Applied Multivariate Analysis

Prof. Amit Mitra

Prof. Shramishtha Mitra

Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

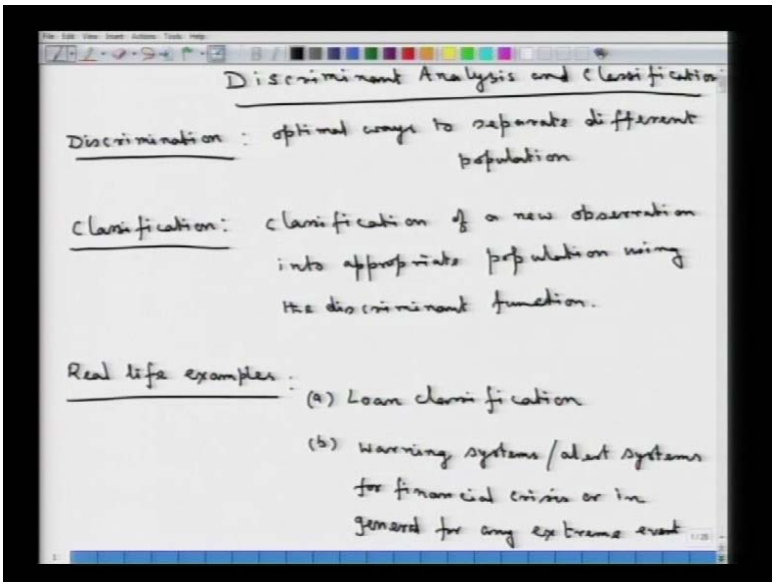
Lecture No. # 30

Discriminant Analysis and Classification

In this lecture, we will start looking at new problem, the problem of discriminant analysis and classification. So, it is a very important technique in multivariate data analysis. We look at actually a population, we look at a collection of multi-dimensional objects, and we use this concept of discrimination and classification. In order to do the following tasks actually, discrimination means it is an optimal way to separate different populations. So, we are trying look at some function that would lead us to discriminating members coming from different populations. So, we are trying to separate out different distinct populations.

Now, once that is done we are looking at next the classification problem; the classification problem is basically going to do the following job. Whenever a new observation is coming, whenever a multidimensional new observation is coming, using the discriminant function that we will be constructing, we would like to assign one of the possible populations to this new multi-dimensional vector. So, discriminant analysis and classification go side by side.

(Refer Slide Time: 01:33)



Let us look at what we are up to in this section, discriminant analysis and classification problem is what we are going to look at. It has two parts, as I said the first is discrimination or discriminant analysis. Discrimination is where we are trying to find out some optimal ways to separate different populations. By stating that we are trying to find out optimal ways to separate different populations; what we mean is what I try to explain at the very beginning that we have got multi-dimensional observations; they are possibly coming from pi 1, pi 2, pi k such k populations.

And we are trying to look at, what is the best way or rather what sort of function would be best in order to discriminate observations, multidimensional observations coming from different populations. Now, once discrimination, a discriminant analysis or discriminant function is in place, we look at classification. And that is basically the problem of classification of a new observation of A or many such new observations of a new observation into appropriate population using the discriminant function.

So, typically in such a problem what we have the set up that we have is the following that once you have a particular set of data, you have those multidimensional observations, as I said possibly coming from different populations. They are pre-classified, so, they have a class membership, they have a population membership that is clearly mentioned in the data, which we will call a learning set of data. And based on that particular previously classified data, we are going to build the discriminant function first, and then that discriminant functions, after calibration would be ready to be used as a classifier. And so that whenever a new observation is coming, one can actually classify it

to be coming from one of the possible populations. Now, it is a very important concept as such and which has got many real life examples. Let me just talk little bit about real life examples, wherein classification comes into picture. Some examples that come naturally to once mind this say a loan classification problem. What is there in this loan classification problem? A financial institution is confronted with the following problem that there are various loan applications. So, some loans applications are sanctioned, some are not, some loan application may be categorized into say following categories; that a loan application is categorized, as the potential high risk loan or a medium risk loan or a low risk loan.

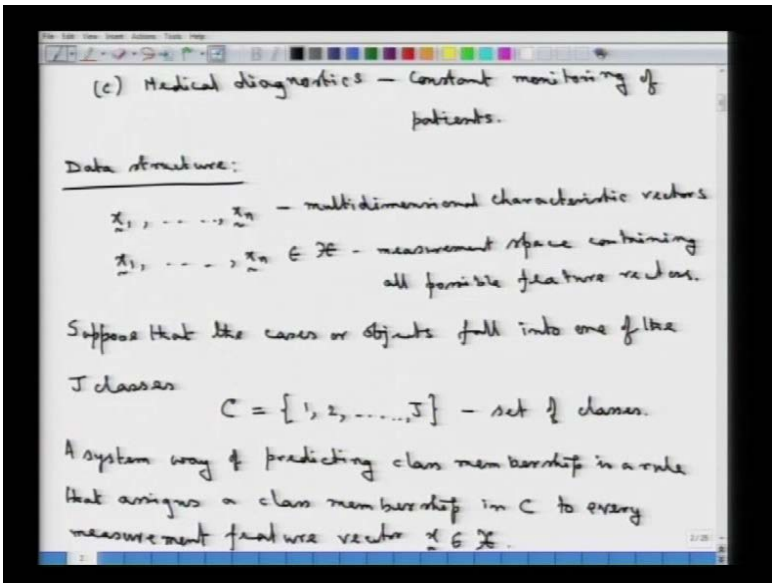
So say in that possible category, we have to classify a particular new loan application that is coming into one of those and decides, whether to sanction loan to that particular application or not. The setup is the following that when a person approaches a financial institution. Certain parameters of that particular individual are basically asked for and then looking at those characteristic features, which form the multidimensional vector based on those characteristics the multidimensional the entire multidimensional vector a decision has to be taken.

Whether to grant loan to that particular individual or firm or not? So, we have basically that particular problem coming down to the problem of discrimination and classification. So first of all, based on the past history, the past experience of what type of loan applications had come? And with what sort of feature vector we would have to build up a discriminant function? We will have to perform this discriminant analysis and then that particular function needs to be used after calibration on the past data, on the learning set data, to the new loan applications and then classify it into one of the possible classes like what I said; it is potentially a low risk loan, a high risk loan, or medium risk loan or things like that.

Now, second example that one can talk about is warning systems or alert systems for financial crisis, or in general for any extreme events. This once again is a very important application of discriminant analysis. Say for example in bank of super diction or prediction of currency crisis or prediction of a say credit card fraud; these type of analysis is very frequently applied. So wherein actually, if we consider say the example of bank of super diction, then looking at the present state of a particular firm, may be

financial firm, may be any other manufacturing firm. Looking at the present state looking at the present state of its financial conditions, one tries to classify that particular state of the firm as one which is potentially dangerous towards a bank of C type of situation that is another important application there are many such applications just looking at couple of such applications.

(Refer Slide Time: 08:31)



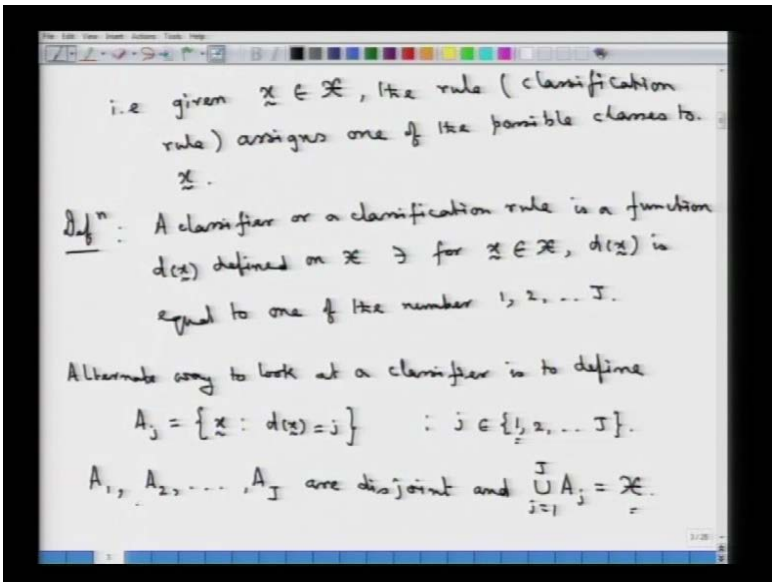
A third application that comes to once mind is a medical diagnostics. In medical diagnostics, actually say constant monitoring of patients conditions, patients parameters, health parameters and then making a classification of that particular patients condition to be critical or otherwise. So, once again one looks at the problem of discriminant analysis and using that as a classifier. In order to classify the state of that particular patients health condition into one of the possible categories. Now, what is the data structure in such a situation? The data structure when we are looking at such a problem is the following that we have got X1 vector, all are multidimensional vectors X1, X2, X n; these are multidimensional characteristic vectors. Now these X1, X2, X n, they belong to some sample space say script X, which is the measurement space containing all possible feature vectors. So these are basically, those multidimensional feature vectors which we are talking about. Say in the medical diagnostic problem, this will be different parameters of the patient. If we are talking about a loan application, loan classification type of problem then each of these would be loan applications. The characteristics corresponding to each of these loan applications, wherein the financial status mostly another social parameters of a particular loan applicant is looked at. So, these are those

multidimensional feature vectors. Now, along with these feature vectors, there is something that is required.

When we are looking at this problem, suppose that the cases these are the means feature vectors of the cases or objects fall into one of the j classes. Say that C is the set containing all such class memberships. So, these are the possible classes and C is a set of these classes. Now, in the loan application case when I talked about, say a particular loan application being low risk, medium risk, or high risk. So, we essentially try, to say that there are three classes in which a particular loan application can fall into. When we are talking about medical diagnostics, we are looking at say two classes. When we are saying that the condition of the patient is critical and the condition of the patient is not critical, so there are two possible classes. So, along with each of these multidimensional feature vectors, there will be a class membership attached to that, which say in general, we are talking about j classes; this is a set containing all those class identifications?

Now a systematic way, in view of this particular data structure and the definition, a systematic way of predicting class membership is a rule that assigns a class membership. In this set of classes to every measurement, which are vector to every measurement feature vector say X multidimensional belonging to the set of all possible such feature vectors? So, in view of this particular data structure and this class of set of class possible classes we are looking at the problem of building a systematic way of predicting the class membership. So, that basically is a rule that assigns a class membership in C. To every possible measurement feature vector X belonging to script X, that is in another words given X belonging to this script X. The rule the classification rule, that this rule is basically that classification rule.

(Refer Slide Time: 13:41)



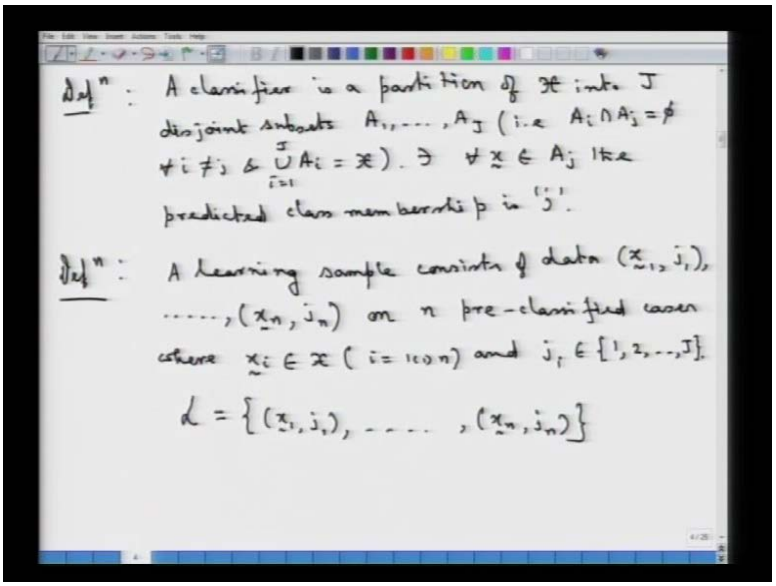
The classification rule assigns one of the possible classes to this X the feature vector. Let me give you a basic definition, which is basically based on what we have been discussing. A definition of a classifier goes like this. That a classifier or a classification rule both meaning the same obviously, a classifier or a classification rule is a function say d x, that function is defined on every feature vector X belonging to script X. Such that, for every X belonging to script X, this function d x which is the classifier is equal to one of the numbers 1, 2, to C that is for every X. This function or the classifier d is going to assign one and only one number in this particular set. And there is of course, no overlap between say assigning X to 2 classes that is not possible. So, we have assigned for every X belonging to the possible space of feature vectors one and only one class membership to that.

Now, there is an alternate way to look at this particular classification problem. Now once we talk about say assigning X to a particular member in the, I am sorry this is going to be one of the numbers 1, 2, up to j is a total number of classes. That we had taken we had denoted by C the set of all such classes. Now, when we say that, for every X, we are going to assign one of these numbers to that. So, that given that d x given that X we will say that it is class membership is d X. Now in doing what we are doing is, we are actually making a partition of the sample space. So, for a particular set of X, we are going to assign for every X belonging to that particular set a number between 1, 2, up to j. One unique identification number and hence the entire sample space X. Thus has got a partition which is induced by this particular classifier.

An alternate way to look at a classifier is the following is to define sets A j, such that A j is the set of all such X vectors; such that d X assigns the number j to that particular set. To all the X is belonging to that set here, this j belongs to this set of numbers 1, 2, up to j. So, this is where we are looking at partitioning the entire sample space X into it is possible partitions. Wherein A j denotes the set of all X is for which d assigns the same value small j. So, if we look at these sets now A1, A2, A j. So, this is the set of X is to which d assigns the number 1. So, the for every X belonging to the set A1, the class membership assign to that is the class one this particular class and similarly this for the second class and this for the j th class. So, these are required to be disjoint are disjoint and what we would require is that union of A j, S j equal to 1 to up to capital j is the entire sample space. Why do we require that? Because there is no ambiguity in assignment, in the way that a particular case X belonging to script X is assigned only one class membership and hence each of this particular A1, A2, A j are going to be mutually disjoint.

There is no common X belonging to any of these A j, S in this particular set and for every X, we are saying that for every X belonging to script X the feature vector space. We have to assign a class membership. So, it cannot be that there is some X S to which in class membership is not assigned. And hence the union of these A j, S j equal to 1 to up to capital j has to be this particular sample space. We have the intersection of A i, A j now this by saying that they are disjoint. What I what we say is that A i intersection A j that is equal to a null set phi. For every i not equal to j and we have union i equal to 1 to up to capital j of this A i elements that is equal to script X. And hence A1, A2, A j form a partition of the sample space and then we can give an alternate definition in terms of these. So, let me give that definition through the partition.

(Refer Slide Time: 19:51)



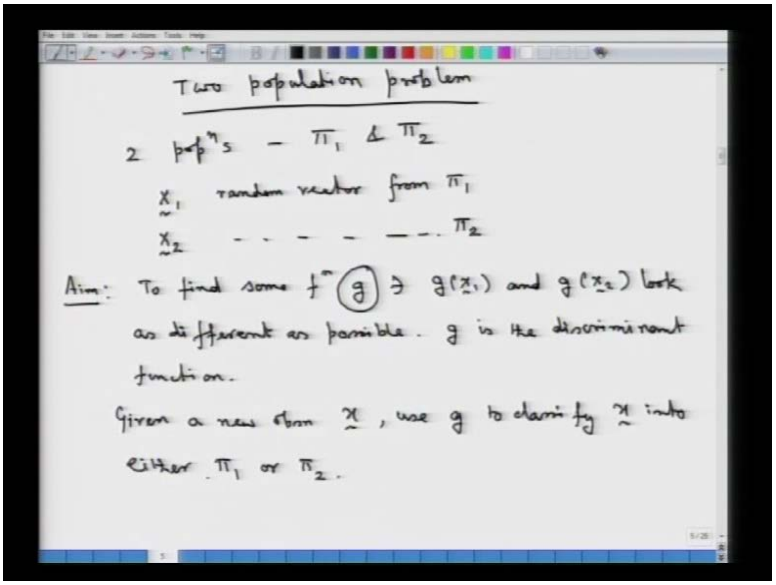
A classifier that is a point of interest or a classification rule we have already written that, so no need to write it again. This is a partition of the possible feature vector space, script X into j disjoint. When we say that its partition its disjoint subsets, A1, A2, A j, that is we have that condition that a i intersection A j equal to pi null set for every i not equal to j and union of these am is i equal to 1 to up to capital j that is equal to script x. It is through this particular partition, such that for every X belonging to A j. A particular A j, the predicted class membership is this small j. This small j of course belonging to one of these numbers 1 to up to capital j. So, this is how a classifier can alternatively be defined through the partition of this feature vector space.

Now, let me give you one more definition basic definition, what we mean by a learning sample consists of the following data, consists of data which is of the type that it is X 1 along with this X 1, which a feature vector is corresponding to the first case. We have A j one which is a class identification number corresponding to this case which is X 1 up to X n and corresponding to X n. We have another class membership j n attached to that X n on n pre-classified cases. Where we have these X is belonging to script X, i equal to 1 up to n and these j is belongs to this our set of possible classes 1, 2, 3, up to capital j. This becomes structure of the learning sample. So, the learning sample s is basically the collection of all such learn learning vectors.

Wherein the class memberships are given that this each of these cases are pre classified examples. This is how a learning sample looks like. Now, this is the history, this is from where one is going to build the classifier. And in future we are not going to have the

class membership. We are just going to have the feature vector and based on the classifier built on this learning sample, which is the set of n pre classified examples, one is going to build the classifier in an optimal way there are various ways of building the classifiers. That we are going to discuss in this concept of discriminant analysis and classification. Now, let us start looking at some such problems how to build up discriminant functions, and how to use such discriminant functions in practice. In order to classify features which are not classified?

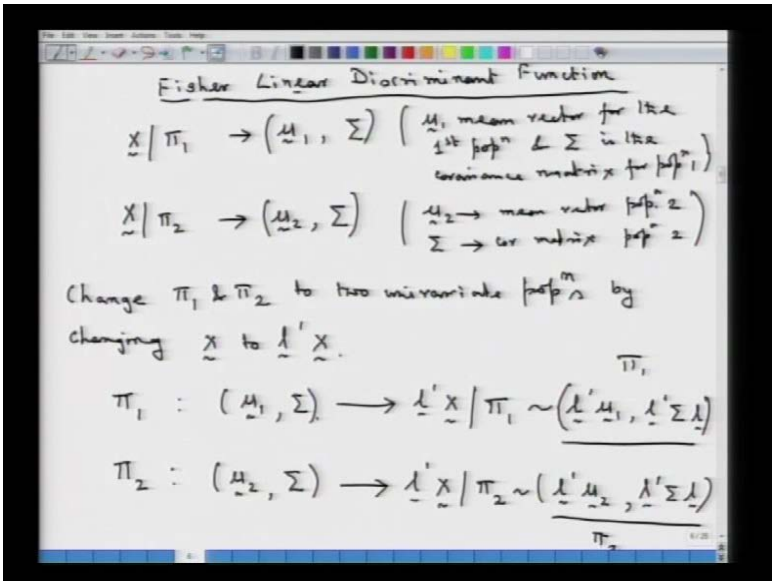
(Refer Slide Time: 24:08)



Let us first look at a simple fundamental problem with two samples. A two population problem rather a two population classification problem we have here two populations say pi 1 and pi 2, both are multivariate populations. Suppose I have X 1, a random vector from pi1. The first population X 2 is a random vector, which is coming from say the second population. When we talk about a discriminant function, we are trying to find out a function which would look as different as possible, when we have observations coming from two different populations. The basic aim here, when we are trying to build the discriminant function is the following the aim is to find some function say g. Such that our g X 1, if X 1 is coming from the first population and g X 2. X 2 is coming from this second population; they look as different as possible. Now, such a g can in that situation, if it is being, if it looks as different as possible for observations coming from two different distinct populations pi 1 and pi 2. Then g is the desire discriminant function and then given a new observation X. We can use that g to classify X into either pi 1 or pi 2 into either pi 1 or this pi 2.

That is basically the classification problem; first of all we will have to look for such a function g which would look as different as possible. So, it will distinguish observations coming from different populations as best as possible. And then once that is done, that is the discriminant function in place then we can use that discriminant function. In order to classify a new observation X, for which the class membership is not known so, we do not know whether X belongs to pi 1 or whether it belongs to pi 2 on the basis of this discriminant function. We will have a rule which would either put X into pi 1 or pi 2. So, that is a classification. Now, let us discuss a very fundamental concept in discriminant analysis which is referred to as the fisher linear discriminant.

(Refer Slide Time: 27:18)



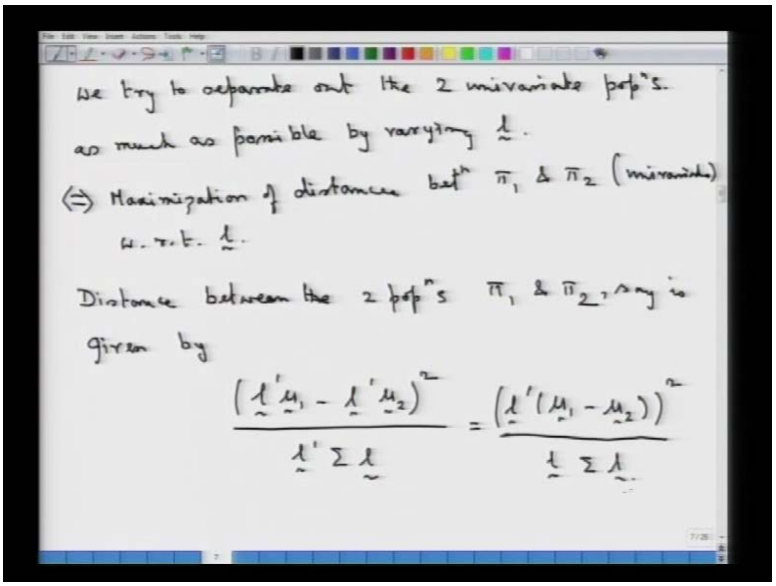
So, for the fisher linear discriminant function we have the following setup that if X belongs to pi 1. It is characterized by a mean vector, which is say given by mu 1 and the covariance matrix sigma. So, that we have this mu is the mean vector for the first population and this sigma is the variance covariance matrix for this population number 2. Similarly, if X belongs to pi 2 then the mean vector is say mu 2 and for simplicity, we will have to look at the sigma matrix to be similar. So, these are once again this is mean vector for population I am sorry this is sigma is the covariance matrix of population 1 only, because we are looking at pi 1 to be the first population. So, this is for the population 1 and sigma matrix is the covariance matrix. This is I am sorry this is mu 1 this is mu 2 and hence this is for the population 2. And this is a covariance matrix for

population 2. This mu 1 and sigma are corresponding to population 1 and mu 2 and sigma corresponding to population 2. So, this is what we have.

Now, these are the characteristics of the two populations. So, this is where the difference between the two populations in their mean vectors is. Now, we make the following change we linearize this population. So, change pi 1 and pi 2 to univariate populations. How we look at that univariate populations by changing X to some L prime x. Now, the point would be to determine, what is this L 1? Such L prime or L vector that the discrimination is best possible. That is what we are now doing is, that this pi 1. We had a population which was characterized by mu 1. The mean vector and sigma, the covariance matrix this is now changed to L prime X that population this given pi 1.

Now, we will have the characteristics as L prime mu 1. So, if X is the multivariate random vector, which has got mean vector as mu 1 and a covariance matrix as sigma. Then, if we have changed it to univariate population that is L prime x, we are now looking at the linear combination of the elements of these X vector. L prime X given mu 1 has got the characteristics that it is mean is L prime mu 1 and it is radiance is L prime sigma L. Now similarly, if we look at the second population, which is pi 2 which was in the multidimensional, setup characterized by, characterized by mu 2 and sigma. This now is a change to the univariate population which is L prime X. This given mu 2, now has been characterized by L prime mu 2 and L prime sigma l. So, these are two univariate populations this and this. So, this is now the univariate counterpart of those populations. This is the characterizing and this is the characterizing parameters of the second population which is pi 2. Now, we have 2 univariate populations where the variances are of the two populations are same. It is differing by the mean quantity for 1 it is L prime mu 1 and for the other its L prime mu 2. So, we will look at what L would separate out these two populations as far as possible that is we look at the distance.

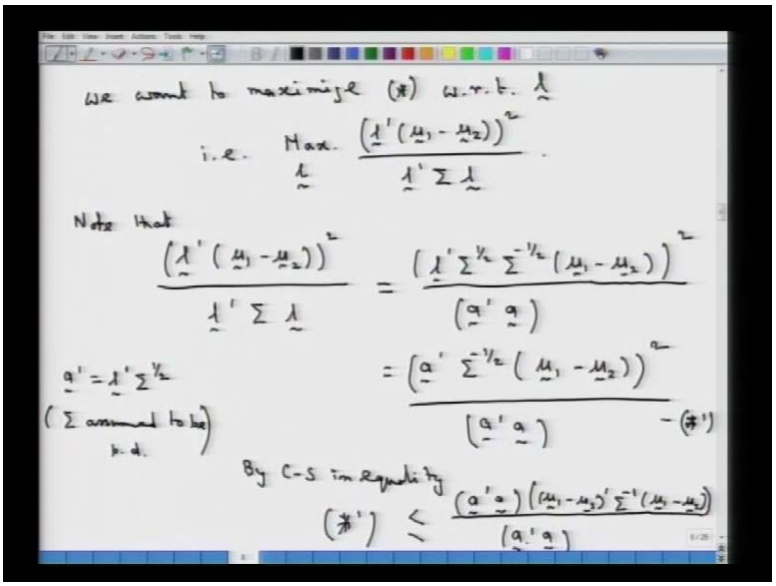
(Refer Slide Time: 32:29)



We try to separate out the two univariate populations as much as possible by varying or by choosing the best possible L by varying L, because L is what is a freedom given to us. Now, this is this problem is equivalent to maximization of the distance, maximization of the statistical distance between the two populations. Between pi 1 and pi 2, univariate with respect to this L, because L is what we are taking in l, so L is a freedom to us. So, we will try to choose L, such that the statistical distance between this univariate populations and this univariate population is maximum possible with respect to the choice of this L. Now, we can propose the following distance between the two population - pi 1 and pi 2. The statistical distance between the two population's pi 1 and pi 2 say is given by the following. So, it is L prime mu 1 minus L prime mu 2 whole square that divided by the variance L prime sigma l. So, this is same as L prime mu 1 minus mu 2 squares this divided by L prime sigma l.

So, this can be taken as a statistical distance between these two univariate populations. One with a mean L prime mu 1 and the other with a mean L prime mu 2 and with the same variance, we are looking at the different square in their means that standardized with respect to the variance the common variance. That is what we have now; if this is the distance between the two univariate populations. In order to have the optimum discriminant function, what we will have to do is to look at what is that L which would maximize such distance.

(Refer Slide Time: 35:25)

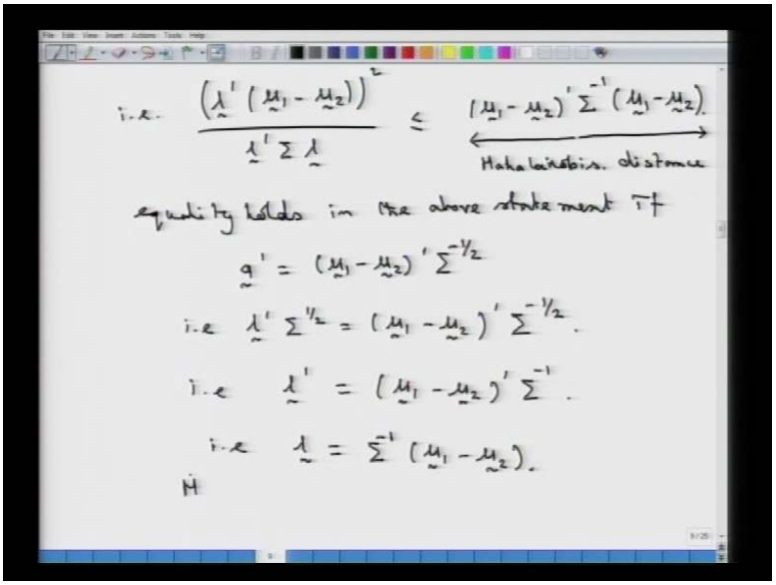


Now the problem thus to find out the best discriminant function, we want to maximize naturally, the distance between the two populations. We want to maximize this quantity. let me give this number say star, want to maximize this star with respect to L. In order to separate out the two populations in an optimum way, that is we try to look at maximization with respect to L of this function L prime mu 1 minus mu 2 whole squares that divided by L prime sigma l. When we are trying to look at the maximum of this particular quantity with respect to L. Note that we are now looking at this quantity, which is the quantity which we are trying to maximize with respect to L, which is the freedom with us L prime sigma, L now with the following definition a prime that is equal to L prime time's sigma half. Now, sigma is assumed to be positive definite sigma assumed to be a positive definite matrix.

So, we have this in terms of the vector, this is L prime sigma half sigma half l, this is a prime a vector and here we will have to introduce this sigma. Let me write it one step sigma L transpose sigma half sigma minus half times this mu 1 minus mu 1 this square. That this term, now can be written as a transpose, so that this is an a transpose sigma to the power minus half then we have mu 1 minus mu 2 whole square this divided by a prime a. Now, by Cauchy Schwarz inequality, this term by Cauchy Schwarz inequality we can say that this is less than or equal to. If this is given a number star 1 then this star 1 equation is less than or equal to. We do not disturb this denominator is just a prime a and then looking at this to be one vector. This to be the other vector this would be less than or equal to a prime, as we have whole square out here. So, it is a prime a and then

transpose of this into this vector itself. So, what will be having is mu 1 minus mu 2, transpose sigma to the power minus half into, sigma to the power minus half will make it sigma inverse that into mu 1 minus mu 2.

(Refer Slide Time: 38:50)

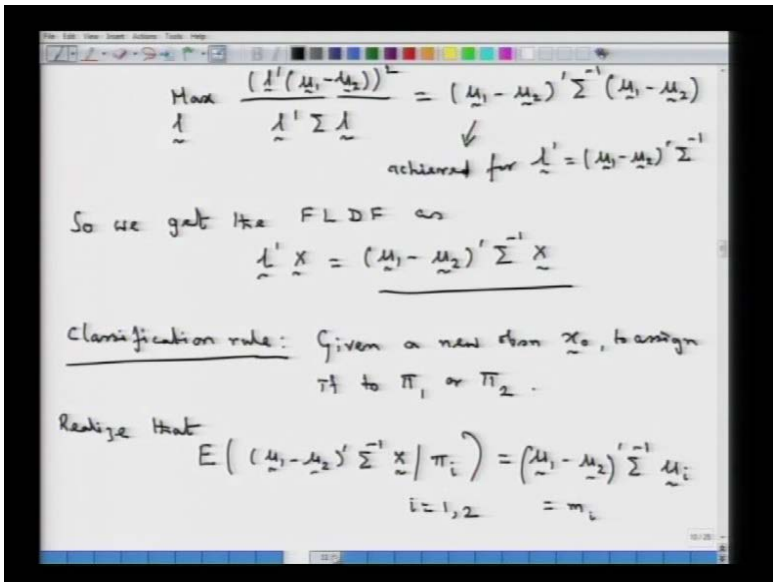


So, this is straight forward by using the Cauchy Schwarz inequality. These two terms cancel out that is this L prime mu 1 minus mu 2 whole squares this divided by L prime sigma. L this is less than or equal to what we have seen is mu 1 minus mu 2 prime this sigma inverse mu 1 minus mu 2. We have this particular term, which is referred to as the mahaloinobis distance also. So, the distance between the two univariate populations characterized by the mean vector L prime, mu i and a sigma to be the common L prime sigma L to be the common variance of those univariate populations is less than or equal to this particular distance which we call mahaloinobis distance between the two populations.

So, once we have this to be less than or equal to this. The mahaloinobis distance this less than or equal to term, here is coming from the application of Cauchy Schwarz inequality at this particular point. So, we know where the equality is going to hold equality holds in the above statement. If we have a prime to be equal to mu 1 minus mu 2 prime. Then sigma to the power minus half why is that? Because, if we look back at this expression. Here this is application of Cauchy Schwarz inequality to this particular term. So, this is one vector U prime V square less than equal to U prime U into V prime v.

That the equality will hold, if the two vectors are same or a constant multiplier of that. And in particular, we will have equality here. If a prime is the vector, that we had chosen out here, if that a prime is equal to this particular term. That is now what is a prime a prime is L prime sigma half, that is this L prime sigma half is equal to mu 1 minus mu 2 prime sigma to the power minus half. That is our L prime is equal to post multiplying this particular equation by sigma to the power minus half. What will be getting is mu 1 minus mu 2 prime sigma inverse. That is the optimum L, which would maximize the distance between the two univariate populations is going to be given by this particular sigma inverse mu 1 minus mu 2. Now, this is what is leading us to the fisher linear discriminant function. So, we get this is the L let me write one more step. Before we conclude, we were trying to find out the maximum distance.

(Refer Slide Time: 42:17)



Let me write it here, that the maximum L belonging to the appropriate dimension space of L prime mu 1 minus mu 2 whole square that divided by L prime sigma L. the statistical distance between the two univariate populations that is the mahaloinobis distance which is mu 1 minus mu 2 prime. This sigma inverse into mu 1 minus mu 2 and this is achieved for this L prime to be equal to the vector that we have derived there. That is mu 1 minus mu 2 prime sigma inverse mu 1 minus mu 2 prime sigma inverse. We get that L, which maximizes the distance between those two univariate populations that we had characterized out there mu 1 minus mu 2. So, we get the fisher linear discriminant function, as L prime optimized X, which is nothing but our mu 1 minus mu 2 prime sigma inverse x. So, this is a desired linear form linearization that we were looking at

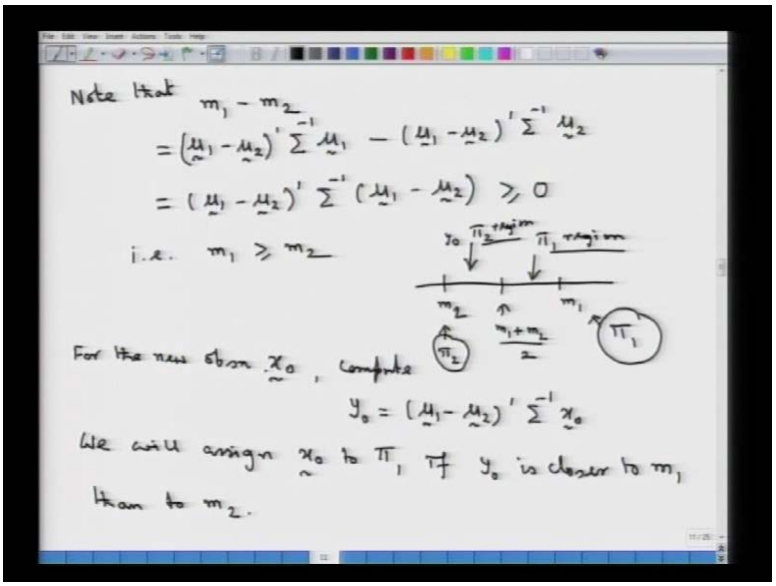
which would look as different as possible in the sense. That when we are looking at the corresponding univariate populations, then this function L prime X, where L prime is given by this is going to lead us to the maximum possible separation of the two univariate populations.

Now, comes the second part of this particular problem. Once we have this as the discriminant function, what would be a classification rule that is going to be based on this discriminant function? So, we are we have to now address the second part of this problem that is what is going to be the best classification rule. Now, what is the problem now? That given a new observation say x naught to affine it to pi 1 or pi 2. So, that is basically is the problem. Is that we well this is the best discriminant function, that we have come up with. Now we will have to frame a rule how this discriminant function is going to be used? When we have a new observation, for which the class membership is not known to us and we are trying to assign a class membership. That is we are going to assign x naught to either pi 1 or pi t2 based on what.

Now in order to derive the classification rule, we look at the following realization. So, realize that expectation of this linear efficient, linear discriminant function, that is mu 1 minus mu 2 prime sigma inverse X given any of this populations pi 1 or pi 2. Let us denoted by pi i. what is the expectation of this linear fisher linear discriminant function? Given pi i for i equal to 1 and 2 that would be given by the expectation of this particular function, when X belongs to the corresponding population pi i.

And what is that, this is a non stochastic part. So, what will be having is mu 1 minus mu 2 prime a sigma inverse and expectation of X, given pi i would be given by mu i simply, this is what is going to be the expectation of the FLDF. When we are looking at its expectation with respect to being belonging to that particular pi i population. Let me write a given notation here, say m i to this particular term. So, we will have an m 1 we will have an m 2.

(Refer Slide Time: 46:35)



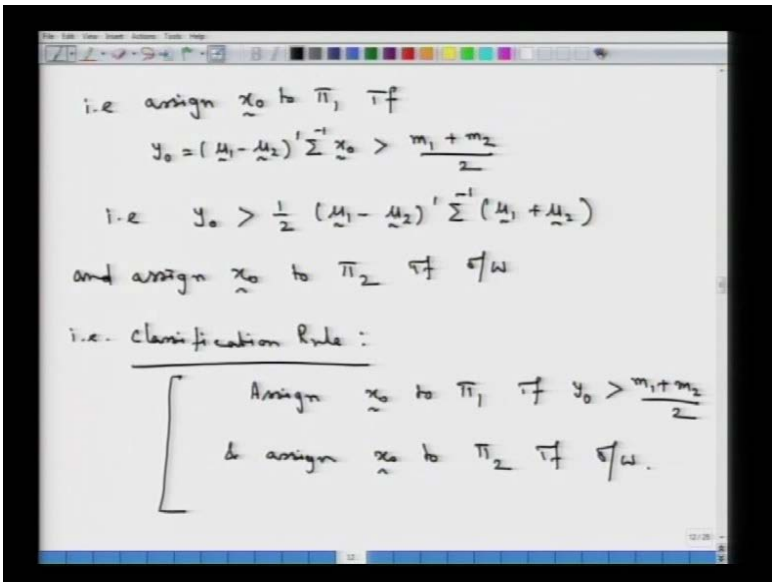
Now, note that if we look at m1 minus m 2. What is that going to be equal to? Now, m 1 is going to be mu 1 minus mu 2 prime sigma inverse m 1. I am sorry mu 1 this is going to be this as mu 1 vector this minus m 2 is the expectation of the fisher linear discriminant function. When it is coming from the second population and hence this is mu 1 minus mu 2 prime sigma inverse is mu 2. So, this is equal to mu 1 minus mu 2 transpose sigma inverse mu 1 minus mu 2. Now, note that sigma is positive definite and hence sigma inverse is also positive definite. And hence this is any vector that is belonging to say p dimensional space and hence this is going to be greater than or equal to 0. This would be equal to 0 only if mu 1 is equal to mu 2. In general, what we can say that if mu 1 is different from this mu 2 we will have this to be strictly greater than 0. That is we will have this m 1 to be greater than or equal to m 2.

From this relationship, that m 1 is greater than or equal to m 2. Let me have m 1 greater than m two. So, that this is say m 2 point and this is m 1 point and this is the midpoint say of m 1 plus m 2 this divided by 2. Now, the for the new observation for the new observation x naught compute say y naught which is equal to mu 1 minus mu 2 prime sigma inverse x naught. So, this is a given value that we are going to compute when we have x naught to be known to us. A following rule can be assigned I will discuss the logic of that particular rule, we will assign. So, we will assign x naught to pi 1, if this y naught is closure to m 1 then to m 2. Now this is a simple logical rule, why it is logical, because we had looked at the expectation of the fisher linear discriminant function. Under the condition, that it is belonging to two different populations. So, the expectation

of the fisher linear discriminant functions, when it is coming from pi 1 is equal to m 1 and if it is coming from m second population pi 2.

Then the expectation of the fisher linear discriminant function is m 2. We have m 1 to be greater than or equal to m 2. Now this is the value of the fisher linear discriminant function for a new observation which is x naught. Now, this is this m 1 is what is corresponding to my first population pi ones expectation of the fisher linear discriminant function. And this m 2 is expectation of the fisher linear discriminant function. When it is coming from the second population pi 2 and hence if the value of this fisher linear discriminant functions with the new observation x naught falls on this side. That is, if it is closer to m 1 than to m 2. If it is on this side of the middle line here, which is m 1 plus m 2 by 2 it is logical to assign the observation x naught towards this particular population which is pi 1. On the other hand if this is the pi 1 region and similarly if the value of this y naught. If the value of this y naught falls here that is if the value of y naught is closure to this m 2 we will assign y naught or x naught. Actually we will assign x naught to this pi 2 population. So, this is what the pi 2 regions are.

(Refer Slide Time: 51:21)

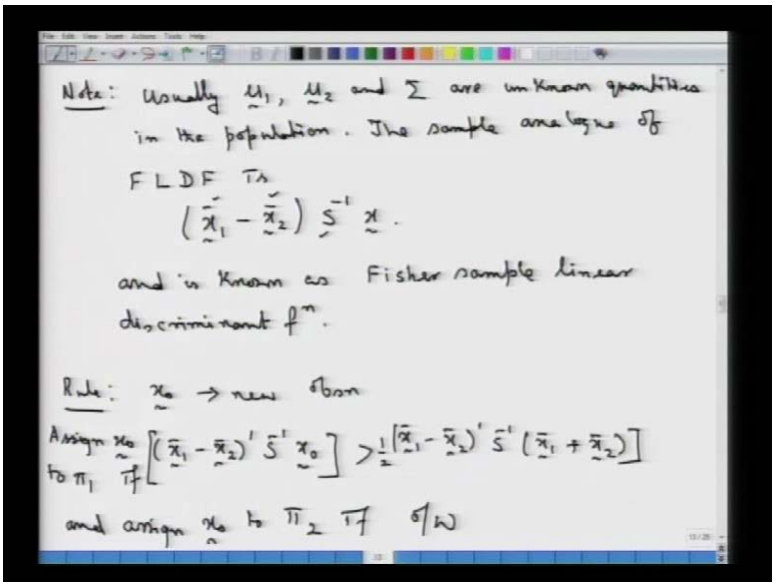


These two are the two regions corresponding to, what possible value that y naught can take here. That is assign x naught to pi 1. If we have the following that, we say that this is the region for pi 1 population that is if we have our y naught, which is equal to mu 1 minus mu 2 prime sigma inverse x naught. This term is greater than the midpoint which

is this one which is m 1 plus m 2 by 2, m 1 plus m 2 by 2 in terms of the values of m 1 and m 2. What is that equal to that is if y naught is greater than m 1 is the expectation of the FLDF under pi 1 and this under m 2. So, that this is half of mu 1 minus mu 2 prime sigma inverse mu 1 plus mu 2 and. So, this is the assignment rule and assign x naught to pi 2 if otherwise. So, it is basically that we are going to divide that particular segment m 1 m 2 through its midpoint. And if the value of y naught is closer to m 1 than to m 2. That is if y naught is greater than this midpoint and it is on the right hand side of the midpoint. We will assign that x naught new observation to pi 1 and if it is otherwise, we are going to assign x naught to the pi 2 population. So, this basically is the rule the classification rule that is using the fisher linear discriminant function. We have the classification rule; the classification rule says assign x naught to pi 1. If y naught is greater than just writing in short and or rather we can just say that it is an assign x naught to pi 2.

If it is otherwise this becomes the classification rule that is based on the fisher linear discriminant function. Now, note that there is something in this particular fisher linear discriminant function. And the classification rule that is going to pose a little bit of problem, the problem is that for any practical situations this mu 1, mu 2 and sigma they are unknown to us. And for computing this y naught, we would require this mu 1. We would require mu 2, we would require sigma inverse and also for computing the right hand side. That is this particular term here; we would also require the values of mu 1 mu 2 sigma also. And for all practical purposes, these are population characteristics. They are unknown and hence we would have to replace these quantities by the corresponding sample counterparts. And that would lead us to the sample fisher linear discriminant function and that would be in the perfectly implementable form.

(Refer Slide Time: 54:44)



We just put it as a note that, usually this mu 1, mu 2 and sigma are unknown quantities in the population. The sample counterparts, the sample analog of fisher linear discriminant function is given by x1 bar minus x2 bar. Where x1 bar is the mean of the first population the estimated mean of the first population; X2 bar is the mean of the second population. And then s is the pooled estimate of the sample variance of the population variance covariance matrix. So, this is what is going to be given by the fisher linear discriminant function. This is the sample analog of the fisher linear function and is known as the fisher sample linear discriminant function.

Now, if this is the fisher sample linear discriminant function the rule becomes the following in the light of this x naught is a new observation. We will compute this term which is x1 bar minus x2 bar transpose s inverse x naught. This is the sample counterpart of what we were talking here, this y naught. We will say that, assign x naught to pi 1. If we have this particular quantity to be greater than half x1 bar minus x2 bar transpose s inverse x1 bar plus x2 bar. That is this particular term here is more towards the first population than to the second population and assign x naught to pi 2. If it is otherwise that is if this quantity here is less than or equal to the right hand side quantity.

Now, since we will be having these estimates x1 bar x 2 bar and the pooled sample variance covariance matrix. as S there is no problem as such an implementing say this rule this classification rule, we will have n of n 1 of observation say coming from the first population, n 2 observations coming from the second population. So, based on n 1 observation we will compute x 1 bar based on n 2 observations we will compute x 2 bar and then pulling n 1 plus n 2 observations, we will have S and using that this is in a perfectly implementable form.

So, in the next lecture what we will see is to look at this classification problem. In a more general setup, wherein we will introduce misclassification, because there is always a chance that if we are looking at the classification problem, that a particular observation may be coming from population number 1 pi 1. And by mistake whatever, we propose as a classification rule, it may get classified into the other population leading us to a misclassification problem. Now, in most practical situation there is always a danger of misclassification. We will have to introduce such concepts, as cost of misclassification, and hence we will have to design or rather design optimum strategies, which would find out what is the best under such general classification problem. Thank you.