Indian Institute of Technology Madras
Presents

NPTEL
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING

Pattern Recognition

Module 02

Lecture 02

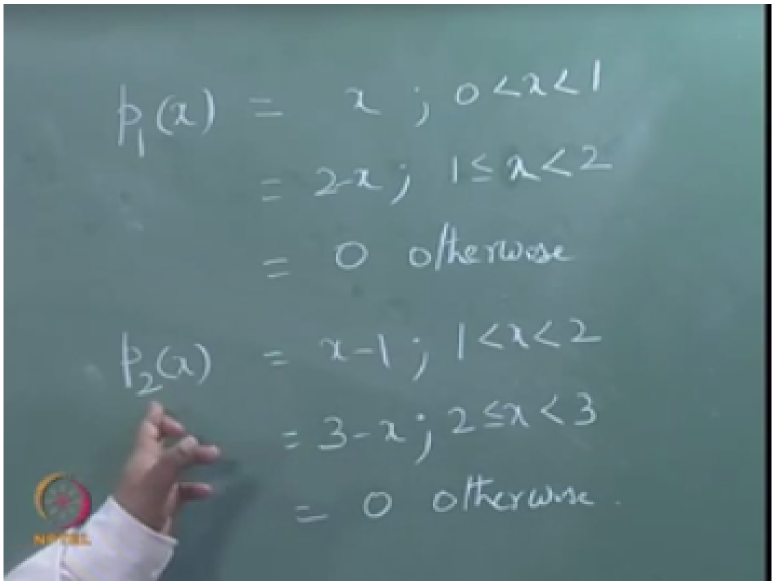
Example of Bayes Decision Rule

Prof. C. A. Murthy

Machine Intelligence Unit,
Indian Statistical Institute, Kolkata

I gave you the base decision role probably it is look completely theoretical to you I will try to go give some example so that it will be more clear view.

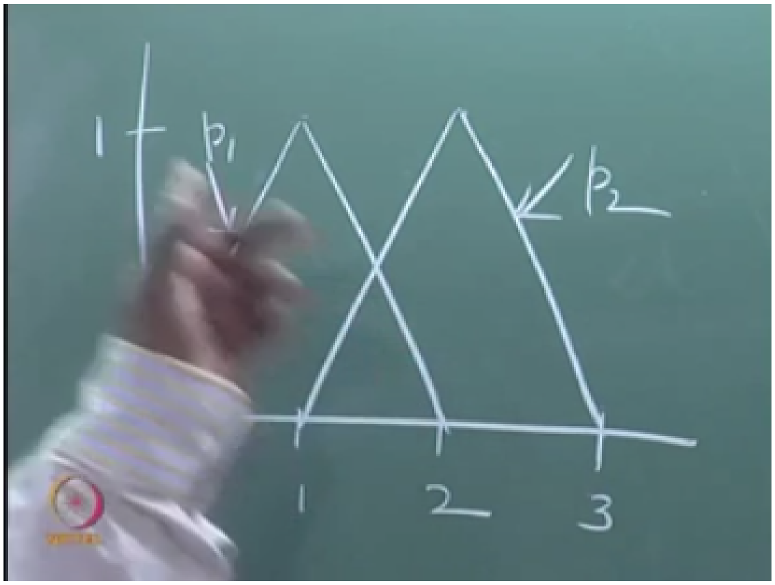
(Refer Slide Time: 00:28)



$$
\begin{aligned}
p_1(x) &= x \,;\ 0 < x < 1 \\
&= 2-x \,;\ 1 \le x < 2 \\
&= 0 \text{ otherwise} \\
p_2(x) &= x-1 \,;\ 1 < x < 2 \\
&= 3-x \,;\ 2 \le x < 3 \\
&= 0 \text{ otherwise}
\end{aligned}
$$

So I will write here let us take the number of class is 2 and the 1st density function is this is one density function, and let me give you another density function P1(x) = x and x lies between 0 and 1, 2-x when x lies between 1 and 2 0 otherwise P2(x) = x-1x lies between 1 and 2 3- x, x lies between 2 and 3, 0 otherwise let us see how this density functions look like okay since I will be using this part of the board for some other calculation let me us this part of the board.

(Refer Slide Time: 02:12)

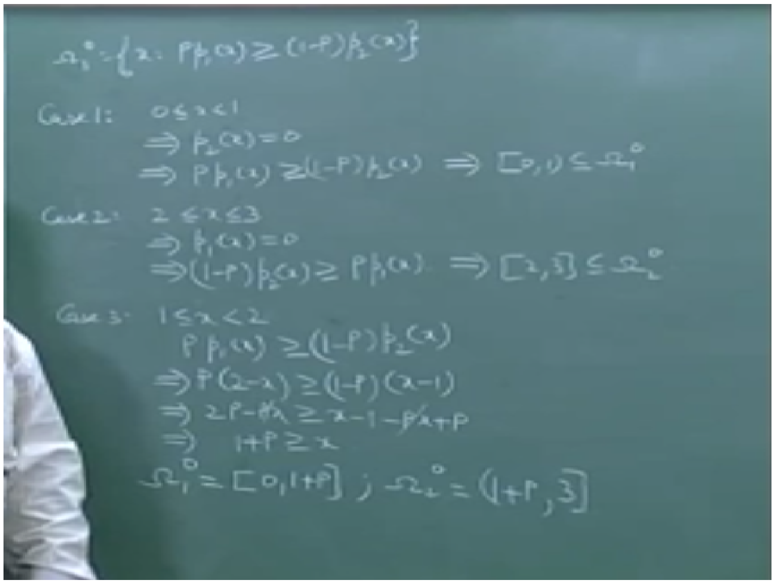


This is x0, x1 this is 2, 3 between 0 and 1 the value is x so if I write this as 1 this is the straight line at 0 it is actually going to 0 at 1 the value is 1 look at this and from 1 to 2 it is 2-x so in fact we can include this also no problem and 0 otherwise okay and look at this is x-1 let us also include this one this is x -1, when x lies between 1 and 2 that means that x =1 the value is 0 at x = 2 the value is 1,. So this again and this is 3 – x again at x = 2 the value is 1 and x = 3 the value is 0.

So we will use a straight line and 0 otherwise if you look at the mathematics that I gave you where Ω is the subset of RL that is what is wrote here our Ω is the subset of R real line in fat what is our Ω here the Ω is the set 0 to 3 because outside that the values are 0 for both the functions okay so here Ω is 0 to 3, now see the prior property of this class is P I can write this thing as P1, this is P2.

Prior property of the 1st class is P prior property this class is 1 – P sum of these 2 things is 1 okay as you can see there is overlap between the classes now these are the density functions.

(Refer Slide Time: 05:06)



$\Omega_1^0 = \{x : P p_1(x) \geq (1-P) p_2(x)\}$

Case 1: $0 \leq x < 1$

$\Rightarrow p_2(x) = 0$

$\Rightarrow P p_1(x) \geq (1-P) p_2(x) \Rightarrow [0,1) \subseteq \Omega_1^0$

Case 2: $2 \leq x \leq 3$

$\Rightarrow p_1(x) = 0$

$\Rightarrow (1-P) p_2(x) \geq P p_1(x) \Rightarrow [2,3] \subseteq \Omega_2^0$

Case 3: $1 \leq x < 2$

$P p_1(x) \geq (1-P) p_2(x)$

$\Rightarrow P(2-x) \geq (1-P)(x-1)$

$\Rightarrow 2P - \cancel{Px} \geq x - 1 - \cancel{Px} + P$

$\Rightarrow 1+P \geq x$

$\Omega_1^0 = [0, 1+P]\ ;\ \Omega_2^0 = (1+P, 3]$

Now I will start now here now Bayes decision rule what is the Bayes rule our Ω1 0 is the set of all x for which p times P1 x is greater than = I am putting the equal to here 1 – P x P2x this is Ω10 okay now let us have some cases, case 1 so x lies between 0 and 1 when x is lying between 0 and 1 but density of the certain class is 0 right look at that one the density of the certain class is 0 to 1 the density value is 0 okay.

So this implies P2x = 0 okay so this implies actually P times P1 x is this is either >0 or = 0 so this is always ≥ – P times P2x right so the P when x is lying between 0 and 1 this as to go to class 1 so this actually implies that this set 0 to 1 at 0 it is closed at 1 this is open it is the subset of Ω10 right now let us take case 2, 2 ≤ 2 < x≤ 3 okay I can as well include equal to also no problem.

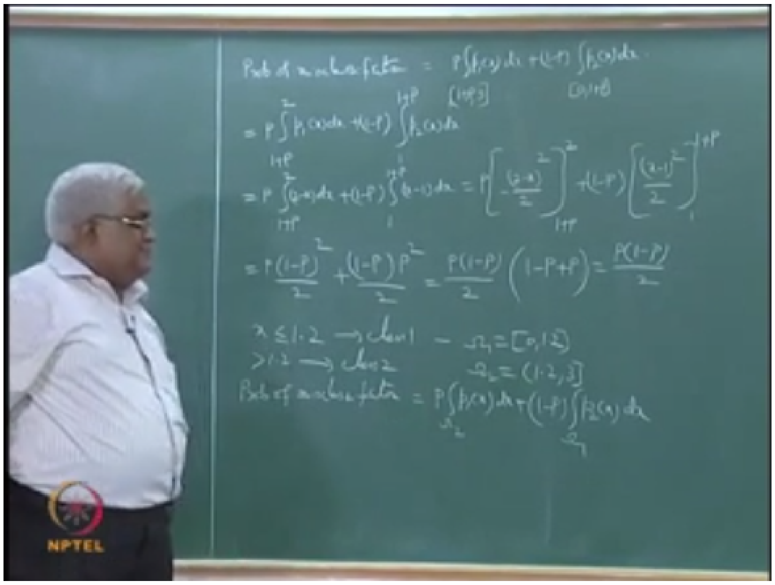
I can include equal to also nor problem here what will happen here when x is lying between 2 and 3 this actually implies that P1 x = 0 is it true at x = 2 the value is 0 and < 2 also the value is 0 so P1x is 0 here this implies 1- P times P to x, P Times P1x so this whole set this will go to class 2 now let us have the main case that is the case 3 0 to 1 so 1 ≤ x < 2 okay now let us say P times P1x it is ≥ 1 – P times P2x this implies P times what is the value of P1x, P1x is x now P1x is 2 – x 1 – P times this is x -1.

P2 x is x -1 okay P2x is x – 1 so this implies 2P – Px is ≥ x- 1- Px+ P, this –Px – Px gets Canceled and 2P – P is P x is ≥ 1 + P 2P – P is P this 1 is coming this side so 1 + P is ≥ x.

Now from these 3 cases what can we say about Ω10 so basically Ω10 is 0 to 1 and 1 to 1 + P, 0 to 1 and 1 to 1+ P so this is basically 0 to 1+ P, 0 to 1 here 1 to 1+ P so this is 0 to 1 + P and what is Ω20 1 + P2 to 2 and 2 to 3 here you have 1+ P to 2 here it is 2 to 3 so this is basically 1+ P to 3 is it cleat when x≥ 1+ P you will put it in class 1 when x is greater than 1 + P you will put in class 2 so here x > 1+ P if it goes to class 2up to what up to 2.

Now you might be having a question is 1+ P <2 always it is <2 because this P≤ 1 this P cannot be greater than 1,s o 1+ P to 2, 2 to 3 so this is 1+ P to 3 so these are the this is the reason for class 1, this is the reason for class 2 now let us calculate the probability of miss classification.

(Refer Slide Time: 12:26)


Prob of misclassification = $P\int_{(1+P,3]} p_1(x)dx + (1-P)\int_{[0,1+P]} p_2(x)dx$
$= P\int_{1+P}^{2} p_1(x)dx + (1-P)\int_{1}^{1+P} p_2(x)dx$
$= P\int_{1+P}^{2}(2-x)dx + (1-P)\int_{1}^{1+P}(x-1)dx = P\left[-\frac{(2-x)^2}{2}\right]_{1+P}^{2} + (1-P)\left[\frac{(x-1)^2}{2}\right]_{1}^{1+P}$
$= P\frac{(1-P)^2}{2} + \frac{(1-P)P^2}{2} = \frac{P(1-P)}{2}(1-P+P) = \frac{P(1-P)}{2}$
$x \leq 1.2 \rightarrow$ class 1 $-$ $\Omega_1 = [0, 1.2]$
$> 1.2 \rightarrow$ class 2 $\Omega_2 = (1.2, 3]$
Prob of misclassification $= P\int_{\Omega_2} p_1(x)dx + (1-P)\int_{\Omega_1} p_2(x)dx$
NPTEL

Probability of miss classification = P times integral P 1x over this set 1+ P to 3 + 1- P times integral P to x 0 to 1 + P look at the expressions that you have the expression is P1, P1x over Ω2 so this is the prior property of class 1 multiplied by this over Ω2 that is 1+ P to 3 the P2, P2 over

Ω1so to 1+ P this is = P times integral P, P1x is 0 when >= 2 s o basically we need not have to take the whole of 1 + P to 3.

This is right + here again P2 is 0 form 0 to 1 this is basically 1 to 1 + P so this = P times integral 1+ P to 2 what is P1, P1 is 2- x what is P2, P2 is x – 1 P2 is x -1 now this is = P times what is the integral of 2 – x this will be 2-x^2/2 and there is a negative sign this whole thing this you need to evaluate it at 2 and evaluated 1 + P from this and you need to subtract and this one integral of x -1 this will be x-1^2 /2 this is also from 1 to 1 + P.

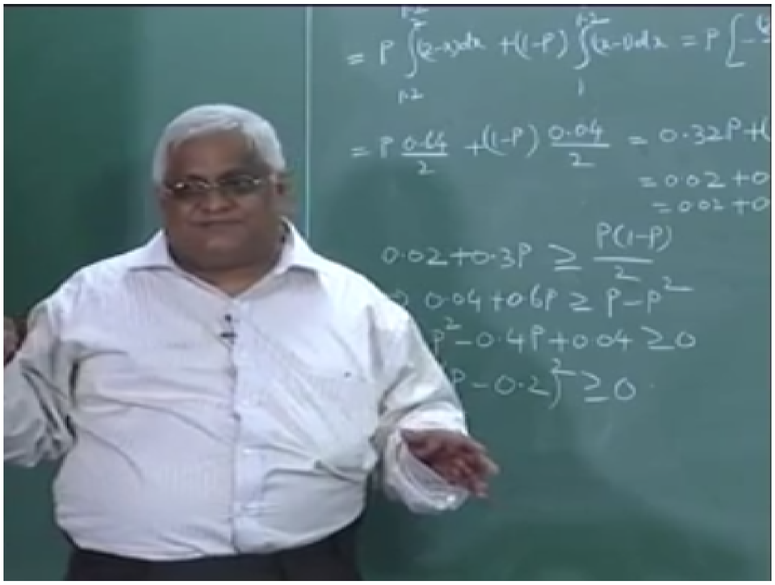
So this will be equal to P times at to the value 0 so this will be 2-1, -P so this is $(1\text{-}P)^2/2$ this negative at to the value is 0 minus and there is again a minus so this is going to be positive, +(1-P) at 1+P the value is $P^2/2$ at one value is 0, and if you take P(1-P)/2 as common what you are going to get is here you are going to get 1-P and here you will get P.

So this will be P(1-P)/2, so this is base error probability that means if you take any other decision rule it is error probability will be greater than or equal to this probability, so now you give me a decision rule one of you, please give me a decision rule, okay I will give one then afterwards you try something else. Say my decision rule is x≤1.2 I will put it in class 1 okay, and ≥ 1.2 I will put it in class 2.

Either decision rule, I just taken some decision rule x≤1.2 I will put it in class 1 ≥1.2 I will put it in class 2. Note that in this decision rule there is a P here but in this rule there is no P, there is no prior probability, x≤1.2 I am putting in class 1, ≥1.2 I am put it in class 2. Now let us find the probability of error I will probability on misclassification for this rule, so that means here our Ω1 is 0 to 1.2 and Ω2 is 1.2 to 3.

Ω1 is 0 to 1, Ω2 is 1.2 to 3, now what is the probability on misclassification here the probability on misclassification is again P (P1x)/Ω2+ (1-P) (P2x)/Ω1, so this is the probability of misclassification, now what is this value this is equal to.

(Refer Slide Time: 20:06)



P times Ω2 is 1.2 to 3 +(1-P) times Ω1 is 0 to 1.2, here again the same thing happens we can take 1.2 to 2 and here 1 to 1.2 right, and this is equal to P time again P1x= how much 2-x and P2x=x-1 and this is equal to P times again let us do the same thing this here it is 1.2 to 2, this is $(x-1)^2$ this is 1 to 1.2, this is equal to P times again at to the value is 0 at 1.2 the value is 0.8^2, this is 0.64/2 okay, + 1.2 the value is $(0.2)^2$ this is 0.04/2.

Okay, so this is equal to 0.32P+ this is 0.02 okay, so this is the probability of misclassification. Now I claim that this is less than or equal to this you have any question. So why did I take it to 2, look P1x look at the definition P1x, anything greater than to the value is 0 so we will not have to take anything from 2 to 3 that is why it is 1.2 to 3.

Here also 0 to 1 we need not take because 0 to 1 the value is 0, 0 to 1 the value is 0 for P2 so we need to take only 1 to 1.2, so this is 0.02+0.32P-0.02P right, this is equal to 0.02 0.32P- this one this is +0.3P am I right, 1x0.02 and this is 0.32P and this is -0.02P3.2- this one this 0.3P right. Now I claim that this is greater than or equal to this let us see whether it is true or not, because that has been my whole claim that you take any other decision rule you calculate its misclassification probability.

That means classification probability has to be greater than or equal to the misclassification probability of base decision rule. Now I claim that this is greater than or equal to this, let us see whether it is true or not. 0.02+0.3P is ≥ this is true if and only if might take this 2 on this side

0.04+0.6P is ≥ this is P-P^2 right, this is if and only if bring this thing on this side this is P^2+P becomes –P, -P+0.6P this is -0.4P, right.

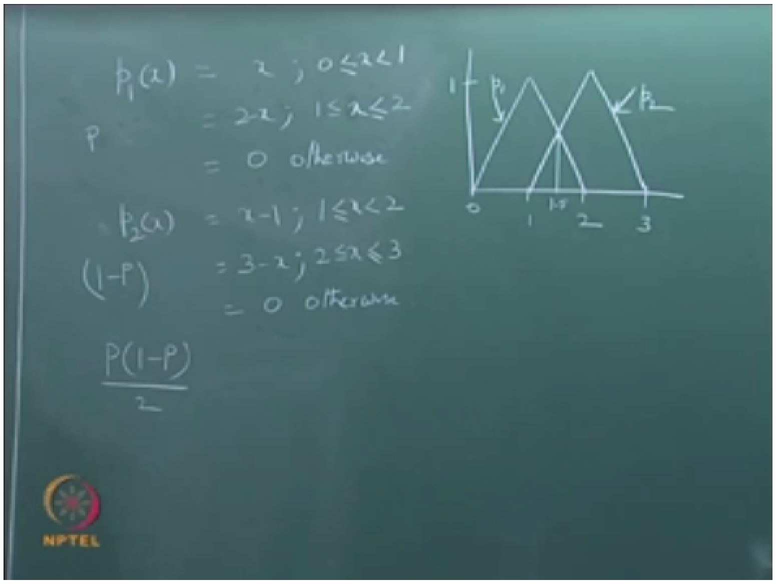
And this is +0.04 this must be ≥ 0, if this ≥ 0 is this ≥ 0 yes or no, yes why, you take any decision rule, take any dame decision rule find its misclassification probability that will be ≥ the misclassification probability of base decision rule, this is true. I have just taken one rule you take any decision rule I mean it just does not matter what your decision rule will be it just does not matter you will be able to show this thing, you will be able to show this.

Okay, in fact in your homes you can try with other such Ω1, Ω2 you will get some such thing and sometimes you will get some such thing plus some positive quantity it depends on what decision rule you have taken. Suppose you take a decision rule such as 0 to 1.1 and 1.7 to 3 class 1, the rest class 2 you can take that 0 to 1.1, 1.7 to 2 that is class 2 the rest is I mean class 1 the rest is class 2.

You can take even opposite also 0 to 1.1 at to 0.7 to 2, that is for class 2 and the rest is class 1 that also you can take. Then also you will get a square term always you are going to get a square term like this and plus some constant, constant is a positive value, this is something that you will find it for this example this is only true for this example which some other distributions you will get some other such things.

It is not always true that you will get a square term like this okay, it is not always true that you will get a square term like this but you are in a position to show always that the probability on misclassification for any other decision rule will be greater than or equal to that of the base decision rule. You see we found the base error probability as P(1-P)/2 what is the maximum value of this.

(Refer Slide Time: 29:58)



$P \quad p_1(x) = x;\ 0 \le x < 1$

$= 2-x;\ 1 \le x \le 2$

$= 0$ otherwise

$(1-P) \quad p_2(x) = x-1;\ 1 \le x < 2$

$= 3-x;\ 2 \le x \le 3$

$= 0$ otherwise

$$\frac{P(1-P)}{2}$$

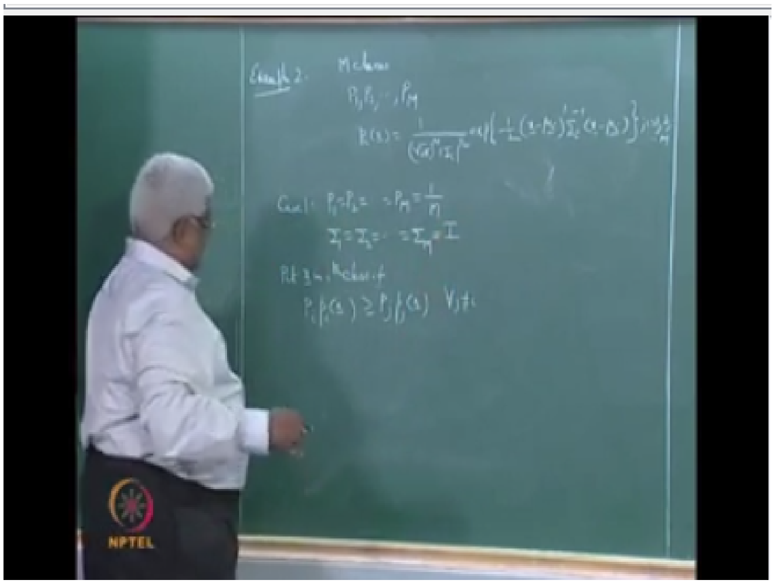
The maximum value of this is when P=1/2 right, when P=1/2 you will get the maximum value and the maximum value is 1/8, P=1/2 means what both the classes they have the same probability of occurrence and the base decision rule says this is 1.5 that means you sort of have the maximum uncertainty here.

If one class has more probability of occurrence then you can put more points in that class, you can put more points in that class but if both the classes have the same probability of occurrence then you have the sort of maximum uncertainty and that is what is happening when P=1/2, what is the minimum value of this, minimum value is 0 when P=0 you will get 0, when P=1 you will get 0, P=0 means what, P=0 means the whole class is a second class, the whole class is second class and you will look at your decision rule P=0 means 1 to 3 you are going to put it class 2.

And P=1 means the whole class is first class and 0 to 2 you will put it in class 1 there, look at your decision rule okay. So this is again true only for this, you have many other distribution you are going to get different formulas for different distributions. This distribution is known as triangular distribution because the shape is a triangular, so it is known as triangular distribution. You might ask me why I have taken triangular distribution there are many other distribution there are normal distribution why did I take triangular distribution?

Well I am going to deal with normal distribution now and number 2 these distributions are easy to handle I can calculate all those integrals, there are normal distribution integrals there are slightly difficult to handle. So now my next example is basically normal distribution example, so this is the example 2.

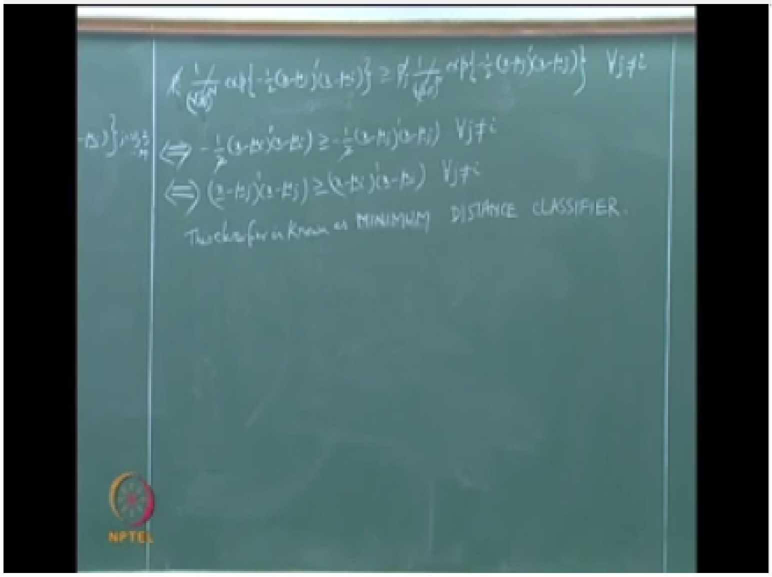
(Refer Slide Time: 32:57)


Example 2. M classes
$P_1, P_2, \ldots, P_M$
$p_i(x) = \frac{1}{(\sqrt{2\pi})^n |\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\}\ i=1,2,\ldots,M$
Case 1: $P_1 = P_2 = \cdots = P_M = \frac{1}{M}$
$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_M = I$
Put x in i th class if
$P_i p_i(x) \geq P_j p_j(x)\ \forall j \neq i$
NPTEL

This is the example 2 let us say you have M number of classes and each of them is multi variant normal okay and prior probabilities are P1, P2 upto Pm, and the density function is 1/√2π^n σi ½ - ½ x- μi' σi x –μi this is I = 1 to m. σi is the varinace3 co variance matrix of the multi variant normal μi is the mean vector of highest multi variant normal and capital P is the corresponding higher probability of ith class.

Now here let us take a case, the case is case 1, let us assume that P1 = P2 = Pm this is = 1/m, let us assume that all prior probabilities are equal okay and let us also assume that σ1 = σ2 = σm this is identity matrix. All prior probabilities are equal and this is diagonal matrix, now let us see what exactly we are going to get. We will put x in the ith if Pi x ≥ Pj x for all the V0 = I okay, Pi x ≥ Pj x V0 = I then put x in the ith class now what does this going to give you here I will write here.

(Refer Slide Time: 36:08)



$$P_i \frac{1}{(\sqrt{2\pi})^n} \exp\left\{-\frac{1}{2}(x-\mu_i)'(x-\mu_i)\right\} \geq P_j \frac{1}{(\sqrt{2\pi})^n} \exp\left\{-\frac{1}{2}(x-\mu_j)'(x-\mu_j)\right\} \quad \forall j \neq i$$

$$\iff -\frac{1}{2}(x-\mu_i)'(x-\mu_i) \geq -\frac{1}{2}(x-\mu_j)'(x-\mu_j) \quad \forall j \neq i$$

$$\iff (x-\mu_j)'(x-\mu_j) \geq (x-\mu_i)'(x-\mu_i) \quad \forall j \neq i$$

This classifier is known as MINIMUM DISTANCE CLASSIFIER.

Capital P_I determinant of σi σ is identity matrix determinant is 1 + √ is also 1 so I am not going to write it, so I am not going to write it okay exp of ½ x - μi' and σi is the identity matrix inverse it is itself identity matrix and multiplication identity matrix does not change anything so I am just going to write this okay, this is ≥ Pj. You might think that this is a very complicated expression actually it is not.

You will see it now this Pi is same as Pj we can cancel it this $1/\sqrt{2}\pi^n$ this there on both sides we can cancel it, the exp -½ ≥ exp -1/2 means if you log algorithm on both sides this is same as -1/2 I am right log than increasing confirmation and I can always apply algorithm that is same as removing the exp term because I have cancel this and this so I can remove the exp this term also it is going to look like this okay.

Then I will cancel this 2 so what is this x - μ' ≥ you put x1 i^{th} class is Pi, Ps ≥ Pj Px and this is same as put x in i^{th} class if x - μi' x – μi is < x – μj for r = i, what is the meaning of this? Is not the case that we are calculating the distance between x and μ and i, and μi is a mean. Os you will put x in i^{th} class if the distance between x and the mean of the i^{th} class is distance of mean of all the other classes then you put x in i^{th} class there is the standard name for this classifier, this classifier is known as minimum distance classifier.

You take x you have these many means calculate distance with x of these means find where ever it is minimum put x in that class and this is derivation of minimum distance classifier. This classifier is known as minimum I think I will write in capital letters distance classifier, this classifier is known as minimum distance classifier this is one of the standard classifiers that you would seen in any pattern book in fact books they start the classifications with this classifiers.

You find the mean and see where to put to class where the mean is there the distance is minimum, many old books they start with this classifier. You might be wondering or you might have thought what the proof of this is, I mean how is it coming? You see the way it is coming note that this classifier is one of the widely used classifiers in patron or and this is the best classifiers under these assumptions.

All the prior probability are equal and we are assuming that distributions are all normal that is the 2nd assumption and under that assumption we are assuming that all the co variance matrices they are equal to the identity matrix then this classifier is the best classifier look at the number of assumption that we have made, 1 the distributions assumption normal distribution, 2 the p0rior probability being same, 3 all the co variance matrices are equal not only equal to identity matrix.

This is a vast range of assumptions and the vast range of constrains under that this is the best classifiers and in fact many times we use this classifiers without really knowing probably all these details, we use this classifiers. Now you might be having a valid question the question is how we know that data that is given to us it follows normal distribution that is the very basic question, how do we know that data that is given to us is normal distribution number 1?

Number 2 if we know that it is the normal distribution how do we get to know the mean of the distribution and the variance co variance matrix that is the 2nd one. Now these questions I will try to answer them partially why only a partial answer? The reason is that we are given a data sets which as finite mini quantity, on the base of this finite observation in general it is very difficult for you to say whether the distribution is normal or not?

Here let me mention a few things to you, one of the things that in any statician when he doing is 2nd year or 3rds year of statistics what he would learn is what is known as fitting at distribution, given a data set, how do you fit? Let us just say bio nominal distribution, normal distribution and you have much other distribution like high square distribution, f distribution, P distribution there is the β distribution there are many number of distribution.

Now he fits and calculate some value which that is known as i^2 value if the observed ki^2 then we would say that the data is fitting the distribution, this is one thing that these statician would do that can be in 1st year of course. There the main assumption is that I said that they calculate the ki^2 value is coming under an assumption of multi nominal distribution going to some ki^2 value and the test that is developed there that is known as approximate test not an exact test.

Exact test means that distribution are exactly known approximate test means the distributions are approximate as the number of observations goes to ∞ it follows some distribution, approximate test exact test means the distribution are exactly known. So one does it that is an approximate test that the student is to do, so let us just say some distribution x is fitting one data sets, now it does not mean that some other distribution y that does not fit the data set have you understand this point?

Suppose one distribution x fits to a data asset it does not mean that some other distribution y cannot fit into that data set. So this actually says that when you finitely many observation given to you, it may be possible for you to fit more than one distribution, may be possible for you to fit more than one distribution then the distribution that you are fitting may not be unique some other person may come and then may fit may be able to fit some other distribution properly same thing to the concept so this is the first thing that people learn about finding the distribution profit given set of point.

Now there is something slightly more advanced that is there is something called estimation of nobility density functions given a data set you would like to estimate the corresponding probability density function now what is the meaning of estimating a density function a curve like this what is the meaning of estimating a curve that leads to properly defined but before going into all those definitions.

First we need to know how to estimate a point say that we assume that a distribution is normal then we have to estimate the mean of the distribution meaning just a single point and then how do you estimate that point variance covariance matrix there are these are finitely many quantities how do you estimate this first we need to note this given the distribution given the type of distribution.

They are normal how do you estimate the mean and how do you estimate a covariance matrix that is here we are estimating points then a slightly more advanced one is how do you estimate that density function directly from the given data set now about this density estimation if you look at pattern recognition test books you would find a book like you would find one chapter devoted there are one of the first papers was by passion in 1960 passerine windows way of estimating density function.

Initially person of taken the histogram as an estimate of density I hope all of you know the meaning of word histogram many of you have background in the image processing sop one of the first thing that you would have done is finding the histogram of your image the histogram is one of the first estimates that will have devotion but then a slightly more advanced one is the method given by passerine.

There he called him that nowadays it is known as passerine density estimation that for the 1960 after 1960 we are in 2010, 50 years have passed there are now even many books which are actually dealing with estimation of functions there are many text books which are dealing with estimation of functions one of the books is written by one of the ex directors of higher side the organization one BLS Prakash Rao he wrote a book on non parametric function estimation.

And find it on internet so there are even many books which are come out on estimating density function but there is a small but the but is that there are many methods of estimating density then you have the next question which method is better okay there are no good answers there are no good answers to this questions still lot of work is needed to be done one needs to do lot of work to find out which method of estimation may be better to some extend sometimes you can say.

But then ultimately one would like to ask well what is the best estimation of density function well for that at least I do not know let me just my inability I do not know the best estimator of this function so that is why my answer to the estimating principle of density function are finding density function that is only partial my answer is only partial then this comes to a very I should say grave question if we do not know.

The density function then how do we apply this rule this is the best rule in the sense it minimizes the probability of misclassification and then the next question is how do we apply this rule if we do not know the density function well the answer is plain and simple you estimate density function or you assume some function or you assume some functional form if you think that it is following normal distribution or some other distribution assume a functional form and estimate the parameter.

And do it but there if you do not know the functional form and if you do not want to do this things then the answer is no will not have any way of applying this rule to any data's if you do not know the density functions if we do not know the prior problems and you do not have any way of getting good estimates of then it is extremely different for you to apply this rule please one can make two types of mistakes one is that you can choose a slightly wrong function.

Because the data could be actually you do not know which is the best function of you are choosing one number one number two is that the estimation of parameter itself is a method it could give you an approximate solution that I will discus at least after such errors approximation one can do in this process is that one so then the very next question is if this is actually like this then why I am teaching this that forgotten the answer is very simple.

The answer is that this is the best rule if you develop a new class fire you can find its performance by comparing your base mixture how do you compare your class fire with base decision rule you generate points officially from known distributions then you can apply base decision rule generate points artificially from known distribution then you can apply base decision rule.

And you apply your class fire the class fire that you have developed and check the probability of misclassification of your class fire and see whether it is very close to the base decisions base class base error problem if it is very close then you can say that at least in these cases my class fire performances is closed to the base class this is what you can tell the reviewer of your paper right that is why this is the very first starting point of actually all pattern based decision rule.

If it is starting point of all pattern number one this is best rule and two unfortunately you cannot implied it for the most of the time so you need to develop your own classifiers that is why there are simply too many classifiers like this now you have you are asking about me SVM it is the latest slightly before that we have multi layer personal which many people use of classification there is a nearest by classifier which professor teach you in the class okay.

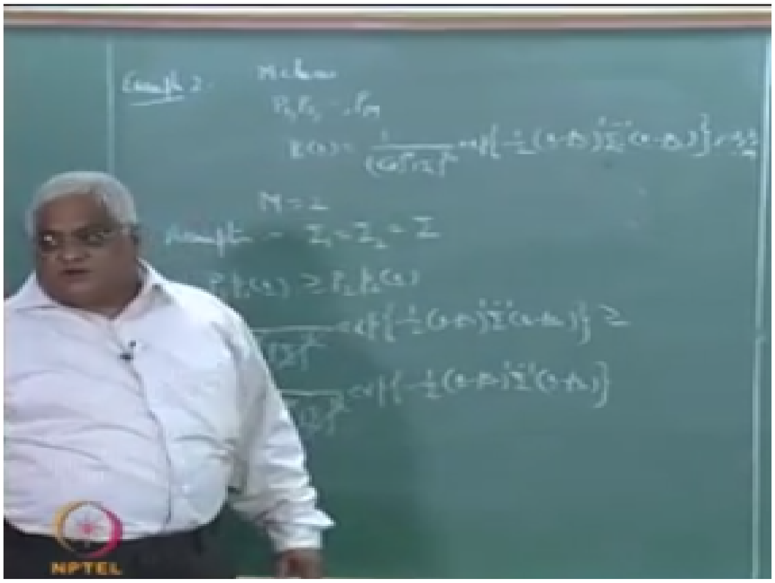
So and you have like this there are very many classifiers each one of them has some relationship with base class each of them some relationship with base class or example multi layer which is any way going to talk in the class there is a particular theorem was the paper was published around 6, 7 years ago I do not remember the exact references they he proved the authors proved that there exist an architecture of multilayer persecution for which the error probably.

He only talked about existence of an architecture he never told how to get that architecture if you apply the MLP then the error rate is very class to base decision that theorem was proved by mean one of the authors I do not remember the exact preferences that is why multilayer prospect is actually working well otherwise I mean you really do not know there are always some relationship between the standard classifiers.

And base classifiers why many people are using this classifiers because at least in some situation it works well here I am telling you the positive point earlier I mention that we had all this constraints but the people are using some situations they got good results okay at least in some situations they got good results.

As for they are applying chain in rule why people are applying it because that is the same situation they got the good results so these are basically problems which are base classifiers basically the problem is lying between density and there is also another problem that problem is

(Refer Slide Time: 59:26)



This is the error expression this is the expression for the error probability of base decision rule so write here o is for optimal i=0 and that is the complement if you apply base decision rule then the probability of misclassification is if you have number of classes Pi is the prior probability of the i^{th} class small pi is the density function of the i^{th} class conditional probability density function σi is the optimal that set for the i^{th} class.

And complement c is the complement you need to find this integral to get a value for the probability of the misclassification this is extremely classification this is an extremely complicated one even for the case of normal distribution with unequal covariance matrix it is difficult to find unequal with equal you can find it with unequal covariance matrix it is difficult so base classifier is good but the basic classifier arr number one knowing the prior probabilities and density functions and number two is one if you know is difficult to calculate.

This the shapes of these sets they might be very horrible shapes I mean you might be having I mean depending on how it may look like this may be reason for one class okay and you might be having some other thing for another class like this and let us just say you have three classes and the third class may be something like this and then how do you do the integration if you have very complicated regions then it is difficult for you to do the integration that is one of the reasons why even.

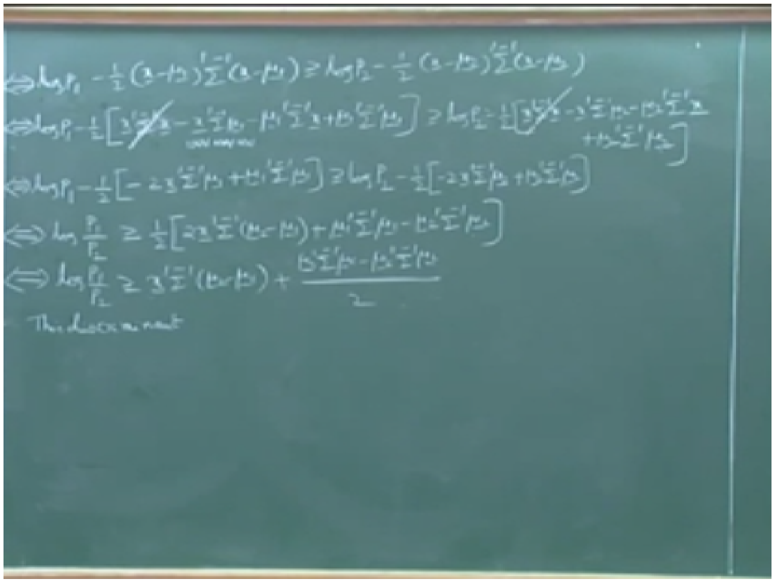
The base classifier is the best classifier because of all this base reasons people have to go in for some other classifier so that they can somehow approximately there and they can somehow approximately say that they developed classifier is performances of the base this is the reason for developing all too classifiers so we derived with case one an example two we did with case 1 now we will deal with case two her instead of m classes any general m.

And take this two classes I will just take two classes and the pix is multi variant normal here I will make an assumption is I am going to assume that σ1 and σ2=σ and I am assume that the variance the covariance matrix are same and I am going to that is the problem okay and I am going to assume that variance covariance matrices are same and I am only in the two class problem okay, and I am going to assume that variant covariant matrices are same.

Now let us see what is going to happen to the base decision row again so this is p1 p1x we only have two classes is greater than or equal to p2 p2x this is define only a there is p1 this is 1 / √ 2 π n determinant of Σ1/2!-1/2 x - μ1 time Σ inverse x - μ1 greater than or equal to p2 this is x - μ2' σ inverse x - μ 2, you will understand why this assumption is made you see you can cancel out this and this and you can also cancel out determinant of $\sigma^{1/2}$.

Now let us apply logarithms on both side, so what is going to happen? This is log p1 log means to the base e natural logarithm log p1 and if you apply logarithm here exponential goes out this will be minus.

(Refer Slide Time: 01:05:46)



$$\Leftrightarrow \log p_1 - \frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) \geq \log p_2 - \frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)$$

$$\Leftrightarrow \log p_1 - \frac{1}{2}\left[\cancel{x'\Sigma^{-1}x} - \underline{x'\Sigma^{-1}\mu_1} - \mu_1'\Sigma^{-1}x + \mu_1'\Sigma^{-1}\mu_1\right] \geq \log p_2 - \frac{1}{2}\left[\cancel{x'\Sigma^{-1}x} - x'\Sigma^{-1}\mu_2 - \mu_2'\Sigma^{-1}x + \mu_2'\Sigma^{-1}\mu_2\right]$$

$$\Leftrightarrow \log p_1 - \frac{1}{2}\left[-2x'\Sigma^{-1}\mu_1 + \mu_1'\Sigma^{-1}\mu_1\right] \geq \log p_2 - \frac{1}{2}\left[-2x'\Sigma^{-1}\mu_2 + \mu_2'\Sigma^{-1}\mu_2\right]$$

$$\Leftrightarrow \log \frac{p_1}{p_2} \geq \frac{1}{2}\left[2x'\Sigma^{-1}(\mu_2-\mu_1) + \mu_1'\Sigma^{-1}\mu_1 - \mu_2'\Sigma^{-1}\mu_2\right]$$

$$\Leftrightarrow \log \frac{p_1}{p_2} \geq x'\Sigma^{-1}(\mu_2-\mu_1) + \frac{\mu_1'\Sigma^{-1}\mu_1 - \mu_2'\Sigma^{-1}\mu_2}{2}$$

The discriminant

-1/2 x - μ1' ≥ log p2 – ½ x - μ2' Σ inverse x - μ2, now I will actually do this multiplication log p1 – ½ this will be x' x x' Σ inverse x – x' Σ inverse μ1 - μ1' Σ inverse x + μ1' Σ inverse μ1 ≥ log p2 – ½ x' Σ inverse x – x' Σ inverse μ2 - μ2' Σ inverse x + μ2' Σ inverse μ2. This is looking slightly complicated but you will see that this will going to be reduce to something slightly simpler, look at this x' Σ inverse x this is -1/2 x' Σ inverse x you will find here and the same quantity will find here, so I can cancel it out okay.

Now let us see log p1 – ½ x' Σ inverse μ1 x is n / n rows and one column x' is 1/n vector Σ inverse is m/n μ1 is n/1, so this whole thing is 1/1 so it is the scalar. What is the μ1' Σ inverse x? That is also a scalar I claim that these two quantities are same, because what is the transpose of this? The transpose of this is this and this is a scalar this is the scalar the transposes are same, so this quantity is same as this quantity.

Similarly here this quantity is same as this quantity, is it clear to you? So then what is going to happen here this is – 2 x' Σ inverse μ1 + μ1' Σ inverse μ1 ≥ log p2 here there is as fn only if I need to write fn only here I need to write fn only here, so log p2 – ½ this is -2 x' Σ inverse μ2 + μ2' Σ inverse μ2 okay. now what I will do is that I will take all this terms on this side and the log p2 on this side so if I bring log p2 on this side this is just going to be log p1 - log p2, can I just write log of p1 / p2 this log p2 when I am bringing it on this side this is log p1 – log p2 I am just writing log of p1 / p2 this is ≥ this -1/2 it is going that side it will become + ½.

So this will be 2x' Σ inverse μ2 - μ1this -1/2 is going inside that will become + ½ okay but this negative sign going to remain, so here you have x` σ inverse μ2 here you have x` σ inverse μ1 it is a negative sign, so x` σ inverse μ2 – μ1 now let me this term also it should be written on this side that will be okay and in many books what you are going to find is people take this two also inside then what we are going to get is x` σ inverse μ2 – μ1 + this is μ1' σ inverse μ1- μ2` σ inverse μ2 is whole thing.

This is what you will find in many books now why did I do all this calculations I will tell you the reason for doing it, if you look at this expressions you will think that x` σ inverse x it is quadratic in x you would assume that the decision boundary would be quadratic in x I hope you understanding the meaning, the decision boundary would be quadratic in x if you have something like x` σ inverse x is that time is there then there is a square time is at some point x` is x1 of 2 N there is a σ inverse.

That is again x1 to x2 N so you are going to get at some point of 10 $x1^{2}$ at some other point of 10 $x2^{2}$ and at some other point of time probably x1 into x2, so these are quadratic terms but because of this assumption the quadratic term is gone this is a constant not that μ1 σ μ2 there are learn to us so this is a constant P1 P2 they are known to us this is a constant μ2 – μ1 this is a constant σ inverse is a constant what you have a something linear in x what you have is something linear in x basically what you are going to get here is.

What is known as linear discriminate function this is a linear discriminant function this is a linear discriminant function this if your distributions are normal and you have two classes and of the correlates matrices are same then the decision boundary between the classes will necessarily be linear if the correlates matrices are not same then you would get non linear decision boundaries because you are going to get x` σ1 inverse x here you are going to get x` σ2 inverse x which is going to be non linear in x.

Here what you have got is something linear in x you basically how got a linear discriminant function in x, so this discriminant function is known as discriminant the over discriminant it came from discrimination you are going to discriminate between two classes from the word discrimination discriminant it has come, so sine the function is linear then this is a linear discriminant function this discriminant function between classes is linear this is linear, so if you have two normal distribution correlates matrices are same you will get linear if you have three normal distribution.

Let us say correlates matrices are same then what are you going to get between 1 and 2 there is a linear function between 2 and 3 there is again linear function between 1 and 3 again there is a linear function right we are going to get piece wise linear decision boundaries 1 and 2 is linear 2 and 3 is linear 1 and 3 is also linear so you are going to get piece wise linear decision boundaries between classes okay so similarly and if you have n number of classes we are going to get those many such piece size linear functions let us stop.

End of Module 02-Lecture 02

Online Video Editing /Post Production
M.Karthikeyan
M.V.Ramachandran
P.Baskar

Camera

G.Ramesh
K.Athaullah
K.R.Mahendrababu
K.Vidhya
S.Pradeepa
D.Sabapathi
Soju Francis
Selvam
Sridharan

Studio Assistants

Linuselvan
Krishankumar
A.Saravanan

Additional Post –Production

Kannan Krishnamurthy& Team

Animations

Dvijavanthi

NPTEL Web &Faculty Assistance Team

Allen Jacob Dinesh
Ashok Kumar
Banu.p
Deepa Venkatraman
Dinesh Babu.K.M
Karthick.B
Karthikeyan.A
Lavanya.K
Manikandan.A
Manikandasivam.G
Nandakumar.L
Prasanna Kumar.G
Pradeep Valan.G
Rekha.C
Salomi.J
Santhosh Kumar Singh.P
Saravanakumar.P
Saravanakumar.R
Satishkumar.G
Senthilmurugan.K
Shobana.S
Sivakumar.S
Soundhar Raja Pandian.R
Suman Dominic.J

Udayakumar.C
Vijaya.K.R
Vijayalakshmi
Vinolin Antony Joans

Administrative Assistant

K.S. Janakiraman

Principal Project Officer
Usha Nagarajan
Video Producers
K.R. Ravindranath
Kannan Krishnamurty

IIT Madras Production

Funded By
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in