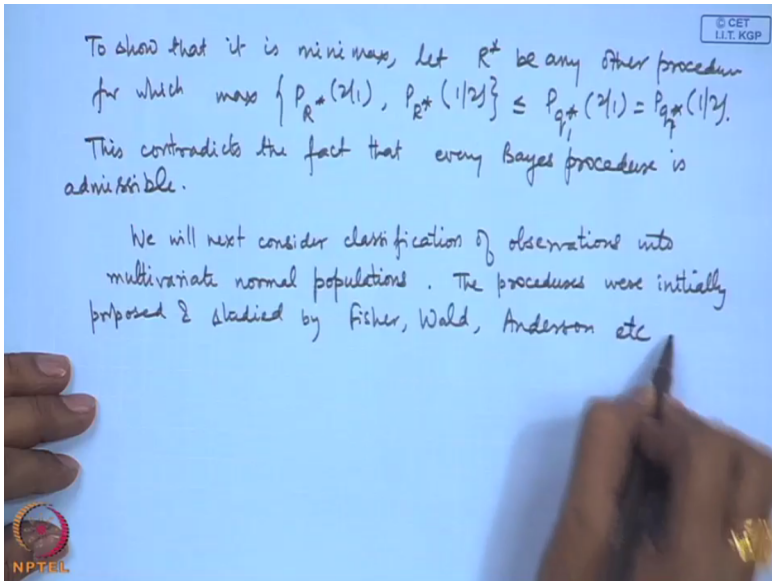
But then we are looking at the worst that can happen and then we are choosing that procedure for which is actually the best. So let R be the Bayes procedure with respect to assignment of probabilities q1 and q2. Let us denote p q1 (i | j) = p R (i | j). That means when q1, q2 is there then the procedure (4) I have written, then that procedure is considered the Bayes procedure and under that probability of misclassification denoting by q1 just to denote that.

So with q1 changing like q1 is half or q1 = 1/4 etc., then this is a continuous function of q1. Actually you can say that p q1 (2 | 1) this will vary from 1 to 0 as q1 varies from 0 to 1 and p

q1 (1 | 2) varies from 0 to 1. So you can see that they are continuous functions and they are varying between 0 to 1 as q1 varies from 0 to 1.

So certainly, they will cross at some point. If I am considering the graph of p q1 (2 | 1) and this is the graph of say p q1 (1 | 2) between 0 and 1 then they will cross at each other. So they cross each other at some point say q1 *. That is p q1* (2 | 1) = p q1 * (1 | 2). This is the Minimax classification procedure that is the Bayes procedure obtained when the prior probabilities are q1 and q2, q1 * and q2 * is the Minimax procedure.

(Refer Slide Time: 53:10)



To show that this is Minimax, let R* be any other procedure for which maximum of {PR* (2| 1), PR* (1|2)} is <= p q1 * (2|1) = p q1 * (1|2). Now if you say this that maximum of this is <= this then both the components have to be <= this. But this would imply for R* procedure the expected probability will be less than the expected probability of misclassification when the rule is assigned by q1 * that is the Bayes procedure.

So this will contradict that this is the Bayes procedure. Therefore, this cannot be true. This contradicts the fact that every Bayes procedure is admissible. So friends today we have considered the basic problem of classification. I have given a very (()) (54:54) consideration of this problem. We considered the costs of misclassification in terms of the probabilities of misclassification p2 and p 1 2.

That means p 1 2 is the probability of classifying 1 when actually the observation is coming from p 1 and similarly p2 1 is the probability of misclassification into the population 2 when

it is actually coming from 1. Now on the basis of this we have considered 2 criteria. 1 is the Bayesian criteria if somehow, we are convinced about the proportions of the observations from the 2 populations say q1 and q2 then based on that we can actually find out the rule, which will minimize the expected probability of misclassification.

So this rule is called the Bayes rule and exact form is obtained here in terms of q1 p1(x) > q2 p2(x) and vice versa that is less as the regions of classification into pi1 and pi2 respectively. 1 can add an additional cost factor also in terms of c (1|2) and c (2|1) and then also the Bayesian procedure is obtained. We also looked at the desirability of the Bayesian rules in terms of the complete class.

For example, we could prove that for every Bayes rule is admissible and every admissible rule is Bayes under certain condition and therefore the class of admissible rule is same as the class of Bayes procedures and therefore the class of Bayes rule is the minimal complete class. So in practice this helps us because if we consider any prior assignment of the probabilities we are doing all right. That means reasonably good rules are available to us.

In fact, whatever rule we propose we will not be able to find the better rule than that, of course we can find the other rules. Second thing is that in the same class we actually determine a Minimax rule also because we can look at, we can vary continuously this q1 and c where probabilities of (2 | 1) and p q1 * (1 | 2) they match. So the point where they match the Bayes- rule at that point will actually give you the Minimax procedure.

So in classical (()) (57:18) formulation we have the solution of this problem. Now in the following classes I will look at the classification procedures for the normal populations rather multivariate normal populations. So the original formulation is by Abraham Wald, 1940s and then we look at the procedures, which are discussed in by Fisher that means when the parameters of populations are unknown. So he estimates that so we consider the classical fisher discrimination.

We consider the (()) (57:54) distance and then we consider the Anderson's classifications, etc. So these are the things that I will be following up in the next classes. That means we will next consider classification of observations into multivariate normal populations. They were initially proposed and studied by Fisher, Wald, Anderson, etc. So we will discuss the

properties of this procedures and how the procedures are actually obtained. So this I will be covering in the next lecture.