Indian Institute of Technology Madras

NPTEL

National Programme on Technology Enhanced Learning

Pattern Recognition

Module 01

Lecture 03

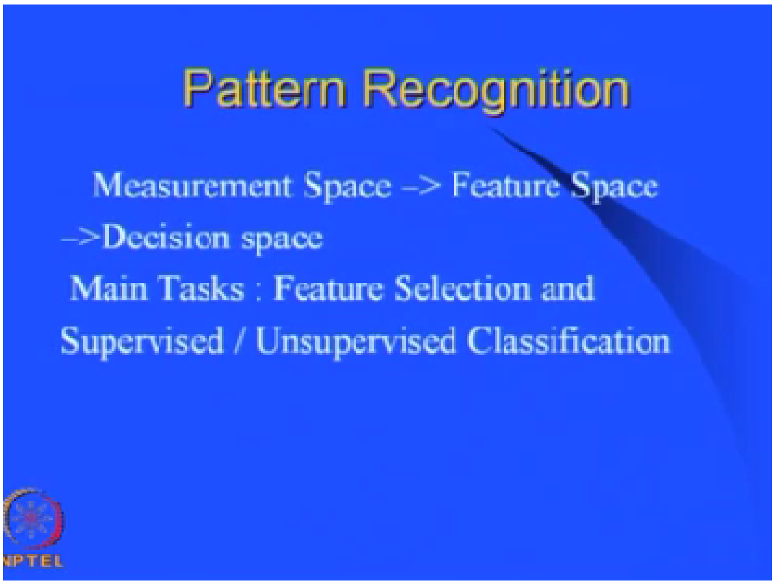
Principle of Pattern recognition III
[Classification and Bayes Decision Rule]

Prof. C. A. Murthy

Machine Intelligence Unit,
Indian Statistical Institute, Kolkata

So I have discussed little bit of mathematical preliminaries for pattern recognition, where I talked about some amount of statistics and some of the connections between the statistics literature and matrix algebra. Now let us actually start going to the subject.

(Refer Slide Time: 00:37)



Pattern Recognition

Measurement Space –> Feature Space –>Decision space

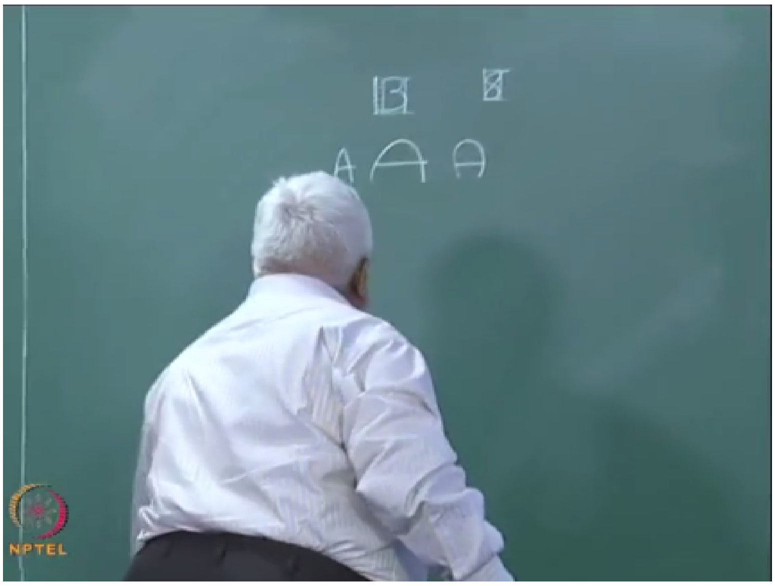
Main Tasks : Feature Selection and Supervised / Unsupervised Classification

Basically, in pattern recognition we have some phenomenon on which we have made some measurements. From those measurements, basically we have to look at what all features that we can think of from the measurement space where we have to go to what is known at that feature space. From the feature space what is known as decision space, we need to make decisions about the phenomenon under consideration.

And the main tasks are one task is feature selection, which features are actually important and the problem under consideration. And the next task we are supposed to do either classification, which in some of the books it is written as supervised classification. Nowadays, it is basically known as classification.

And unsupervised classification, this is slightly old terminology and nowadays it is known as clustering. We need to do either classification or clustering and the other task is feature selection. So basically from the measurement space we are suppose to go to feature space and from there we are suppose to going to the decisions. Now again let me just give you an example. Suppose the problem that you are dealing with is a character recognition problem. What is a character recognition problem?

(Refer Slide Time: 02:20)



See, this is a character B, this is the number eight. This is character B and this is the number eight. We know this is B and this is eight. You want the computer to say that this is B and this is eight. So this is the recognition we need to recognize this. So, how does one do it? There are several ways of doing it. I am just mentioning a procedure, which is easily understandable to you.

What you need to do is that, you enclose the character into a rectangle where the sides of the rectangle are parallel to X and Y axis. So, it should be probably like this and here it is like this. Similarly, here also parallel to, so it must be, should be like this. I wrote it slightly slant way. So here this line is actually touching this that is why it is like this. So, now what one is going to is what is the main difference between B and eight?

This is sort of perpendicular to this line. Whereas if you look at this character, this one and this portion is same, but basically it is this portion where there is a difference. So, if you want to differentiate between B and eight, what one needs to do is that, we get some points on this line and you measure the distance of this point to the one that is very close to this and you draw a line parallel to this and see where it is actually touching.

This is touching at this place, this is touching at this place, this is touching at this place, just measure the distances. If these distances are all more or less the same, then you are going to get a B okay. If they are not same, you are going to get eight. Here this will be very, very deep and here may be small and here this is actually zero. So, these are the features that we are measuring.

Earlier we were given these two characters, which we have digitized, we made a binary image, and this is a binary image. Then what we have done is, we have enclosed this one in this and then from here we have found this features. So, note that from the measurement space, we went to the figure space. We calculated these values. Now, after we have calculated those values, we made a decision.

If these values are all very close, then we call the character as B. If they are not very close, then we call the character as eight. Are you understanding? What I am trying to say? If they are not very close, so this is just for you people to understand the thing. There are much better methods; in fact those methods not only dealing with B and Eight, they deal with all the characters in the English language, okay.

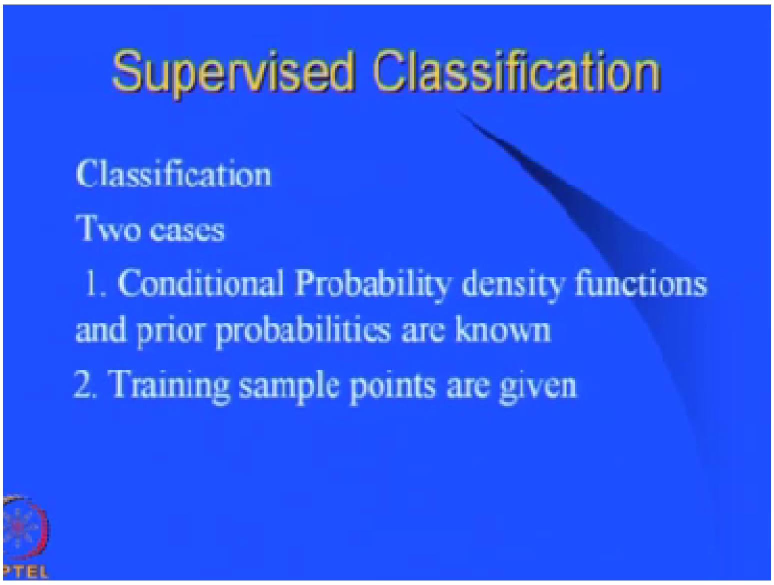
Not only capital letters. There are lower case letters and upper case letters. You have all the numbers zero to nine all the digits, and we have all possibility and they take care of all the things. For English language, the character recognition software is you have really developed software's available in the world now, character recognition software for English language.

So, whatever I am telling is basically a rudimentary thing just to make you understand the mean of measurement space, feature space, and decision space. Measurement space is the given two characters, which we have digitized them and we made a binary image. Then feature space, that we measured this features and decision of those measurements we have a made a decision on the basis of those distance value.

So, this is just an example of what measurements space is, and what feature space is, and what decision space is. It is not necessarily true that in every example, measurement space and feature spaces they are different. Sometimes, they may be same, sometimes a portion of measurement space you will put into features. So, it depends on how you look at it, and then we are going in for decision and the decision can be of various types.

If we are looking at some sought of dissimilarity and similarity and within a cluster, the similarity should be more between two clusters that they are, the dissimilarity between one cluster or other cluster should be more than the similarity, between two cluster the dissimilarity should be more and in a cluster the similarity should be more. That is for clustering and while interested in doing classification. So there are main problems that we are going to encounter here are mostly their feature selection, clustering, and classification.

(Refer Slide Time: 09:37)



Supervised Classification

Classification

Two cases

1. Conditional Probability density functions and prior probabilities are known

2. Training sample points are given

Now, let us just see in classification there are basically two cases that we are going to consider in these classes. But in one case, we are going to assume that conditional probability density functions and prior probabilities are known for the classes. In another case, we are going to assume that 10 sample points are given. Now, what is the meaning of this two cases.

The meaning is the following, in the case one here; we will be knowing the number of classes. What is meaning of we will be knowing the number of classes? Note that in my previous lecture, I was giving you several applications of pattern recognition, where I was talking about land cover type the satellite images. The number of different land cover types are classification of pixels that is the number of classes.

The number of different land cover types, I think these kinds of number of classes in that problem. Now in the problem of digits, you have zero to 9, ten digits and you are given a new digit. This new digit should be put into one of these 10 digits. If that is problem, then the number of classes is 10, and those classes are that correspond to the digit zero, digit 1, digit 2, digit 3, up to digit 9.

If there a classification of characters, the upper case letters of English language, then there are 26 classes, because there are 26 letters A, B, C, D, up to Z. When someone writes A, or someone writes one of the character, you should put the characters into one of those classes. Now one may write the letter A, it may be like this, it may be like this, it may be like this, so you have several such things.

Each one of them, you should into the class of letter A. Each one of this thing, you should put into the class of letter A. So, we are going to assume that the number of classes is known. Now the second one is that you see conditional probability density function. At the present moment, let me forget about the meaning of the word conditional. Probability density function.

You see, we have to talk about the probability density function, when we know the features and their consideration. If the number of features is small and numbers are small, then the probability density function is defined over the space Rn. So, when we are going to talk about the probability density function, it means the first question is how many variables do you have? Are the same as how many features do you have?

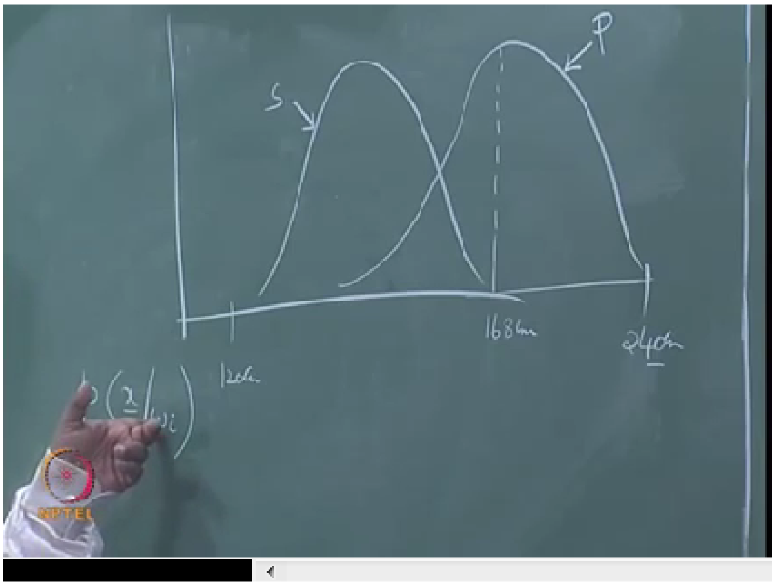
Let me at the sake of convenience, again mention an example. The example that I am going to mention is classification of say, persons belonging to two different communities. One community is south Indians, another community is Punjabis. So basically you have two classes. One class is Punjabis, another class is south Indians. Now, we are going to measure features.

What features are you going to measure? Let us just say one feature is height, another feature is weight, and you might measure some other features also, and I have just given two features. You might measure some other features also. Then the probability density function, it comes in with, how you take see, let just say we are going to be interested in classification of males.

Let us for the present moment forget about females, and let us also for the present moment assume that the males under consideration we have at least the age as 20 years, so that let us not think of small children. Age, months, and year, let us not think of them okay. So, then our main problem is we have some males, each one of them, the age is greater than or equal to 20.

Each male is either a Punjabi or a south Indian, and what you need to do is that, you should put this male into either the Punjabi class or the south Indian class. That is the problem under consideration. Now what are the input that is given to you, the input is for the Punjabi class we know the density function. For the South Indian class we know the density function. What is the meaning of that? The meaning is following.

(Refer Slide Time: 15:40)



Let us say you only have one feature height we only have one feature height. Let us say height we are measuring it is in cm, let us say the maximal height that we have is 240cm. 240 cm means how many feet it is 8 feet. And let us say the minimum height is 120cm. And the maximum height is 240 cm.

And now let us look at the class of Punjabis, mostly and let us assume that we only have one feature. So that I can draw the density functions if I have only one feature. If I have two features then I need to draw to three dimensional spaces which I cannot do okay. So I am going to assume that for the present moment in order to make you understand I am going to assume that I only have one feature that is height.

So basically what may happen is with respect the height may this is the density function of the Punjabis say this thing it may correspond to something like may be around 5 feet, 8 or 5 feet 9, so how much it will be, so it is around 167, 168cm or so something like this. Whereas South Indians, because South Indians are generally shorter than Punjabis that is one of the reasons why I took this two communities. Am I right?

South Indians are generally shorter than Punjabis. Am I right? Yes, so I think for South Indians it may look like this. It is this side of the curve how much this side that I do not know but it is on this side of the curve. So this is for say Punjabis, say this for South Indians the density function. That means area under this curve is one, area under this curve is also one.

So you can have some density functions when you have the numbers of features to be two, three, four like that you are given the density functions of the classes. You have two classes the example that which you have taken, you have two classes and for each class the density functions are given. Since the density function is given for a class that is known as class conditional density function.

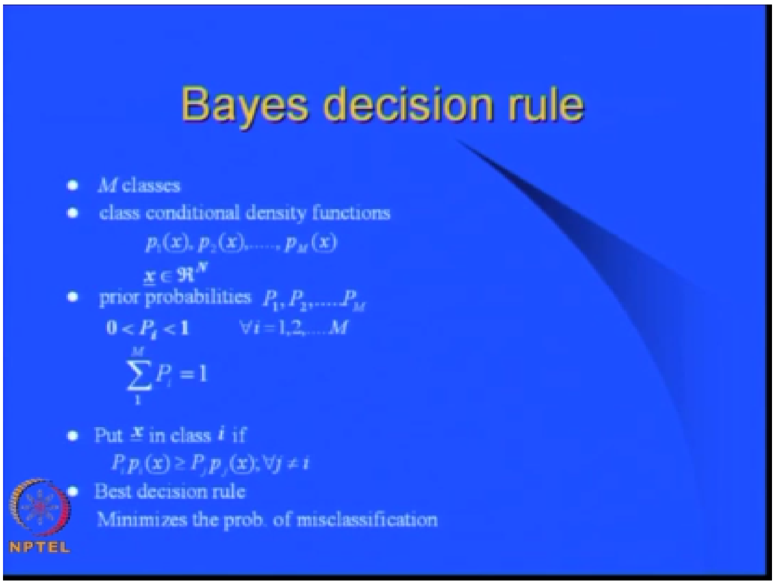
And it is represented by, if you look at the books it is represented by this is the general representation that you find, P represents probability density function, this is for the ith class and for the point x the density value is PX/ωi, this is the notation that you will find in books. I have followed; I did not want to write this wi always, so I have just followed this notation.

You will see in my slides and I have followed this notation, but there is one another quantity is there. Prior probability, what is a prior probability? You put Punjabis and South Indians together all Punjabi's and South Indians together. Suppose in the whole mixture, let us say 40% of Punjabis, 60% are South Indians then the prior probability of the class Punjabi is 0.4 and the prior probability of the class South Indian is 0.6. Are you understanding?

Let me repeat, you put Punjabi and South Indians together in the whole mixture if you get 40% as Punjabis, 60% as South Indians, then the prior probability for the class Punjabi is 0.4 and the prior probability of the class South Indian is 0.6. So these are prior probabilities.

Now what we know is the conditional probability density functions are known and the prior probabilities are also known.

If you know these two, then how does one get the rule of classification? There is a second case where training sample points are given. I will explain to you slightly later. I will explain this case to you slightly later.
(Refer Slide Time: 21:57)



Bayes decision rule

- M classes
- class conditional density functions
  $p_1(\underline{x}), p_2(\underline{x}), \ldots\ldots, p_M(\underline{x})$
  $\underline{x} \in \Re^N$
- prior probabilities $P_1, P_2, \ldots\ldots P_M$
  $0 < P_i < 1 \qquad \forall i = 1,2,\ldots M$
  $$\sum_{1}^{M} P_i = 1$$
- Put $\underline{x}$ in class i if
  $P_i p_i(\underline{x}) \geq P_j p_j(\underline{x}); \forall j \neq i$
- Best decision rule
  Minimizes the prob. of misclassification

Let us assume that we are given the conditional probability density function and prior probabilities. Now this is what is known as Bayes decision rule, well what is this rule? I have assumed here that you have N number of classes, and the density functions are naturally have many classes are there those many density functions you would have. So the density functions are p1, p2 up to pm.

Now the number of features are N, since I have assumed that all the points that belong to N dimensional Euclidean space in the number of features are N, and the prior probabilities has many classes are there those many prior probability you need to have. So the prior probability are p1, p2 up to pM, these are all known. And naturally the summation of this probability is one, and all the probability is must slightly between zero and one their probabilities.

Now what is the Bayes decision rule says is put x in class I if pi, pix is greater than or equal to pj, pjx for all j not is equal to i. This is what is known as Bayes decision rule. Now I think the remaining part of this lecture and the next lecture I am going to surely concentrate upon this rule which is Bayes decision rule. You see whenever you are doing any classification that is some misclassification.

Whenever we are doing any classification there is generally some misclassification. Let me tell you this thing by using a few examples. The examples that have we have considered Punjabi and South Indians. We have taken the features as let us just say only two features height and weight. Let us assume that we have only two features height and weight.

Now if you put a threshold that if the height is greater than this and the weight is something, then you take this person as Punjabi, otherwise it takes South Indian okay. Now then you can ask me a question on behalf South Indians whose heights are greater than that. How do we say that we can never have one, that I cannot say that. Are you understanding?

Whenever there is a decision rule it is not necessarily true that every point in one class it satisfy that rule and other points and other class it satisfy the complimentary rules that is not necessarily true. In general you have some amount of misclassification. Let me tell another example, this example concerns images, suppose you are working in a bank, you would like to see to it that people do not steal money from the bank that is what is your aim.

Now what you would like to do is that you would like to prevent people carrying rifles, pistols, and guns etc.., like that says 10 to 15 such objects, now what you have done is you have put a camera at a specific location. The camera is a video camera. Whenever someone is entering that particular portion of the bank the camera is looking at the person and it is a recording. Actually it is taking photography.

And the recording is going to the one who analyzes the video naturally earlier there was a human being who is sitting next to the monitor to see whether anyone is carrying this thing and he would make his own judgement, but it is difficulty expecting human being to be really nice about judging this thing in each and every minute of the day, probably the first few minutes or the first 20, 30 minutes or the first one or two hours he would be really alert.

Afterwards you would start seeing the same thing he would not be alert. Why we are blaming other, even if you sit or I sit in next to the monitor start observing the same thing we would not be alert. This is the property of all of us human beings. So it is better that a machine does this analysis and comes out with someone is comes out with judgement of miscarrying one of these objects are not.

Now in order to make this judgment there are again two problems involved. One is someone is actually carrying and the machine says that he is not carrying, this is one error. You have another error, someone is not carrying it, the machine is saying that he is carrying it. There are two errors, let me repeat the machine says that this man is carrying it, but actually he is not carrying, that is one error. The machine says that he is not carrying, but actually the man is carrying, there are two errors.

Now which error you would not like to have. And which error you do not mind. The answer is if you are the manager of the bank you would say that well I do not mind if you falls alarm, if someone is not carrying any such explosive device, but if the machine says he is carrying it, I do not mind it, but on the other hand if someone is actually carrying and if the machine says he is not carrying then I do mind.

So I do not want this sort of error to happen, are you understanding? I do not want this sort of error to happen. So it is not necessarily true that the errors would have the same weight depending on the situation the errors would have different weights. Now in this slide you please note the last line it is best decision rule minimizes this is the probability of misclassification.

Yes, it minimizes the probability of misclassification by assuming that the errors would have the same weight. I am repeating, it minimizes the probability of misclassification by assuming that the errors would have the same weight. If the errors do not have the same weight then this rule is not applicable. If the errors do not have the same weight then this rule is not applicable.

One can prove that if the errors have the same weight then this rule minimizes the probability of misclassification. That is why this rule is known as the best rule for classification. This rule is known as the best rule for classification, because it minimizes the probability of misclassification. I will give you another example where the errors would not have the same weight.

The example is the following there are two statements, so there is actually one statement. The statement is India is going to attack Pakistan tomorrow, and you have to attribute this statement to one of the two persons. One person is CA. Murthy, another person is Manmohan Singh. Now if the statement had actually been made by CA. Murthy and if you attributed to Manmohan Singh, then you can see the consequence are enormous.

On the other hand it was really made by Manmohan Singh and you have attributed to CA. Murthy, well may be your Boss, you are journalist your boss would say fine, you have made a mistake, does not matter that much. Nothing would happen to you. It is not really that serious an offence than the first one. So here again the weights are not same. This example is from speech.

The earlier example was from images and in any classification you are going to have these sort of thing and I observation belonging to one class if you put it into another class then what are the consequences. The consequences can be first you are looking at the consequences only from the point of you all I mean how much misclassification it is going to occur without really looking at the weight of misclassification.

Then the next one is that if the misclassification actually happens how much cost you are going to get, if the costs are same then the weights are same, if the costs are different, then the weights are different. Then the second one is cost of misclassification. The cost of misclassification is the thing that actually decides that actually makes you many situations to have your own decision rule of the problem.

Because cost of misclassification may vary, you would like to have some cost of misclassification some set of weight for the misclassification may be some other person for the same problem may have different set of weights. It depends; you might be knowing that cancer is a disease which is not actually easily detectable in the early stages even now.

Now say you are a doctor and a patient comes to you, now you have to say let this person is having cancer or not. You have a problem, your problem is if the patient is actually having cancer and you say that patient is not having cancer, because the knowledge is not completely known to you, then that creates a problem. Not having cancer you would say that the patient is having cancer you are going to give medicines according to that and that will creates several complications on the patient.

On the other hand if the patient is having cancer and you say that he does not have. That will create another difficulty. Then what is that you are going to do as a doctor, what do you say that the patient because its not quite clear to you, then probably the best option is you directly admit that you are in the position to judge whether the patient is having cancer or not. Otherwise, if you make a wrong judgement nowadays people can file cases in courts, probably you are aware of these things.

This is again, so the doctor is facing the classification problem, whether the patient is having cancer or not. You will see very many classification problems in actually every minute and every second of our life when you walk from here to downstairs how do you talk the step at this place or that place. It starts from there every minute we are making decisions.

Every minute and every second we are making decision by going out, what path are you going to chose. What path there is lot of space why do you take one such path, why do not you take another path? If I ask you do you have an answer look at that place there? The door and the place there, we have an answer to that note that you really do not have an answer, because you are not actually optimizing anything.

Whichever is reasonable you are just following so for the same problem you have different solutions at different points of time or the same problem you have different solutions at different points of time. And every problem there is pattern recognition problem though you may not have unique solution, there are several solutions you can take anyone of them. You can take anyone of them, I suppose I need to stop here.

End of Module 01 – Lecture 03

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
S. Subash
Selvam
Sridharan

Studio Assistants
Linuselvan
Krishnakumar
A. Saravanan

Additional Post – Production
Kannan Krishnamurty & 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
Santosh 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