Applied Multivariate Analysis

Prof. Amit Mitra

Prof. Shramishtha Mitra

Department of Mathematics and Statistics

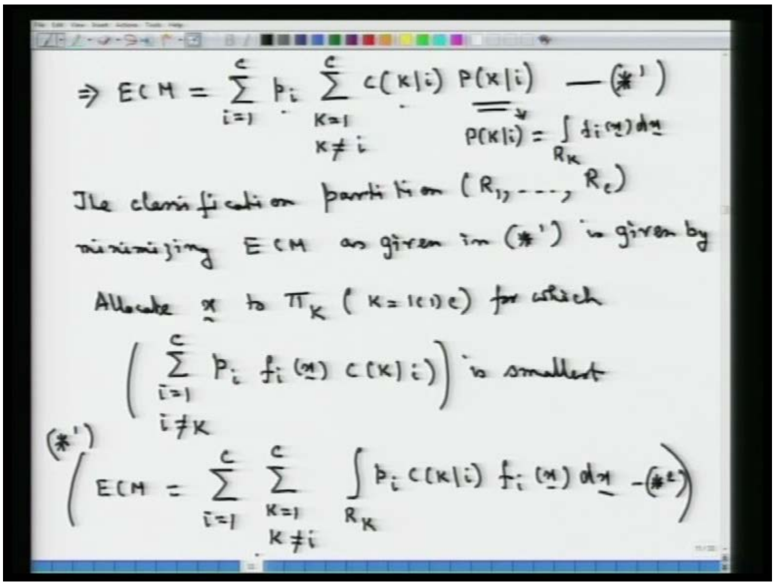
Indian Institute of Technology, Kanpur

Lecture No. #34

Discriminant Analysis and Classification

In the last lecture, we had looked at various types of discriminant functions, the rules that one can derive when we have multi population problem. So, specifically, we had discussed the problem like this that if we have got c populations pi 1, pi 2, pi c. And then, under a general classification problem, if we consider cost misclassification prior probabilities and so on. Then what sort of optimum rules are desirable.

(Refer Slide Time: 00:45)



$$\Rightarrow ECM = \sum_{i=1}^{c} p_i \sum_{\substack{K=1 \\ K \neq i}}^{c} c(K|i)\, \underline{\underline{P(x|i)}} \quad — (*')$$

$$P(K|i) = \int_{R_K} f_i(\underset{\sim}{x})\, d\underset{\sim}{x}$$

The classification partition $(R_1, ---, R_c)$ minimizing ECM as given in $(*')$ is given by

Allocate $\underset{\sim}{x}$ to π_K $(K = 1(1)c)$ for which

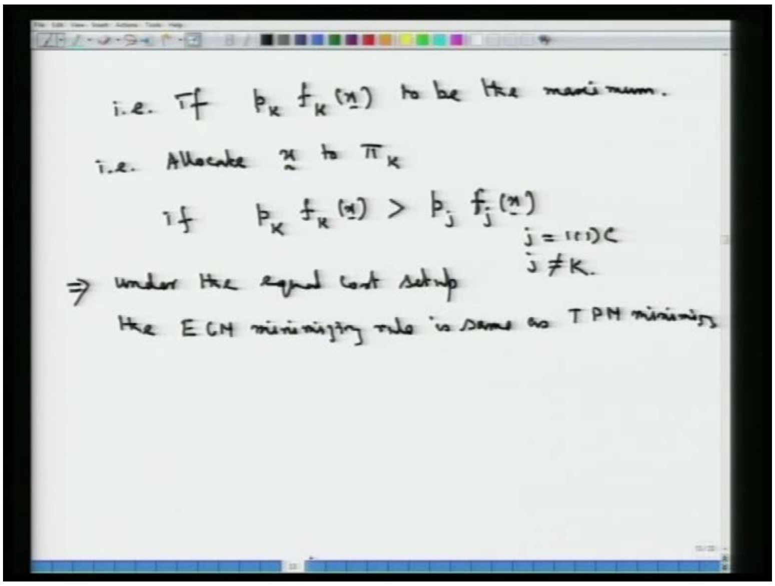
$$\left(\sum_{\substack{i=1 \\ i \neq K}}^{c} p_i\, f_i(\underset{\sim}{x})\, c(K|i) \right) \text{ is smallest}$$

$(*')$

$$\left(ECM = \sum_{i=1}^{c} \sum_{\substack{K=1 \\ K \neq i}}^{c} \int_{R_K} p_i\, c(K|i)\, f_i(\underset{\sim}{x})\, d\underset{\sim}{x} \quad -(*') \right)$$

Say for example, we had derived the rule which would minimize the expected cost of misclassification, appropriately derived under such a c population setup. And we had also looked at the rule which would optimize, minimize actually. The total the total probability of misclassification, the two rules were seen to be the same, when we have got the following formulation that we were looking at a rule that would minimize the expected cost of misclassification, that was given by this particular expression. And then, we had seen that under a small note, that under the equal cost setup, if we assume that the costs of misclassification quantities are all assumed to be equal, then naturally one is going to get the rule which is going to be the total probability of misclassification minimizing rule. And that is what we had seen in the last lecture.

(Refer Slide Time: 01:37)



i.e. If $p_k f_k(\underset{\sim}{x})$ to be the maximum.

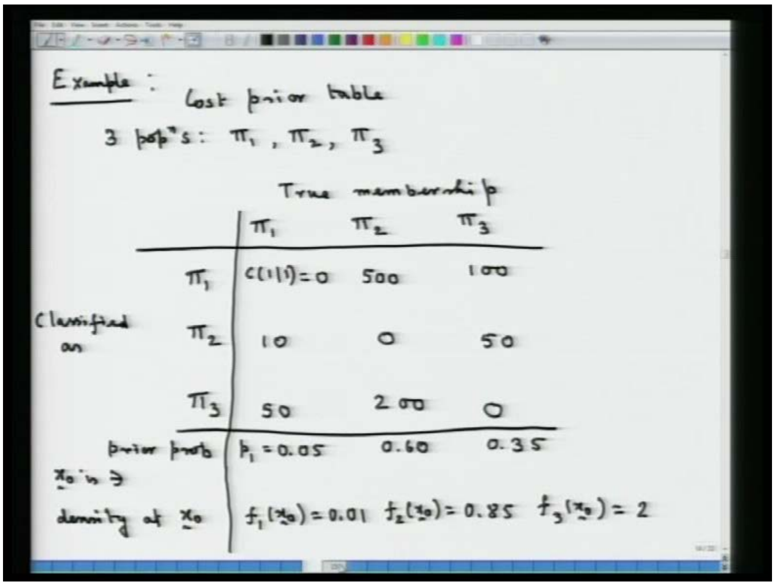
i.e. Allocate $\underset{\sim}{x}$ to π_k

if $p_k f_k(\underset{\sim}{x}) > p_j f_j(\underset{\sim}{x})$ $\quad j = 1(1)c$, $j \neq k$.

$\Rightarrow$ under the equal cost setup
the ECM minimizing rule is same as TPM minimizing

So, in this lecture, what first we are going to look at is an example of the type of theory that we had developed in the last lecture.

(Refer Slide Time: 01:53)



Example : Cost prior table

3 popns: π_1, π_2, π_3

True membership

		π_1	π_2	π_3
Classified as	π_1	$c(1\|1) = 0$	500	100
	π_2	10	0	50
	π_3	50	200	0
	prior prob	$p_1 = 0.05$	0.60	0.35
$\underset{\sim}{x_0}$ is $\ni$ density at $\underset{\sim}{x_0}$		$f_1(\underset{\sim}{x_0}) = 0.01$	$f_2(\underset{\sim}{x_0}) = 0.85$	$f_3(\underset{\sim}{x_0}) = 2$

And then we will also talk about other popular type of classification functions. Let us first, and then look at this example. We have got the following example, this is the cost prior table cost prior table corresponding to the three population setup. So, we have got three populations, which are say denoted by pi 1, pi 2, and pi 3.

Let us make this particular table, which is going to give us the cost of misclassification and also the prior probabilities of the corresponding populations. So, suppose we have on this side true membership, true membership of that individual. Now an individual can either belong to pi 1 or it can belong to pi 2 or it can belong to pi 3, because these three are the three possible populations. And corresponding to the true membership, corresponding to our classification rule, we can get that object being classified as coming from either pi 1, this is going to be the classification that is based on the classification rule in place. So, it can be either of these three pi 1, pi 2, and pi 3.

Now, if we have got an observation classified into pi 1, and if the true membership is really pi 1. Then there should not be any cost of misclassification. So, this entry here, which is c 1 1 that is it is coming from pi 1, it is getting classified into pi 1 itself. So, that should be equal to 0. So, all these three diagonal entries c 2 2 and c 3 3, all these three quantities are 0, the off diagonal elements are cost of misclassifications.

Suppose we have for this hypothetical example that this is equal to 100, this is equal to this is 500, this is equal to 100, this is equal to 50, this is equal to 10, this is equal to 50, and this is equal to 200. So, this is the cost of misclassification table, note that this cost of misclassification table, this matrix actually 3 by 3 matrix need not be a symmetric matrix, because we do not have the same cost like this one is c 2 1. And, the other one here on this matrix is c 1 given to and it is not really natural to assume that the two will be equal. And hence, we have 2 different misclassification cost corresponding to do such situations.

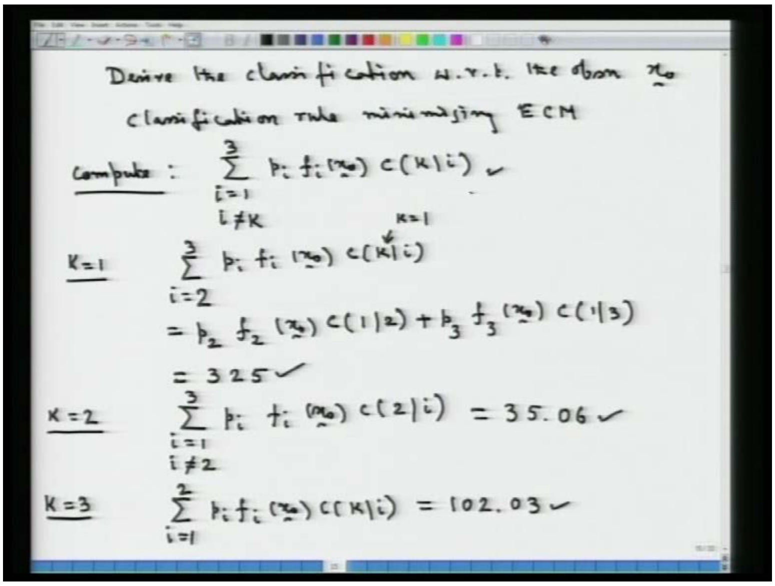
Now, let us have these prior probabilities also. So, the prior probabilities of pi 1, pi 2, and pi 3 are given by say p 1, which is in our notation p pi 1. So, that is equal to say, 0.05, this is equal to 0.60 say and this is equal to 0.35.

So, these three are the three prior probabilities, the sum of these three of course, is going to be equal to 1, because we are assuming that there are only three populations. This c here is equal to 3 anyway, in general we had denoted that by pi 1, pi 2, pi c and this we have as 3, c equal to 3.

Now, suppose x naught under such a situation caused by prior table, x naught is a new observation is say such that we have got the density at x naught corresponding to the respective populations pi 1, pi 2, and pi 3 are say given by f 1 x naught, x naught is a multivariate observation, say this is given 0.01, this is f 2 x naught. So, that is the density of this variable of this multivariate vector x naught, under the second population pi 2 that is equal to say given by 0.85, and this is f 3 x naught. This is for the third population, this is say given by the difference or it can be anything else also, this is going to be given by say 2.

Now, corresponding to such setup, we will implement what the type of optimum rule optimum classification rules that we have learned. So, we have got the crossed structures to be not equal.

(Refer Slide Time: 06:29)



Derive the classification w.r.t. the obsn $\underset{\sim}{x_0}$

classification rule minimizing ECM

Compute : $\sum_{\substack{i=1 \\ i \neq K}}^{3} p_i f_i(\underset{\sim}{x_0})\, c(K|i)$ ✓

K=1 $\sum_{i=2}^{3} p_i f_i(\underset{\sim}{x_0})\, c(\overset{K=1}{\overset{\downarrow}{K}}|i)$

$= p_2 f_2(\underset{\sim}{x_0})\, c(1|2) + p_3 f_3(\underset{\sim}{x_0})\, c(1|3)$

$= 325$ ✓

K=2 $\sum_{\substack{i=1 \\ i \neq 2}}^{3} p_i f_i(\underset{\sim}{x_0})\, c(2|i) = 35.06$ ✓

K=3 $\sum_{i=1}^{2} p_i f_i(\underset{\sim}{x_0})\, c(K|i) = 102.03$ ✓

So, we have the cost of misclassification to be taken into account, and hence we are going to compute the following quantities, in order to implement the expected cost of, I will first say that what we are going to do.

Derive the classification derive the classification with respect to the observation, which is x naught. So, in order to do that we are trying to find out the classification rule. So, this classification rule is one that is going to minimize the expected cost of misclassification, classification rule minimizing the expected cost of misclassification. So, we need to compute the following quantities, one after the other. This is going to be given by i equal to 1 to up to 3, there are 3 possible populations for i not equal to k. So, for specific choices of k, we are going to compute the quantity which is going to be given by this multiplied by c k given i.

Now, why do we need to compute this, we will have to find out for which k such an expression from among possible choices of k, which is 1, 2, and 3. This quantity is what we are going to get the smallest, why is that so. Because in the general classification rule, when we were looking at expected cost of classification minimization rule, we had r k to be that partition of the sample space r k is the region of all x s such that this quantity here needs to be the smallest. And hence, we had got this classification rule that x to be allotted or allocated to pi k for which we have this quantity for i not equal to k is smallest.

So, in order to implement this expected cost of misclassification minimizing rule, what we need to do, is to compute the this quantity here for different choices of k, k equal to 1 to up to this quantity c. And then, find out which is giving us the minimum and that is going to give us the allocation corresponding to the rule which is minimizing the expected cost of misclassification.

So, we compute this quantity for k equal to 1, k equal to 2, and k equal to 3. It is trivial to find out this term here for k equal to 1, 2, and 3. So, this is going to be i not equal to k, and hence if k is equal to 1. Then, I would start from 2 and then we will look at this particular product here, which c k is given i. So, this can be computed very easily this is p 2 f 2 x naught.

Now, this k is equal to 1 in our case, because we are choosing k equal to 1. So, this is going to be 1 given to this plus the term corresponding to i equal to 3. So, that is p 3 into f 3 x naught that into c 1 given 3. So, using this cost prior table, one can compute what this quantity is equal to it turns out that this is equal to 300, 25.

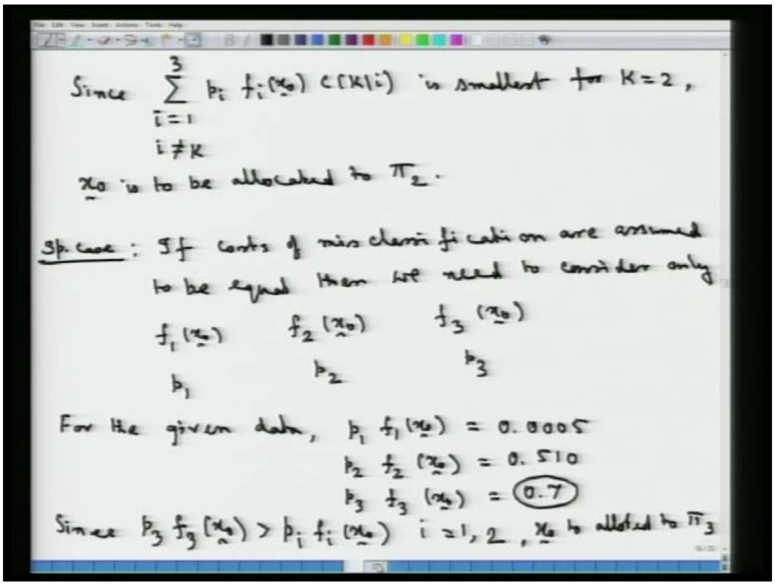
Now, similarly this is. So, this is the value of this quantity here, this quantity here for k equal to 1. Similarly, we need to find out, what is this for k equal to 2. So, for k equal to 2, we will have an expression which is summation i equal to 1 to up to 3. And then, this i is not equal to k that is i not equal to 2. So, the 2 terms in this summation would correspond to i equal to 1, and i equal to 3 of this expression, which is p i and f i x naught into c k k here is equal to 2. So, we can write it straight away as 2 given i.

So, one can similarly computed from the previous cost prior table, and distance out to be that it has the value that it 35.06 and for k equal to 3. This expression now, summation for k not equal to i , not equal to k. So, k being 3 this summation is from i equal to 1 to up to 2, 2 terms. So, this f i x naught into c k given i, this term here from the given data turns out that it is 102.03.

So, we have computed this quantity here for every value of k, for k equal to 1 this turns to be this for k equal to 2, it turns out to be this, and for k equal to three, this turns out to be this. So, if we have got a multi population set up, say for k greater than 3 the same logic can actually be applied. So, one can look at all the values of k, k equal to 1 to up to c, and then look at the one, which is giving us the minimum.

In this in the present situation here ,what we have is for k equal to 2 the expression which minimizes the expected cost of misclassification rule that it is giving us for k equal to 2, the minimum possible that is 35.06 is minimum among these 3 observations. And hence, what we have is that x naught.

(Refer Slide Time: 12:21)



Since $\sum_{\substack{i=1 \\ i \neq K}}^{3} p_i f_i(\underset{\sim}{x_0}) c(K|i)$ is smallest for $K=2$,

$\underset{\sim}{x_0}$ is to be allocated to π_2.

Sp. case : If costs of misclassification are assumed to be equal then we need to consider only

$f_1(\underset{\sim}{x_0})$ $f_2(\underset{\sim}{x_0})$ $f_3(\underset{\sim}{x_0})$

p_1 p_2 p_3

For the given data, $p_1 f_1(\underset{\sim}{x_0}) = 0.0005$

$p_2 f_2(\underset{\sim}{x_0}) = 0.510$

$p_3 f_3(\underset{\sim}{x_0}) = \boxed{0.7}$

Since $p_3 f_3(\underset{\sim}{x_0}) > p_i f_i(\underset{\sim}{x_0})$ $i=1,2$, $\underset{\sim}{x_0}$ to allotted to π_3

For which we have got the this density and x naught to be given by this values, that x naught is going to be allocated to the second population. And that is how this rule works.

So, this implies that since we have got this summation i equal to 1 to up to 3 here, i not equal to k of the quantities p i f i x naught into c k I, this is smallest for k equal to 2. This, we will have x naught being classified k equal to 2, x naught the new observation is to be allocated to the second population that is pi 2. So, this how this type of a same minimizing rule works.

Now, if we take a special case in the previous example, suppose we take the cost of misclassification to be same, if costs of misclassification are assumed to be equal are assumed to be equal. Then we need to just consider, we need to consider only the quantities, which are f 1 x naught, f 2 x naught, f 3 x naught, and the corresponding prior probabilities that is p 1, p 2, and p 3. Because, if in this particular setup, we now assume for simplicity that all these quantities are equal except those which are on the diagonal.

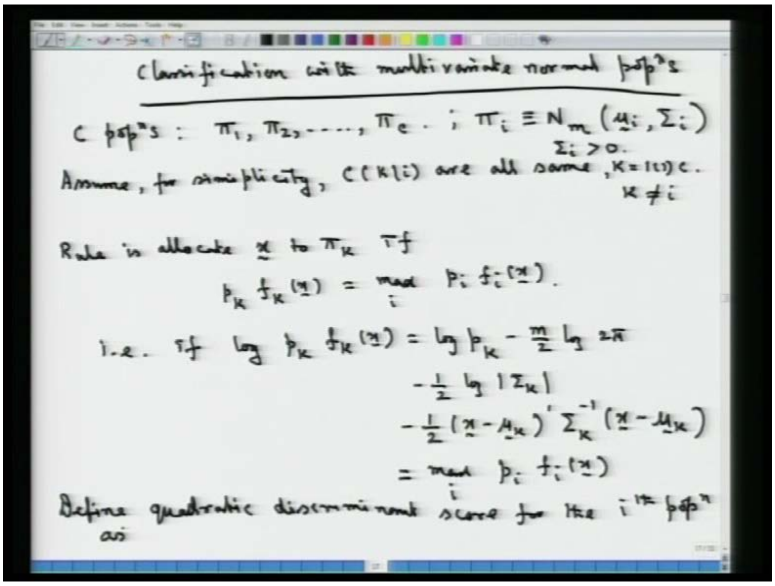
Then, the special cost structure does not make any difference, and we will have this special case out here, what we said in the note that under the equal cost setup. There is same minimizing rule, this is what is going to lead us to the total probability of misclassification minimizing rule. And then we will allocate x to pi k, if this quantity is smallest wherein the cost structure of course, comes out because they are the same. And, when we are basically going to allocate x to pi k the k th population, if this p k f k x is greater than p j f j x for j equal to 1 to up to c j not equal to k. And hence, we would just be looking at under such a situation these 6 quantities.

Now, what we have is the following quantity that this is for the given data for the given data, what we have are the following quantities that this p 1 f 1 x naught that is equal to 0.0005 p 2 f 2 x naught, for the given x naught to be equal to 0.510, and p 3 f 3 x naught that is equal to 0.7.

So, we have these three quantities to be given by this. Now, we are going to allocate x naught to that pi, for which we have got one of these quantities to be the highest, and what is that this one is the largest. Since we have got in the present situation, this p 3 f 3, x naught is get greater than p i f i x naught for i equal to 1 and 2, that is leaving a side that k equal to 3 are termed on the right hand side, x naught is allotted to the third population which is pi 3.

So, this is in line with what the theory had suggested us that we are going to allocate an observation x 2 pi k, if this quantity on the left hand side is greater than all the c minus 1 quantities on the right hand side, and that is what was happening out here. That this is the highest among all the three, and hence this x naught the new observation is allotted to that third population, which is pi 3.

(Refer Slide Time: 16:54)



Classification with multivariate normal pop^n s

c pop^n s : $\pi_1, \pi_2, \ldots, \pi_c$. ; $\pi_i \equiv N_m(\underset{\sim}{\mu}_i, \Sigma_i)$ $\Sigma_i > 0$.

Assume, for simplicity, $c(k|i)$ are all same, $k = 1(1)c$. $k \neq i$

Rule is allocate $\underset{\sim}{x}$ to π_k if

$$p_k f_k(\underset{\sim}{x}) = \max_i p_i f_i(\underset{\sim}{x}).$$

i.e. if

$$\log p_k f_k(\underset{\sim}{x}) = \log p_k - \frac{m}{2} \log 2\pi - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_k)' \Sigma_k^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_k) = \max_i p_i f_i(\underset{\sim}{x})$$

Define quadratic discriminant score for the i^{th} pop^n as

Now, depending on the problem, this can the type of implementation of the theory can be simple, it can be difficult also. Now let us look at some specific examples, when we have classification problem for multivariate normal populations. So, this is what we are going to look at.

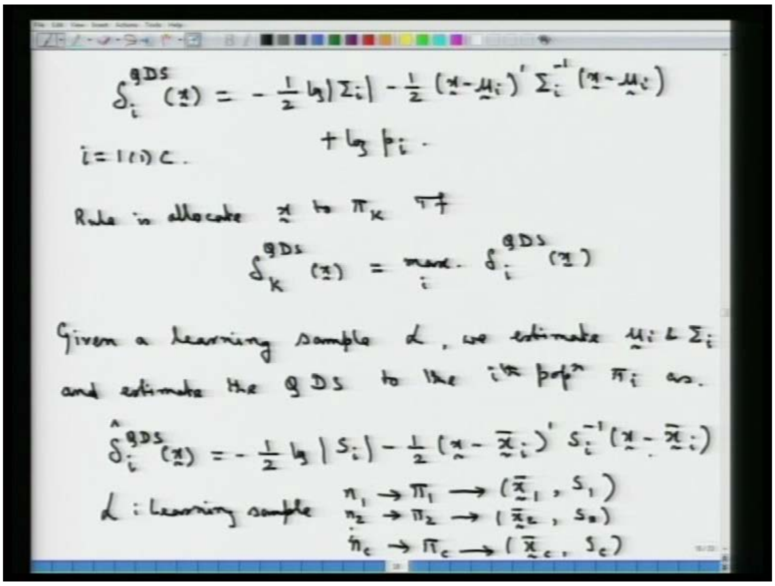
So, suppose we still have c populations, they are denoted by as before pi 1, pi 2, pi c. So, these are the c populations, assume for simplicity that the cost structures are same assume for simplicity. Otherwise it is going to be a bit more complicated, nothing is going to change much. Assume for simplicity that this c k i are all same for k equal to 1 to up to c and k not equal to i.

Now, if we look at the rule, which minimizes the expected cost of misclassification or the one which is going to minimize the total probability of misclassification, the two are going to be the same. Because we have got we have assumed for simplicity that the cost structures are same rule is allocate x to pi k, if we have got this p k f k x to be equal to maximum over i of the quantities of the form which is p i f i x that is if we have got log of this quantity log of p k f k x.

Now, we know what is f k x, because the k th population is a multivariate normal distribution. So, that let us, at this point of time take this pi i to be a multivariate normal population say m dimension with a mean vector equal to mu i, and the covariance matrix equal to sigma i, wherein we assume that this sigma i for every value of i is positive definite matrix.

So, we will have to look at what this quantity is equal to that is simple, because we have got the multivariate normality. So, this would be log of p k, then we will have minus m by 2 log of 2 pi term, then we will have minus half log of determinant of pi k. And then, the negative of the exponent or rather the log of the exponent minus half x minus mu k transpose sigma k inverse x minus mu k. So, that is our this log of p k times f k this is equal to maximum over i of the respective quantities p i f i x quantities.

(Refer Slide Time: 20:52)



$$\delta_i^{QDS}(\underset{\sim}{x}) = -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_i)'\Sigma_i^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_i) + \log p_i .$$

$i = 1(1)c.$

Rule to allocate $\underset{\sim}{x}$ to π_K iff

$$\delta_K^{QDS}(\underset{\sim}{x}) = \max_i . \ \delta_i^{QDS}(\underset{\sim}{x})$$

Given a learning sample $\mathcal{L}$, we estimate $\underset{\sim}{\mu}_i$ & Σ_i and estimate the QDS to the i^{th} popn π_i as.

$$\hat{\delta}_i^{QDS}(\underset{\sim}{x}) = -\frac{1}{2}\log|S_i| - \frac{1}{2}(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_i)' S_i^{-1}(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_i)$$

$\mathcal{L}$: Learning sample

$n_1 \to \pi_1 \longrightarrow (\bar{\underset{\sim}{x}}_1, S_1)$

$n_2 \to \pi_2 \longrightarrow (\bar{\underset{\sim}{x}}_2, S_2)$

$n_c \to \pi_c \longrightarrow (\bar{\underset{\sim}{x}}_c, S_c)$

Now, let us define the quadratic discriminant function as following, define a quadratic discriminant function quadratic discriminant function or quadratic discriminate score say, let us denote that by quadratic discriminant score for the i th population as the quantity, say delta i Q D S quadratic discriminate score. This is for this point x to be equal to the quantity which actually matters, which is minus half log of determinant of sigma i minus half x minus mu i transpose sigma i inverse x minus mu I, this is what it comes there and this plus log of p i.

Why do we take this and define it as a quadratic discriminate code, because if we look at this log of p k f k x. This quantity is a constant which does not depend on p k sigma k or mu k, which are the characteristics of the k th population. And hence, what is going to matter when we are going to look at the maximum of this p i f i xs is the quantity, which is comprising of this term, this term and this particular term. So, taking all those terms the quadratic discriminant score for the i th population is defined in this way.

So, we have got i equal to 1 to up to c, and then rule is basically to allocate x to pi k, if in terms of the quadratic discriminant scores this delta k of x quadratic discriminate score is the maximum over i of this delta i Q D S terms. So, the classification rule is going to be just based on these quadratic discriminant scores.

Now, in this particular setup, in the multi population set up as, what we have in here. If we take that the costs c k is are not the same, then we will have to go back to the expected cost of misclassification minimizing rule, the general rule that we had derived which was given by this following term. Then r k would be the region for all x s such that this we will have the term here to be minimum. In such situation, this c k is which if they are not assumed to be the same or they cannot be taken to be the same. We will have to look at this particular expression, and find out for which value of k left out in this summation here. The quantity is going to be the smallest, but we have chosen it to be for simplicity, that this cost structures are same. And hence, the entire classification is going to be based on this quadratic discriminant score.

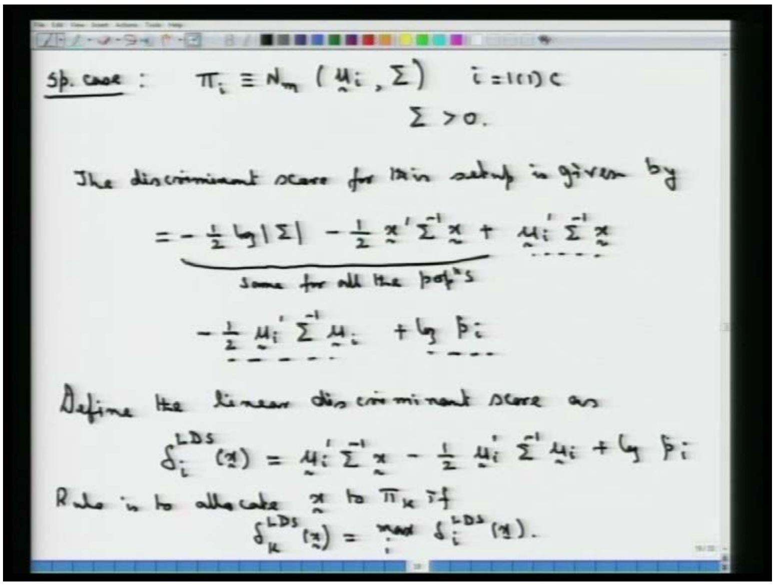
Now, it may be noted that when we are talking about this quadratic discriminant score, and we are saying that the rule is to allocate to pi k, if we have got the quadratic discriminant score corresponding to the k th population is the maximum among all such quadratic discriminant scores. A few quantities here are unknown actually, there is mu i terms, sigma i terms, sigma i inverse or sigma i matrix in general actually not known to us. So, how do we implement for a given data.

Given a learning sample remember that we had talked about the learning sample, learning sample is a sample for which we have got actually the cases being pre classified. So, it is basically the training sample, wherein we have the feature vectors and the corresponding classifier classify or rather the population marks, the population identification marks of the respective feature vectors. So, these are containing these are containing actually these the learning sample contains the pre classified examples, on which we are going to build our classification function.

So, given a learning sample l, we estimate mu I, and sigma I, and estimate the Q D S for the i th population pi i as the corresponding say delta i u d s cap. So, that is the estimated quadratic discriminant score. And that is going to be given by minus half log determinant of s i hat or just s i s i is basically the estimate of sigma i matrix. So, this minus half x minus x bar i transpose s i inverse x minus this x bar i, wherein this is the estimated sample mean based on the observations coming from pi I, because if we have got l this is our learning sample. So, this is our learning sample.

Suppose in this learning sample, we have got n one observations from pi 1 population, n 2 observations from pi 2 population, and similarly n c observations from the pi c population. Then based on these n 1 observations, which are pre classified which have a which have a tag identification tag corresponding to the first population pi 1. So, based on these n 1 observations, one is going to compute. Because these n 1 observations are all multivariate observations. So, one is going to compute this x bar r i which is the mean vector, which is the estimate rather of the mean vector of the first population. And s i that one upon n i minus 1 x i minus x bar x bar i into x i minus x bar transpose.

So, this is the estimate of sigma i matrix, and similarly for each of this sets of populations. From the respective populations, we can actually find out. So, this would be x 1 bar. So, this is going to be x 1 bar, and this is going to be our s 1, this we have x 2 bar, this is the mean vector computed from the n 2 observations. And similarly, for the c th population based on n c observations. We have the mean vector being computed at x bar c, and s c is the corresponding estimate of the variance covariance matrix sigma c corresponding to the pi c population. So, this is what is now an implementable form. So, one looks at the estimated quadratic discriminants scores of the respected population, and the one which is going to give us the maximum. We will have the new observation to be allocated to that corresponding to that maximum of the estimated quadratic discriminant scores. (Refer Slide Time: 27:55)



Sp. case : $\pi_i \equiv N_m(\underset{\sim}{\mu}_i, \Sigma)$ $i = 1(1)c$

$\Sigma > 0.$

The discriminant score for this setup is given by

$$= \underbrace{-\frac{1}{2}\log|\Sigma| - \frac{1}{2}\underset{\sim}{x}'\Sigma^{-1}\underset{\sim}{x}}_{\text{same for all the pop}^{\text{n}}\text{s}} + \underline{\underset{\sim}{\mu}_i'\Sigma^{-1}\underset{\sim}{x}}$$

$$\underline{-\frac{1}{2}\underset{\sim}{\mu}_i'\Sigma^{-1}\underset{\sim}{\mu}_i} + \underline{\log p_i}$$

Define the linear discriminant score as

$$\delta_i^{LDS}(\underset{\sim}{x}) = \underset{\sim}{\mu}_i'\Sigma^{-1}\underset{\sim}{x} - \frac{1}{2}\underset{\sim}{\mu}_i'\Sigma^{-1}\underset{\sim}{\mu}_i + \log p_i$$

Rule is to allocate $\underset{\sim}{x}$ to π_k if

$$\delta_k^{LDS}(\underset{\sim}{x}) = \max_i \delta_i^{LDS}(\underset{\sim}{x}).$$

Now, if we take a special case of the previous multivariate normal populations, if we have the special case that this pi i populations are multivariate normal m dimensional with mean vector as mu i. And the same covariance matrix for all the populations i equal to 1 to up to c, wherein the common sigma matrix is a positive definite matrix. So, we have got the c populations pi 1, pi 2, pi c, all multivariate normal with the mean vector being different for different populations, the covariance matrix sigma remaining the same for all of these populations.

Then the discriminant score as in the previous example, the discriminant score for this setup is given by following exactly the same type of approach, that we can look at p k f k x. Now these f k xs for the k populations will only differ by the corresponding mean vector quantity. And hence, the discriminate score is going to be given by minus half, this is log of determinant of sigma, because the sigma matrix was assume to be the same. And then if we split up that quadratic term there, then we will see that this is there is a term x transpose sigma inverse x, which does which does not vary from population to population. Because sigma matrix is same for us. And hence, this entire quantity here is going to be same for all the populations. So, we will have this quantity to be same for all populations.

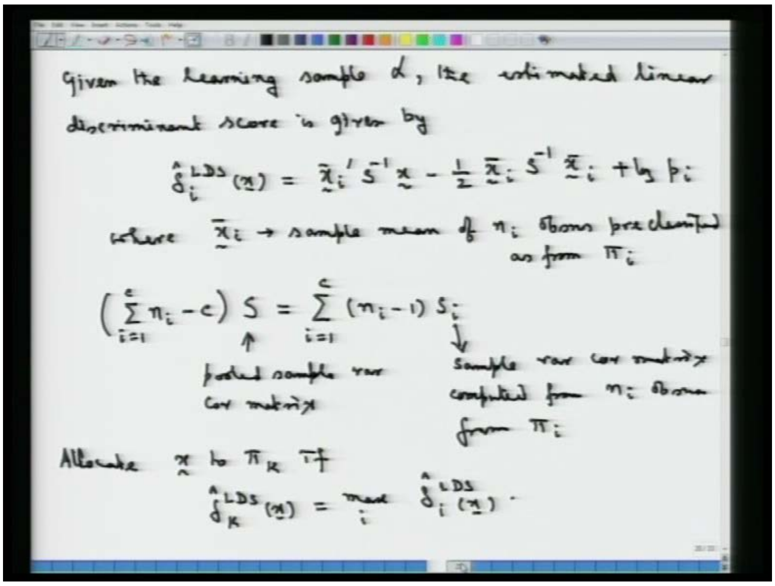
Terms that are going to be different is mu i transpose sigma inverse x, there will be a term which is twice here, because we will have this product here. Once we open up this term here, we will have an x transpose sigma inverse x, that is what the first term which is independent of k that is what we had written. And then, we will have this as x transpose mu k transpose sigma inverse x, and there will be a same term with a transpose coming out from x transpose sigma inverse mu k; however, there is a factor 2 in the denominator. So, eventually this 2 is going to be cancelled out, because we have a factor minus half sitting outside. This minus the term which is dependent on mu i vector. So, it is mu i transpose sigma inverse mu i and this plus this log of p i.

So, what we have cleverly done is to look at the simplifier version of this quadratic discriminant score, because the sigma matrix is same for all the populations. And hence, this is going to be in the present situation, going to be the same for all the populations. And the term x transpose sigma i inverse sigma i is sigma here, and hence the that term is going to be independent for the various choices of the population i.

So, the terms that are going to vary now is this term here, this term here, and basically this term here. So, what we observe is that now the discriminant score can be written as a linear discriminate score, because the quadratic term here is independent of the i th population. And hence, in this situation what one does is to define the linear discriminant score as say delta i linear discriminate score with respect to this observation x, which is now going to be given by this mu i transpose sigma inverse x minus half mu i transpose sigma inverse mu i this plus log of p i.

So, if this is the linear discriminant score corresponding to the i th population. We will allocate an observation x to pi k, if the linear discriminate score corresponding to that k th population is going to give us the maximum. And hence, we have the following discriminant classification rule is to allocate x to pi k, if we have delta k linear score with respect to x to be the maximum overall i of this delta i linear discriminant score with respect to all these x s. So, this is what is the rule, which is basically a linear discriminant score based rule.

(Refer Slide Time: 33:57)



Given the learning sample $\mathcal{L}$, the estimated linear discriminant score is given by

$$\hat{\delta}_i^{LDS}(\underset{\sim}{x}) = \bar{\underset{\sim}{x}}_i' S^{-1} \underset{\sim}{x} - \frac{1}{2} \bar{\underset{\sim}{x}}_i S^{-1} \bar{\underset{\sim}{x}}_i + \log p_i$$

where $\bar{\underset{\sim}{x}}_i$ → sample mean of n_i obsns preclassified as from π_i

$$\Big(\sum_{i=1}^{c} n_i - c\Big) \underset{\uparrow}{S} = \sum_{i=1}^{c} (n_i - 1) \underset{\downarrow}{S_i}$$

pooled sample var cov matrix

Sample var cov matrix computed from n_i obsns from π_i

Allocate $\underset{\sim}{x}$ to π_K if

$$\hat{\delta}_K^{LDS}(\underset{\sim}{x}) = \max_i \hat{\delta}_i^{LDS}(\underset{\sim}{x}).$$

Now, once again, we will have to look at this particular term here, and then look at the sample counter part of it. Sample counterpart in the sense that when we have got the learning sample, then based on the learning sample we will have to compute each of these quantities. Now note that mu I, mu i vectors that is specific to the i th population. And hence that is going to be computed from the n i observations, which belong to the pi i population; however, this sigma matrix we have assume to be the same for each of the populations, and hence this sigma matrix is going to be computed as the pooled sample variance covariance matrix based on n 1, n 2, n c observations.

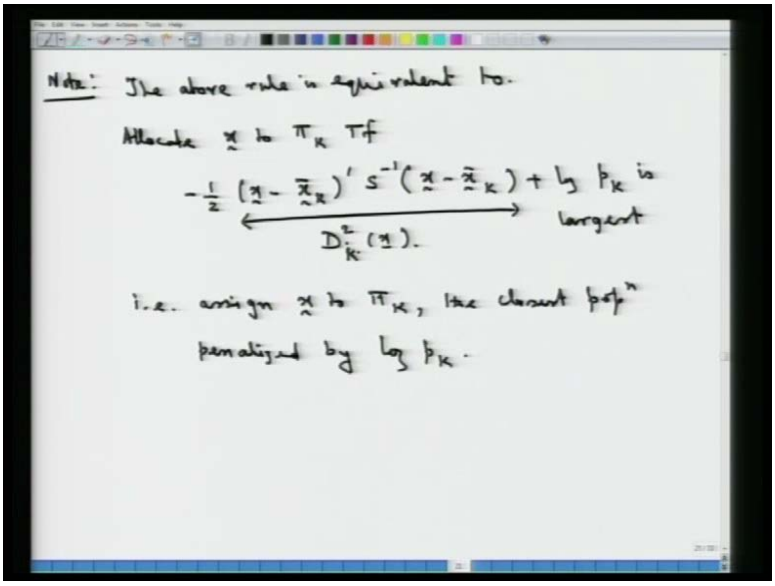
Now, once again given the learning sample, given this learning sample or learning set learning sample script l say the estimated linear discriminant score, the estimated linear discriminant score is going to given by say delta i hat l d s at x, that is going to be given by the corresponding sample quantities x i bar transpose s inverse. I am going to define what these terms are. x this minus half x bar i s inverse x bar i, this plus of p i, where this x i bar is the sample mean of n i observations, pre classified as from the pi i population. And this s matrix is going to be given by the following summation n i minus c i equal to 1 to up to c times, this s matrix is going to be given by the pooled n i minus 1 time s i i equal to 1 to up to c.

So, this is basically the type of pooled sample variance covariance matrix, when we have two population, we say that n 1 plus n 2 minus 2 times s, the s being the pooled sample variance covariance matrix, that is equal to n 1 minus 1 times s 1 plus n 2 minus 1 times s 2. And hence, this is what is the generalization of that. So, we have got this s matrix to be the pooled sample variance covariance matrix.

Now, what are these s i quantities, these s i quantities are the sample variance covariance matrix sample variance covariance matrix computed from the n i observations from pi i population. So, those are basically the sample variance covariance matrices from each of these pre classified n i observations from the pi i population.

So, based on this sample variance covariance matrix what we can say is that, we have the rule as to allocate x 2 pi k. If we have got this delta k cap linear discriminant score corresponding to this x to be equal to the maximum overall i of all such estimated linear discriminant scores at this point x here.

(Refer Slide Time: 37:59)



Note: The above rule is equivalent to.

Allocate $\underset{\sim}{x}$ to π_k If

$$-\frac{1}{2}\underbrace{(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_k)' S^{-1}(\underset{\sim}{x}-\bar{\underset{\sim}{x}}_k)}_{D_k^2(\underset{\sim}{x}).} + \log p_k \text{ is largest}$$

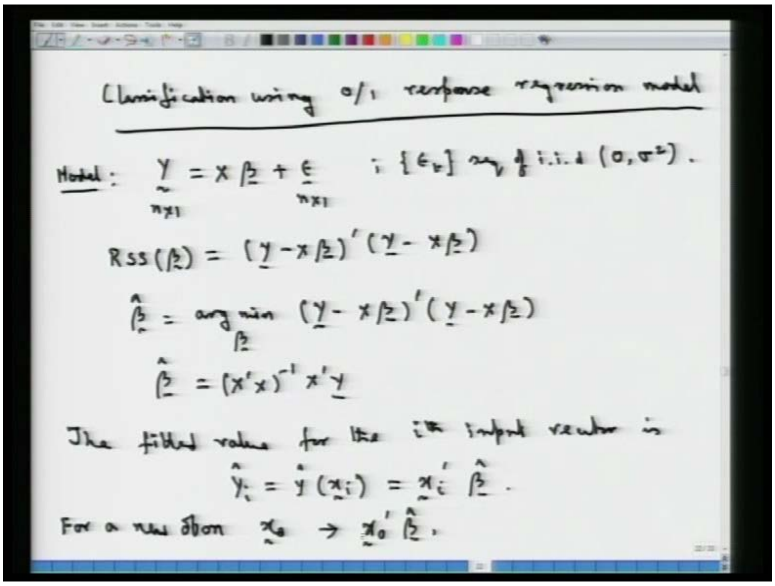
i.e. assign $\underset{\sim}{x}$ to π_k, the closest pop^n penalized by $\log p_k$.

Now, we just put it as a note that this linear discriminant score can be expressed as the distance, square distance actually. This above rule is equivalent to because of the simple and obvious reasons, is equivalent to say that allocate x to pi k. If we have got this minus half x minus x k bar say transpose s inverse x minus x k bar, this plus log of p k is largest this is. So, because what extra terms we have added here is an x transpose s inverse x, that is going to remain the same for each and every population k. And hence just by adding that particular term, we are able to express it in terms of this is basically the distance square d i square x, that is assign this x to pi k the closest population. This of course, closest population penalized by this particular term here penalized by log of p k term.

So, it is equivalent actually, because we the term that we have added out here is basically not going to vary from population to population. So, this is say d d k square, and hence we have got this to be having the interpretation that we are going to assign x to pi k, if that is that x is closest to the pi k population. So, this is about the type of classification rules that we are going to get, if we have the possible c populations to be multivariate normal either differing only by the mean vector or different both by the mean vector and the variance covariance matrix.

Now, let us now look at some other popular type of classification rules, the discriminant functions based on say 0 1 type of regression models classification models, which are going to be based on nearest neighbor classifiers. Then logistic discriminate functions.

(Refer Slide Time: 40:11)



Classification using 0/1 response regression model

Model: $\underset{n\times 1}{\underset{\sim}{Y}} = X\underline{\beta} + \underset{n\times 1}{\underline{\epsilon}}$; $\{\epsilon_r\}$ are of i.i.d $(0, \sigma^2)$.

$$RSS(\underline{\beta}) = (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta})$$

$$\hat{\underline{\beta}} = \arg\min_{\underline{\beta}} (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta})$$

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$$

The fitted value for the i^{th} input vector is

$$\hat{y}_i = \hat{y}(\underset{\sim}{x}_i) = \underset{\sim}{x}_i' \hat{\underline{\beta}}.$$

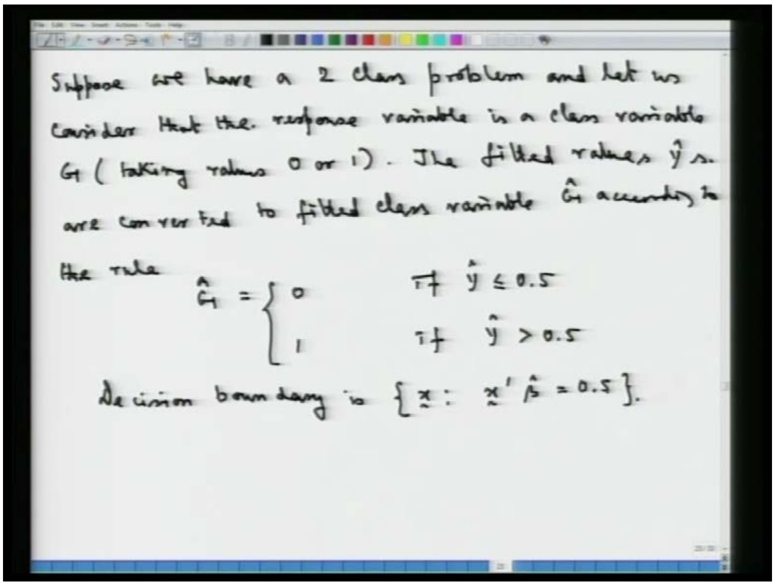
For a new obsn $\underset{\sim}{x}_0 \rightarrow \underset{\sim}{x}_0' \hat{\underline{\beta}}$.

So, let us start looking at those special cases. The first one that we are going to look at is classification model using a 0 1 response regression model. This is a very simple and intuitive way of looking at the classification problem. Suppose, we represent the regression model as the following in the standard multiple regression model that it is x beta plus epsilon.

Now we have got here. Suppose this is n by 1 and hence this epsilon also is n by 1. And we have got epsilon t, this is the sequence of say i i d with mean 0, and some variance equal to sigma square term. Then we all know that the residual sum of squares with respect to this unknown vector regression vector beta is given by y minus x beta transpose y minus x beta. And then, beta hat which is the argument minimum with respect to beta of this residual sum of squares. So, beta hat is one that minimizes this y minus x beta transpose y minus x beta is going to be given by x transpose x inverse x transpose y.

So, we have got this beta hat to be come to this. Now corresponding to this beta hat, the fitted value for the i th input vector is say given by y i hat, which is equal to say y hat of this x i vector that is the i th input vector. It may have a constant 1, in the first entry if this model has got the constant term and it will not be having that if we do not have a constant term there. So, this is going to be given by x i transpose times this beta hat. So, this is what we have at the i th input vector for a new observation x naught, the predicted value is going to be x naught transpose times this beta hat.

(Refer Slide Time: 42:53)



Suppose we have a 2 class problem and let us consider that the response variable is a class variable G (taking values 0 or 1). The fitted values $\hat{y}$ s. are converted to fitted class variable $\hat{G}$ according to the rule

$$\hat{G} = \begin{cases} 0 & \text{if } \hat{y} \leq 0.5 \\ 1 & \text{if } \hat{y} > 0.5 \end{cases}$$

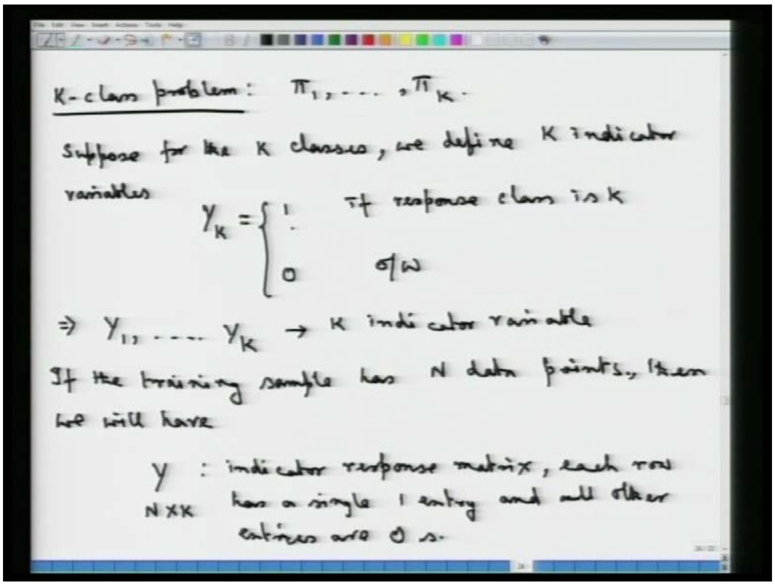
Decision boundary is $\{ \underset{\sim}{x} : \underset{\sim}{x}' \hat{\beta} = 0.5 \}$.

Now, based on this value of x naught transpose beta hat, we are going to give a classification rule. Suppose we have got a two class problem that is there are two possible populations pi 1 and pi 2. And let us consider that let us consider that the response variable is a class variable say denoted by g taking values either 0 or 1, the fitted value y hats are converted to fitted classes in the following way, class variable say g hat according to the rule say, I put g hat to be equal to 1 or 0, this is going to be 0. If we have the predicted y hat to be say less than or equal to 0.5. Now this is nothing special about 0.5, one can vary that also. So, this is equal to 1, if we have got y hat to be greater than 0.5.

So, such a type of rule, which is going to be based on the predicted value y hat. And then from that continues scale y hat, what we have observed. I f we look at, using that y hat to frame this g hat, the fitted class variable to take the value 0. If y hat is less than or equal to 0.5 and is equal to 1, if y hat is greater than this, then based on this regression model and a new given observation x naught a value of this being as at a predicted value y hat say, we will be able to then give the class membership 0 or one based 1, the predicted value of y hat.

This is the simple rule, this has got a decision boundary given by the following. So, the decision boundary is basically the set of x s, such that we have got x prime beta hat that equal to 0.5. One can vary this particular constant here, in order to get other type of decision boundary. It is a simple rule of what one assumes that the response variable is zero one then apply a simple multiple regression type of setup, and then get to the predicted class membership through the mapping as what we have given in here.

(Refer Slide Time: 46:19)



K-class problem: $\pi_1, \ldots, \pi_K$.

Suppose for the K classes, we define K indicator variables

$$Y_K = \begin{cases} 1 & \text{if response class is } K \\ 0 & \text{o/w} \end{cases}$$

$\Rightarrow Y_1, \ldots, Y_K \rightarrow$ K indicator variable

If the training sample has N data points, then we will have

$\underset{N \times K}{Y}$: indicator response matrix, each row has a single 1 entry and all other entries are 0 s.

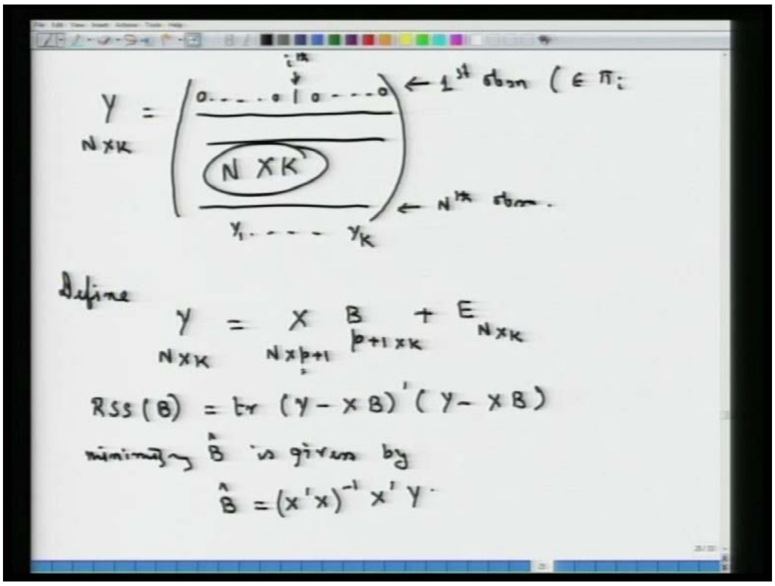
Now, this is simple concept here, what we have for this 0 1 response variable can be generalized for a k class problem. Let see that particular generalization to the k class problem. Suppose, we have got a k class a k class problem meaning thereby we have got h populations pi 1, pi 2, pi k, these are the k possible populations. And then, suppose under such a setup, we are going to have the following. Suppose for the k classes, we define k indicator variables in the following way that we have got a y k, the k th indicator variable taking the value 1 and 0, 1 if response class is k and is equal to 0, if it is otherwise.

So, we will have the corresponding k indicator variables y 1, y 2, y k. These are the k indicator variables, which are going to give us the class memberships. Now if we have the training sample consisting of n data points, if the training sample has n data points, then what we are going to have, then we will have corresponding to these k defined indicator variables, a matrix y which is going to be an n by k matrix. So, this is the indicator response matrix.

(No audio from 48:30 to 48:34)

Now, each row here, each row has a single 1 entry, and all other entries are 0, entries are 0, why is that so, because if we have defined this y y k; k th indicator variable to take the value equal to 1, if the responses in class k, and 0 if otherwise.

(Refer Slide Time: 49:14)



Define

$$\underset{N\times K}{Y} = \underset{N\times p+1}{X}\ \underset{p+1\times K}{B} + E_{N\times K}$$

$$RSS(B) = tr\,(Y - XB)'(Y - XB)$$

minimizing $\hat{B}$ is given by

$$\hat{B} = (X'X)^{-1}X'Y.$$

And then, let us look at this y; this is this y matrix this is an n by k matrix. So, how is that going to look like? Now these are corresponding to this is the first observation in the training sample, and this is the n th observation in the training sample. Now, we have here corresponding to the first observation, the k indicator variables. Now suppose that the first observation is belonging to say i th, suppose this is belonging to pi I, then the y i corresponding to that this if this is the i th position. Then the y i corresponding to that would be equal to 1 and all others will be equal to 0.

Similarly, for the second observation here, if that is belonging to some pi i prime, then at the i prime position this vector this row vector will contain the 1 entry and all other positions will have 0 entries. So, each of these rows of this matrix that we are going to have, which are going to be based on those y 1, y 2, y k the k indicator variables. So, for the j th observation in general, we will have the corresponding entry in the column of y 1 y 2 y k this particular row to be equal to 1. If the j th observation is belonging to that particular population in the training sample. And hence, this is how we are going to get this n cross k matrix of 0 s and 1s, wherein we will have 1, a single entry 1 in each of these rows here.

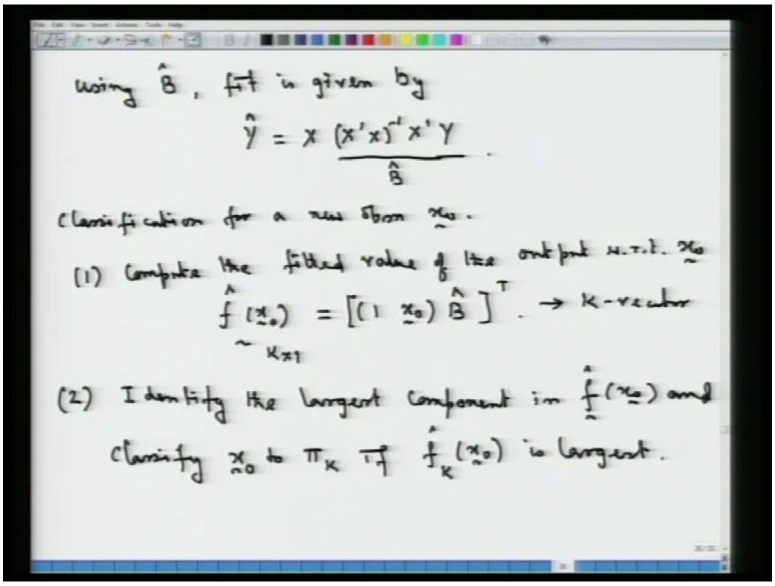
So, the entire data using the indicator, the k indicator variables y 1, y 2, y k. We are able to write it in terms of this n cross k matrix, which is the indicator response matrix. Each row of this particular matrix y has a single 1 entry and all other entries are 0 that is because of this simple reason out here.

Now, what we are going to write is the following, define this model through this matrix y n cross k, which is going to be given by an x matrix, which is now having say p plus 1 columns n rows corresponding to the n observations of this data. So, this we have n cross p plus 1, we have taken as say a constant term here. So, that the first column of this x matrix would be constant. And hence, people ask 1 and we assume that there are be more explanatory variables, explanatory in the sense of giving us some clues about or rather explanatory for the corresponding response variable. And this b is a matrix of unknown constants, which is of dimension p plus 1 rows and k columns. And then, corresponding to this y matrix here, we will have an e matrix here which is the matrix of random variables, which is n by k.

Now, if this is the part, if this is a explanatory part and if this is a response matrix part. This is in the matrix regression setup, where b matrix p plus 1 by k is the matrix of constants. And then, we define the residual sum of squares with respect to this b matrix as the trace of this y minus x b transpose y minus x b matrix. So, this is what the residual sum of squares now would be, and the minimizing b matrix matrix of constant which would minimize this r s s b. Minimizing this b hat is given by it is basically corresponding to a particular column of this y matrix that we are going to have this is given by this b matrix as in the before. We will have x transpose x inverse x transpose this y matrix.

Now, as you can see from this form here, if this matrix y, y matrix is consisting of these k columns. And if you consider the i th column of this y, then this matrix x transpose x inverse x transpose being multiplied with the i th column of this y matrix is going to give us the i th column of this b hat matrix.

(Refer Slide Time: 54:04)



using $\hat{B}$, fit is given by

$$\hat{Y} = X \underbrace{(X'X)^{-1} X' Y}_{\hat{B}} .$$

Classification for a new obsn $\underset{\sim}{x_0}$.

(1) Compute the fitted value of the output w.r.t. $\underset{\sim}{x_0}$

$$\underset{\sim}{\hat{f}}(\underset{\sim}{x_0})_{K\times 1} = \left[(1 \;\; \underset{\sim}{x_0})\, \hat{B}\right]^T . \rightarrow K\text{-vector}$$

(2) Identify the largest component in $\underset{\sim}{\hat{f}}(\underset{\sim}{x_0})$ and classify $\underset{\sim}{x_0}$ to π_K if $\hat{f}_K(\underset{\sim}{x_0})$ is largest.

So, once we have got this b hat matrix. Using this b hat, what we can get, fit is given by y hat which is equal to x times x transpose x inverse x transpose this y matrix. This basically is that our b hat. So, this is x times b hat. So, this is the fitted value.

Now, let us see that if we have really now a k class problem. If a new observation now comes, how we are going to actually look at classifying that particular observation into one of the k possible classes. So, using this set up, under this setup classification for a new observation x naught is going to be done in the following steps. In the step 1, we compute the fitted value, the fitted value of the output, this is with respect to this x naught variable. Now remember, we had a constant term. So, let us denote this output fitted output by f x naught vector. Now this itself is going to be a vector because if we look at how that is going to be obtained then this is nothing, but 1 is the constant term into this x naught. So, that is augmented by this term that multiplied by this b hat vector. And then, we are going to take the transpose of this particular quantity.

So, this is going to be a k vector. Now if this is a k vector, then corresponding to the i th indicator variable. This is going to give us the i th entry of this k by 1 vector, and hence what we are going to do is to identify the largest component in this f hat vector x naught. And classify x to x naught 2 pi k if we have say the k th component of this f hat x naught vector is largest, the logic is simple, because we were looking at y 1, y 2, y k to be the k indicator variables.

The value for which that is going to be equal to 1, in the population was the class membership. And from the sample, if we are going to look at the fitted value corresponding to this x naught. Then this is going to be this k vector, k vector corresponding to the corresponding fitted values of y 1, y 2, y k. And then for the component for which we have the largest entry in this particular vector, we are going to classify x naught to pi k, if that corresponding entry is going to be the largest.

So, this section here, the small section here wherein we just try to use this 0 1 type of response variable type. So, it is classification based on 0 1 response regression models. We have seen that how simple that classification rule is based on a simple regression model, when we have got a 2 class problem as in this. And we also have a k class problem in general.

So, I will stop at this particular point in this lecture. In the next lecture, what we are going to look at is nearest neighbor classifiers. And we are also going to look at a logistic discrimination rule. Thank you.