Pattern Recognition and applications

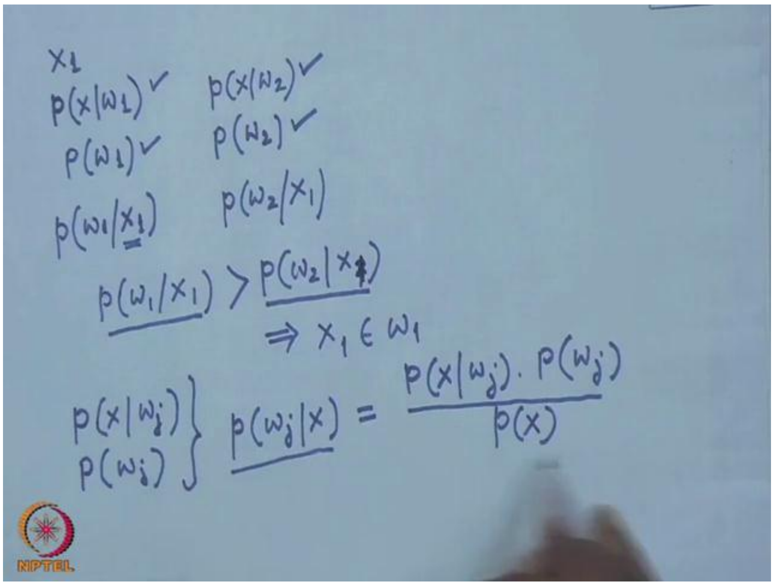
Prof. P.K. Biswas
Department Electronics and Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Lecture - 11
Probability Density Estimation

Hello, so welcome back to this video lecture series on pattern recognitions and applications. So, today what we are going to discuss is given a set of examples, how we can estimate the probability density function. The problem is very important, because you might have noticed that the last few classes, we have discussed about the Baye's decision theory, we have also discussed about the maximum likelihood estimate of the parameters.

So, when we talked about the Baye's decision theory we have said that if you are given say two class, class omega 1 and class omega 2, and given an unknown sample say x of feature vector x, we have to classify this feature vector x into one of these two class. That is I have to decide feature vector x which has to be classified to class omega 1 or it has to be classified to class omega 2. And for that what Baye's decision theory says that I have to estimate what is the probability, that this unknown feature vector x belongs to class omega 1 and what is the probability that this unknown x belongs to class omega 2. So, for each of our class the probability is more, I have to classify this unknown feature vector x to a corresponding class.

(Refer Slide Time: 02:02)



In other words what we have done is, so we have this unknown feature vector x and I know what is the probability p of x given class omega 1 or let me put it as unknown feature vector x 1. And I have this conditional class density probability function which is given by p x given by 1, that is the probability density function of samples which come to the omega 1. And I also know what is the class condition ling probability, that is p of x given omega 2 and it is also assumed that I know what is the apiary probability of class omega 1 and class omega 2. That is I know what is the apparel probability of class omega 1?

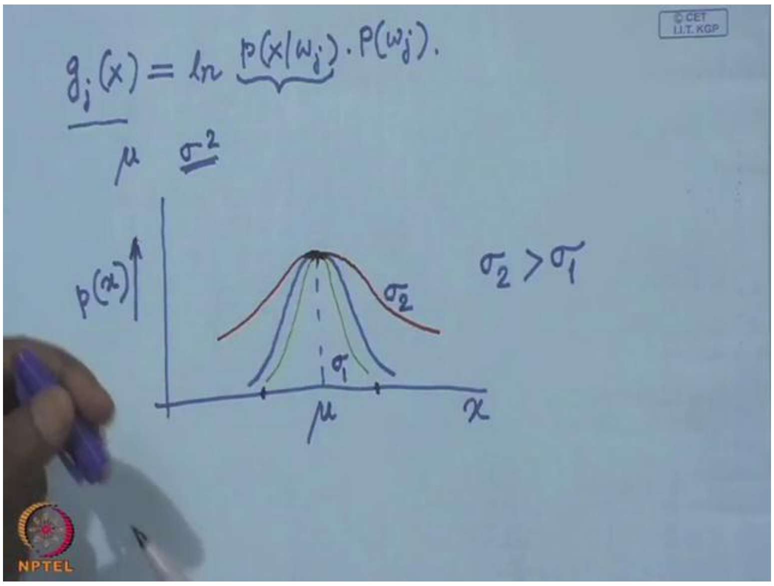
I know what is p of omega 2 which is nothing but the apparel probability of class omega 2 and even this I have to find out the probability of what is p of omega 1 given x 1, x 1 is the unknown feature vector. So, once I have this observation, this unknown feature vector x 1 I have to compute what is the probability that belongs to class omega 1, which is nothing but p of omega 1 given x 1.

I also have to compute what is p of omega 2 given x 1, that is given this unknown feature vector x 1 what is the probability that this x 1 belongs to class omega 2. And then the Baye’s decision rule says that if p of omega 1 given x is greater than p of omega given x will be, sorry p of omega given x. Then my decision will be that x belongs to class omega. Otherwise, if p of omega given x is greater than p of given x, when decision classification decision will be on x belongs to class omega.

So, that is how the Baye's one works and when I compute this p of given omega x or of p given x, that is basic rule. That is as I know that what is p of x given omega and what is the apparel probability, what is p of omega. Similarly, I know what is the class conditional probability density function, that is p of x given omega. I also know what is p of omega, what is which apparel probability or in general I know what is the class conditional disable function, that is p of x given say omega j, which is the class probability density function. Class conditional function of the samples belonging to the class omega j, I also know that what is the a parable probability a class omega j. That is p of omega j and from these two I can compute the apery probability, what is p of omega j given x which is nothing but p of x given omega j into a parable probability p of omega j upon p of x or probability of x.

So, this is the one which has to be computed or apery probability which has to be computed for all the classes. And the class for which this apery probability that p of omega j given x will be maximum, this x will be classified to the corresponding class. So, this was the basic Baye's classification rule and following this in last two classes we have also derived a number of functions called decision functions or decimating functions. So, the decimating functions for a class omega j.

(Refer Slide Time: 06:47)



We have defined that as g j given of x, which is the decimating function of class omega j and this we have defined as log of the of x given omega j into apery probability p of omega j. And while defining this you see that we have ignored one term which was in the denominator, that is p of x. This we have ignored simply, because this denominator term is equal in all the apostasy probabilities expression. So, whatever the omega j this p of omega j will appear in all these apostasy prosperity probable expressions. As a result this p x does not influence the decisions that to which class the x will belong.

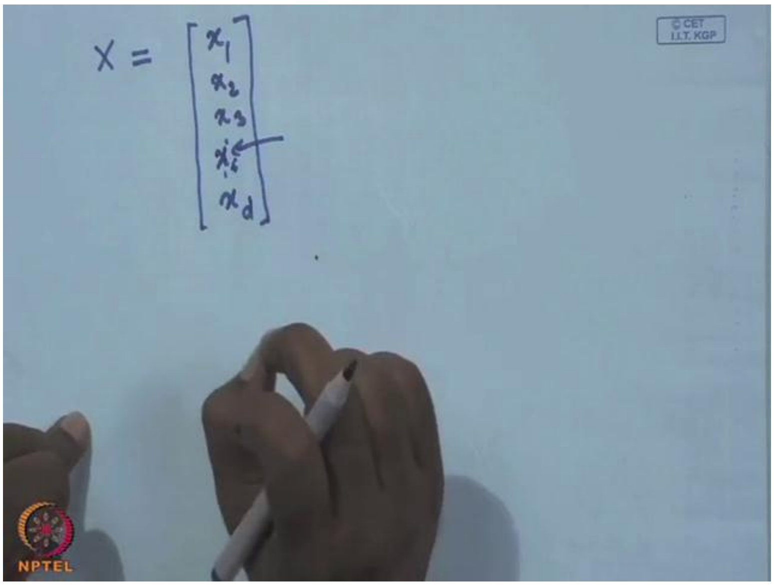
So, when we have found out the expressions for g x, that is the discriminating function, we have assumed that p of x given omega j. That is the class conditional probability density function that follows a predefined form or the expressions that we have derived for this g x. For that we have assumed that the form is a Gaussian distribution or normal distribution. So, when we have a Gaussian distribution, you know that the probability density class that is internally defined by two parameters. One is the main mu and the other one is the variance, which is sigma square.

So, if I go for if I look at the probability density function which is defined over a single variables say x or a scalar variable say x, then if I know mu that is the mu of the samples. And if I know the variants which I know that is sigma square of the samples, I have defined what is the class, which represents this probability density function. So, if I plot, it will be something like this. This is my variable x and along the vertical direction I plot p of x and the curve. If the mu is the mean of sample x then the carve will be something like this.

So, you all know that the position of the pic is defined by is decided by what is the mean of the samples, which is mean mu and the shape of this probity density called that depends upon the variants which is sigma square. So, if sigma square of the variants is very small in that case the curve will be very sharp. It will be something like this when the various is small and if the variants is large then the curve will be a flat one, which is like this. So, if this curve is for sigma is the standard deviation and the sigma is the variants and this is the carve, which is for sigma mean value remaining. The same this indicates that sigma is greater than sigma and if the mean is changed then the position of the peak will be prior city function that will shift along its access.

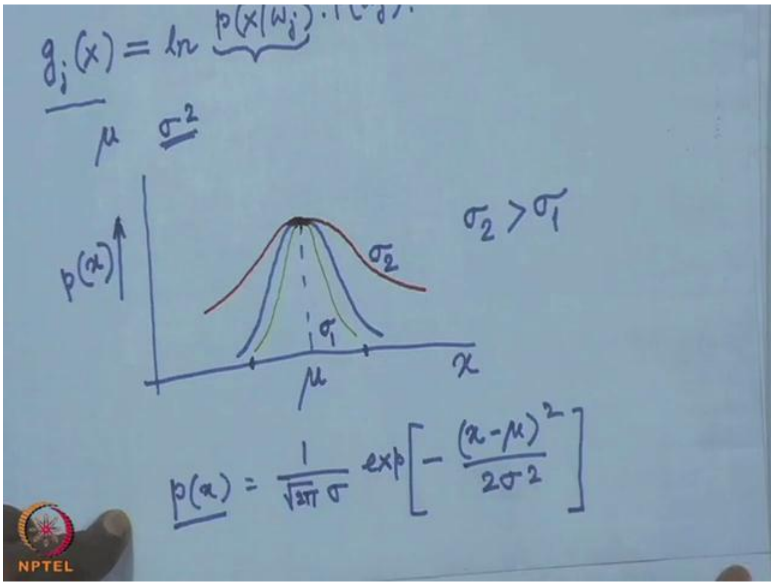
So, for me one if mu is greater than mean mu, then this will be shifted to this location for another mu value. If that is less than mu, this carve will be shifted to the left. So, this is what we get for a symbol variable. If I have multi variable in which gives the feature is represented in the form of a feature vector.

(Refer Slide Time: 11:41)



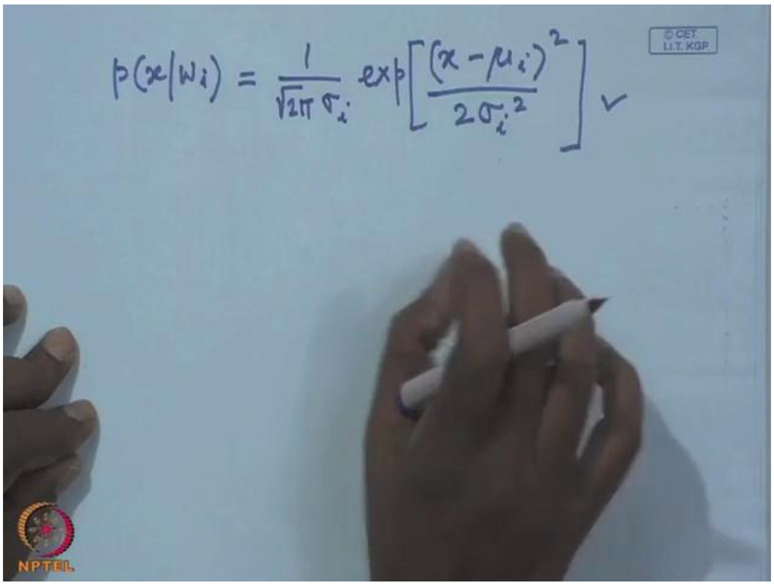
So, such a feature vector is represented by capital X. If this feature vector has say d number of components, then feature of vector capital X will be given by x 1, x 2, x 3 up to x d which becomes a vector having the number of components. And each of them whether its x 1 or x 2 or x 3 or x d in general say ninth element, say x i. This represents a particular element of the feature vector element x or it gives a particular characteristic or a particular feature of the pattern, which you want to classify and then generalizing this Gaussian expression say in this case.

(Refer Slide Time: 12:35)



You all know that the Gaussian expression, in this case you all know that p of x will be simply given by 1 over square root of 2 pi into sigma exponential of minus x minus mu square upon 2 sigma square. So, this is well known form of an Gaussian expression and over here this represents the probability density functions of x. And if I want to make it a class conditional probability density function, that is instead of taking all the x, if I take the samples which belongs to class say omega i.

(Refer Slide Time: 13:30)

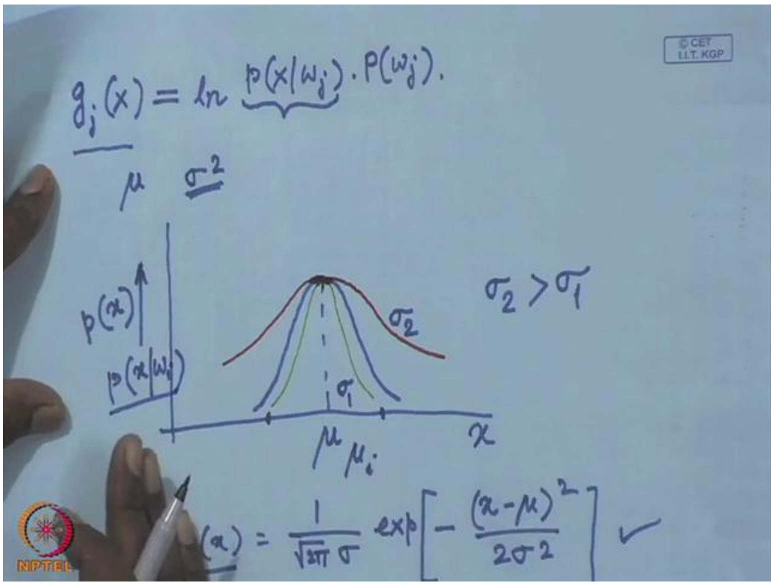


So, I want to make it p of x given omega i. So, the expression of given omega i will be similar. It will be 1 upon square root of 2 pie into sigma i exponential x minus mu i square upon 2 sigma i square. So, the only difference between this expression and this expression is that here, well we have taken different sample values of x. We have now considered whether these samples are taken from class omega or the samples are taken from class omega or the samples are taken from the class omega 3 and so on. We put all the samples together irrespective of what is the class belongings of the sample, but when I look at this expression, that is p of x given omega i, I take only the samples from class omega i. I do not take samples from class omega j, but j is not required.

So, naturally if I want to find out what is p of x given omega, I will take samples only from class omega naught. Take any sample from class omega, I will not take any sample from class omega 3 and so on. Similarly, if I want to compute what is given p of omega 3, I will take only samples from class omega 3. I will not take any samples from omega or omega or omega 4 and so on.

So, once I take only samples from class omega, this mu I simply becomes mu and the radios that I compute that is sigma i square that becomes sigma square. So, it is computed only from the samples which are from the corresponding class, that is what is class conditional probability and in such case this density function probability function.

(Refer Slide Time: 15:56)



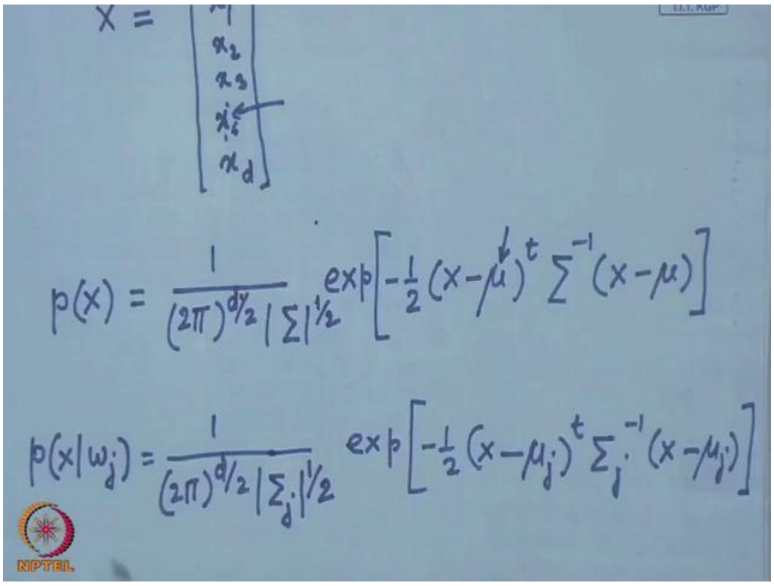
That I thought it is simply become p of x given omega i and this mu will replace by mu i, that is mu of the samples taken from class omega i and this the priority density form function of the samples taken from class omega i. This is what we have in case of a single variable. Now, how this expression changes if I go for multiple variable, that is instead of x being a scalar. Now, x is a vector, so were the vector are replaced in this form it is represented by capital X. Suppose, this vector has d number of components from x to x d.

So, this expression we have written in expression of a variable. Now, it has to be converted, it has to be for a vector. So, in that case obviously this sigma i which is the variance of single variable of the samples taken from class omega i. Now, this sigma i or the variance sigma i square has to be replaced by another term, which is called co variance metrics, alright?

Now, I am dealing with number of variables are every variable represents one component of my feature vector. So, instead of simple variance I have consider a co variance, that is how what is the variance of x to respect x, earlier it was x now I have two multiple components. Now, I have consider the variance. So, x and x taken together, one is the x and x 4 taken together.

So, instead of a simple variance it will be covariance metrics. Similarly, this new I instead of being a scalar value, now it will be a vector value because this has to capture what is the mean of the components x 1, what is the mean of the components x 2, what is the mean of the components x 3 and so on. So, this mu i will again be a vector which is the mean vector of the samples taken from class omega i.

(Refer Slide Time: 18:27)



So, as a result the probability density function for a vector will be simply p of x is equal to upon. So, if you look at this particular expression, in the vector form the priority function becomes p of x is equal to 1 upon i to pie d to the pie by. And this d is the dimensional of this diameters that is the number of components that I have within this feature vectors. And sigma hello lecture and this sigma is nothing but the convenience metrics, sorry this will be this to the power of half.

So, I have to take the determinant of the convenience metrics and square root of that and within this exponential I will have minus half into x minus mu transpose sigma. In words x minus mu at this mu is nothing but the mean vector of all the samples which are of the sample values of the variable x. So, know the x is a vector variable, it is not a scalar variable and in the same manner if I want to put it is the class conditional probability that is p of x given omega j, which will be given by 1 upon pi. Now, this sigma has to be replaced by sigma j because. This is the convenient metrics of all the sample vectors which are taken from class omega j.

So, this sigma will now be replaced by sigma j. So, it will be determent by sigma to the power half and the exponents shall come will simply be minus half into x minus mu j transpose sigma j, in words x minus mu j. So, this will be the expression of the class conditional priority density function for all the vectors x or sample vector x taken from class omega j.

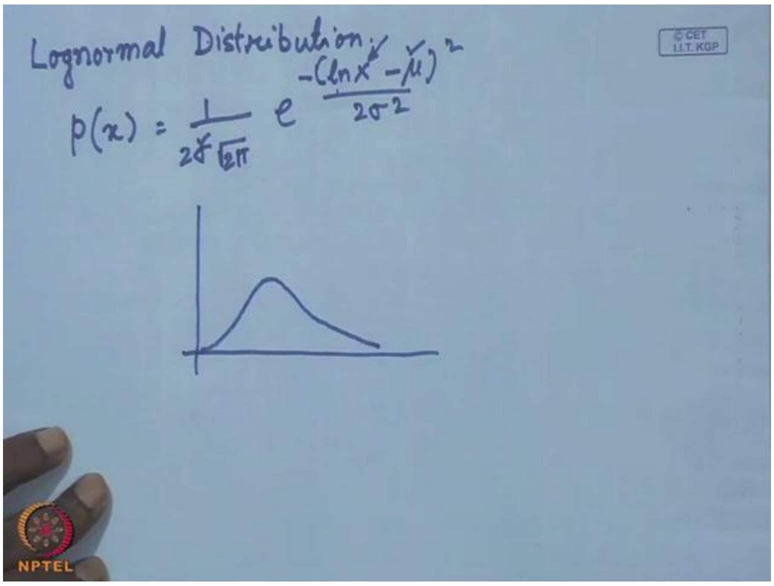
Now, with that I have this expression for a single variable or I have for a vector variable the expression which is something like this. As we said that you find shape of the probability density class is internally defined, decided a number of diameters. In case of a single variable it is decided by mu, which is the variance of x sigma square. In case of vector variables the shape of the probability density half is decided by the mean vector which is mu.

So, this mean vector decides the vocation of the probability and in case of a vector instead of a half it will be a. So, this mu sigma j decides that what is the position of that probability density, which the probability density function and the shape of the surface will be decided by covariance metrics, which is mu theory which is sigma theory. So, these are the parameters, the mean vector and the convenience metrics which determines, which is the position and the shape of the priority density function. So, these are the expressions whether it is this or this. These are called parametric expressions.

So, the probability density function has a parametric form and we find that previous lecture when we talked about the maximum likelihood estimate of parameter values. We assume that the probability density function follows a parametric form. That means I have or more parameters which decides what is the position and shape of the probability density surface or probability density function and knowing those parameters, what are the parameters we have tried to estimate.

We have tried to find out the Baye's estimate of those parameters form a same samples which are given. So, this is one such form of parametric representation of the probity density function and obviously this Gaussian is representation of the representation of the probability density function by mean and the variances, the only way in which a prior a parametric density priority function can be represented. There are many other parametric density functions.

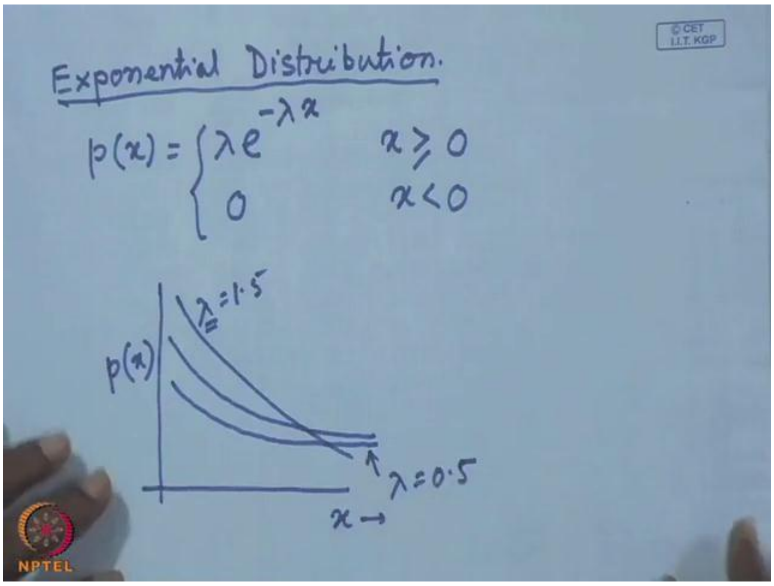
(Refer Slide Time: 24:47)



One of them is what is called log normal distribution. The log normal distribution is given by p of x is equal to over sigma square root of pie into 3 or minus log x minus u square f upon sigma square. So, this is an expression for what is called log normal distribution, and here in case you again find instead of log mu distribution, we have two different parameters. One of the parameters is sigma as before which is the standard deviation. And the other parameter is mu, as before this is the main of the sample values.

So, a notable difference between a Gaussian distribution or a normal distribution and a log normal distribution is that in case of normal distribution. Then you have found that what we have put is the exploration contains its minus mu square. In case of log normal distribution, the exponentiation of the exponent contents log of x minus mu square. If you do this distribution, the distribution will look something like this, which is similar to or Gaussian distribution, but with difference. The slops on this side and the slop on this side, they may be different or in general they are different. Similarly, there is another form of parameter distribution which is called exponential distribution.

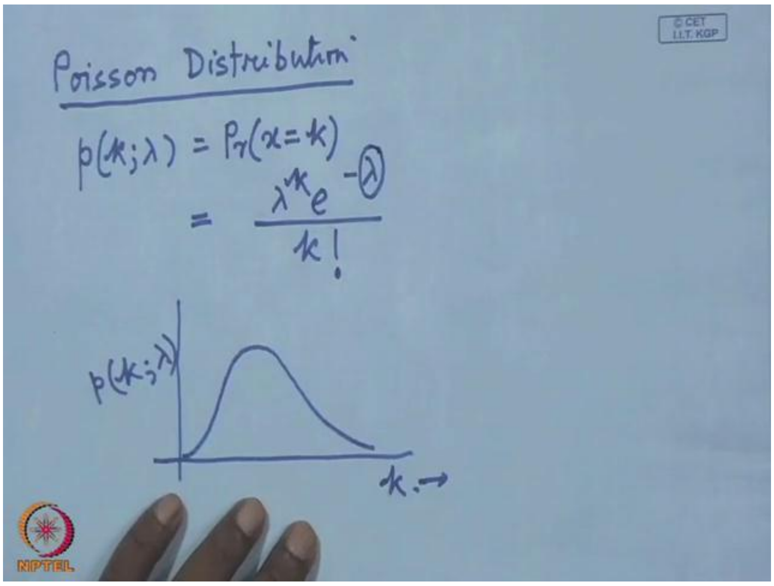
(Refer Slide Time: 26:57)



The exponential distribution is given by p of x is equal to lambda to the power lambda minus x, for x greater than or equal to 0. And this is equal to 0 for x less than 0. If I plot this exponential distribution, the plot will be something like this. For define values of lambda I will get plots like this for i increase, the lambda as the lambda is equal to 1.5. This may be for lambda is equal to 0.5.

So, this is my x and this is p of x, so again in case of exponential distribution you find that your probability distribution to the function is internally defined or decided by a single parameter, which is this lambda. So, if I can estimate for what is the value of lambda, then I know that what is the priority of what is the function of the priority distribution to the function, similarly another parametric distribution be class.

(Refer Slide Time: 28:53)

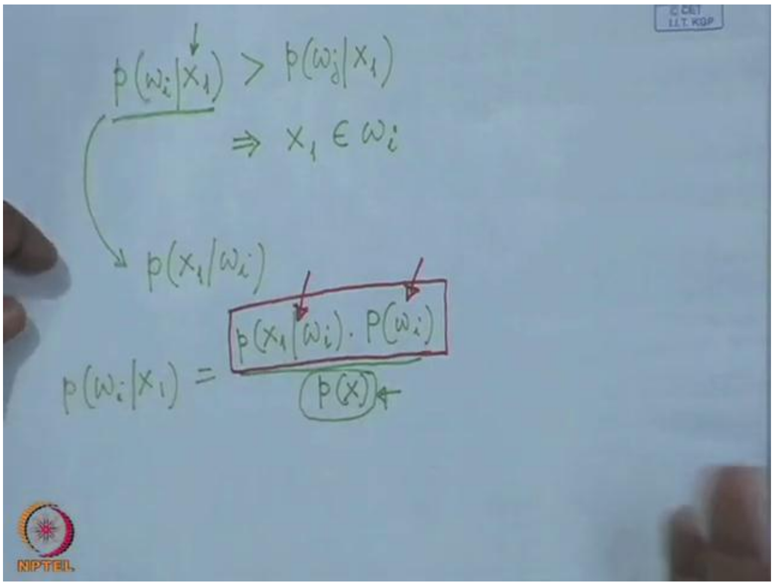


The distribution which is widely used in some of the applications is what is called Poisson distribution and the analytic form of the portion distribution is p of k. If an lambda is simply this actually represents the probability that the variable x takes a value is equal to k. And the analytic expression for this is simply the lambda to the power e to the power minus lambda upon a factorial and if I plot this portion distribution, the portion distribution again appears something like this. So, again will find, so this is what is k and this is what is p of k given lambda? So, for different values of lambda I get such different process and here again you find that the shape of this priority distribution functioned is internally defined by a single parameter, which is lambda.

So, my whole purpose of telling you this is that when I go for say like maximum likelihood estimate of the priority density functions, I have to know that what is the parametric form of distribution function, unless I know what are the parameters which distributes the priority function. I cannot use the maximum likelihood estimate, after all gives you the Baye's estimate of the parameters. Computed the Baye's estimates, that now all problem is when I go for classification problem. Now, particularly of Gaussian distribution or to this parametric distributions, all parameter this parameter values hided, I have large number of samples representing this distribution.

Now, if the samples the number of samples is not very large, then what if the parameters I would do that may not be faithful representation of the actual. And the major problem is I do not know beforehand that what is the parametric form that this distribution is going to follow, he parameters are not known, so unless I know the parametric form I cannot estimate the parameters and in most of the situations I do not know the parametric form. So, given such situations how to still classify the unknown samples following this classification? So, if you look at what this Baye’s classification rule says in case of this column rules, what we have said is simply p of omega I given an unknown sample say x.

(Refer Slide Time: 32:32)



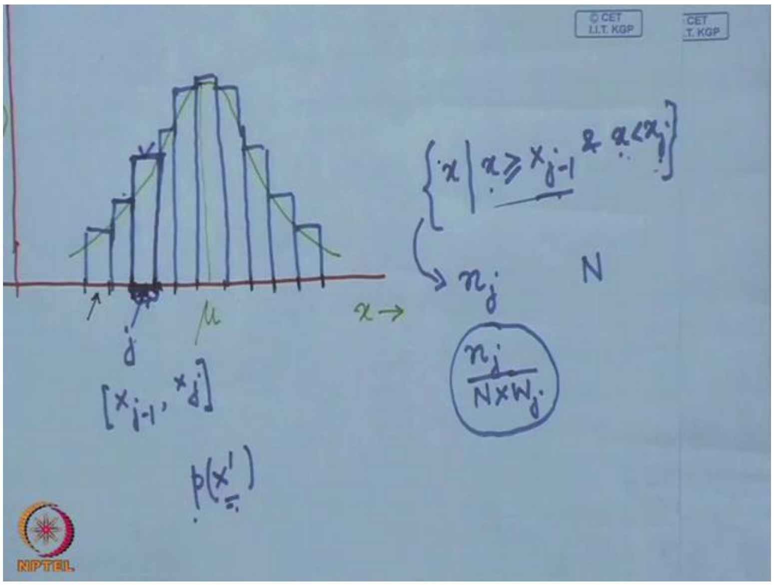
If it is greater than p of omega j for the same unknown sample say x, then we decide that x belongs to class omega i. So, what I need to know is what is this aprostery probability p of omega given i x. So, I need to compute this only for this unknown sample which is x, and to compute this the probability density function or the probability density value that I have to use is what is p of x given omega i because as we said earlier that p of omega i given x is nothing but p of x given omega i into probability p of omega i given p of x. As we said that p of x appears in the denominator for such expressions, for different value of i defined class omega i.

So, this p of x does not contribute anything throughout the section because it is same for all the class. So, what I really have to compute is what is p of x given omega i into p of omega i being apparel probability, it is always assume that is known. So, what I have to compute effectively is this value, what is p of x given omega i. Now, the advantage is that I have if I have a parametric representation something like this.

So, this being an analytic expression for given any x, I can immediately compute this value. So, that is the advantage that I have, but for that I have to know that these are the parameters which defines the parametric, that particular parameters function. And if I know that these are the parameters, then only I can go for the Baye's possible estimate. This parameters from the known side parameters, but all problem is what is the parametric form of the probability distribution. And if I do not know what is the parametric form, then I cannot form any analytic expression like this.

So, my job will be that I have to compute p of x given omega i, where x is an unknown sample making use of the known samples which are provided for supervising. So, my problem is that, so I have a set of samples and using those set of samples I have to estimate what is p of x, what is given on a. Obviously, I do not assure that this probability distribution function follows any known parametric form. So, how we can do that that is the problem that we are going to discuss today and may be tomorrow and day after as well.

(Refer Slide Time: 36:48)



So, let us start with this, let us start with the known form or Gaussian distribution function which was something like this. I had this mean location mu and for certain variance sigma square, the density is something like this. So, this is my variable x, so this access represents variable x and this access of x for p of x which I know, what is mu. I know what is sigma? And this is defined that was called for this, I have analytic expression for this is nothing but Gaussian expression. Now, what I can do is this analytical half I can represent by peace wise constant approximation.

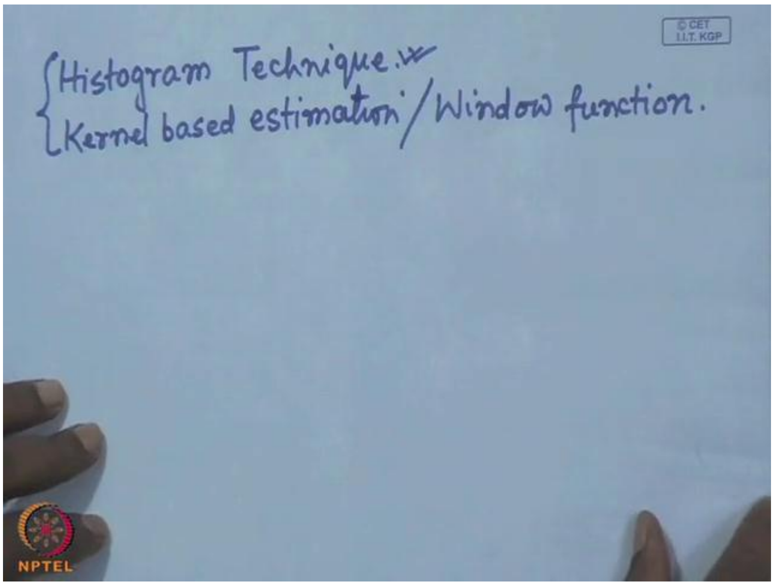
So, what I mean by peace wise constant approximation? What I do is this access representing my variable x i, divide this access into on state of intervals, something like this. I have intervals given like this. So, these are the different intervals in which the x access or the access presenting the variable is divided and I assure that in each of this interval, the probability density is constant. So, the way I can represent this contain was called by a peace wise constant representation is something like this.

So, this is a peace wise constant representation of this continues probability density. Now, was I do that if I take any of this intervals j th interval defined that the j th interval the height of this bar that have that indicates what is the fraction of the samples that falls under this particular interval w, right? So, for all the sample x, suppose the boundary of the interval is given by x j minus 1 to say x j. So, these are the limits of this particular interval, so all x are x is greater than or equal to x j minus and x is less than x j. If I take the set of all such samples, other the sample value is greater than x j minus or the x j minus is the left boundary of the j th interval. And it is less than x j or the x j is the right boundary of the center and the number of such samples if represent by in j.

So, j is in the number such samples which follows this condition and suppose in capital N is the total number of samples that I have. Then this height actually represents in j by N into width of this interval, if i say w j that is the width of the interval w j. So, this is nothing but what is the probability that a sample falls within this j th interval. Now, given any other value of x pine which is unknown and I have to compute what is p of x pine. What we have to look for that in which of this is intervals x pine falls.

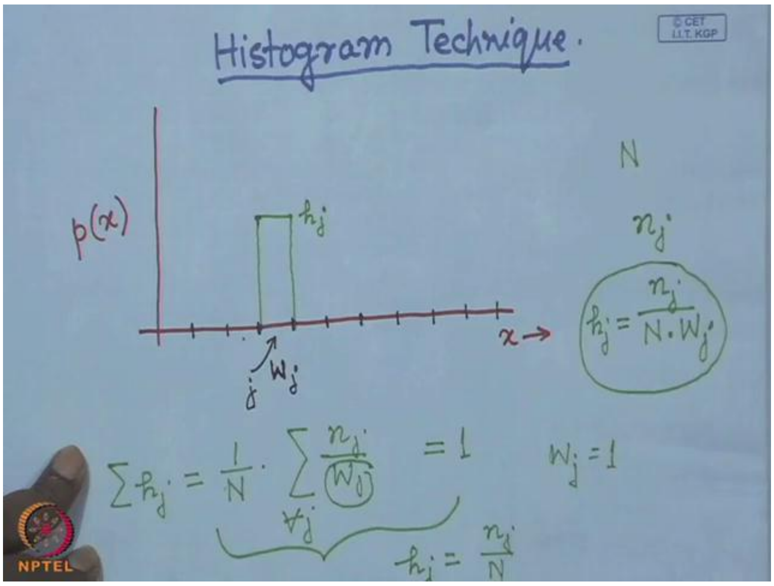
So, it x pine falls within this j th interval, then immediately I can say this p of x pine is given by the height of this bar. So, given any analytic expression I can always make a peace wise constant expression of that analytic expression of that probability distribution which is given like this. So, this gives as int that even if I do not know what is the parametric form of the probability density function, but still I can estimate what is the probability of a particular unknown sample, so there are two ways.

(Refer Slide Time: 43:32)



So, there are many ways discuss mainly two of them, one of them is called histogram technique and the other one is called canal Baye's destination or this is also called making use of in the function. So, we will mainly discuss about these two different techniques for probability density function estimation. So, first will talk about this histogram Baye's technique.

(Refer Slide Time: 44:31)



So, let us see what is the histogram Baye's technique for probability density function? So, we will talk about, so will first explain the concept using a scalar variable x or I have a scalar variable x and then we can extend the concept to multiple dimensions, when we have features which are not scalar, but we feature vectors having multiple number of components. So, let us see that how we can do it for a single variant x or scalar variables. So, first what we do is because it is a scalar variable. So, I will have horizontal access which have variable x or different samples of the variable x, and the vertical access will represent the probability density estimate which is p of x.

So, as we have seen earlier that when we have gone for peace wise constant approximation of a continues probability density class were we have said that we divide the variable access or the x access into a number of intervals. Now, width of different intervals may be same or width of different also be different. In general intervals have different width, but for simplification we can make the way of width equal all the widths to be same. So, this is the kind of situation that we had, following the same concept what we do over here is we divide this x access into a number of intervals.

So, the intervals are like this, out of this I take a particular interval say j th interval for which the width will be given by W j. Then what I have is the situation is something like this, I have large number of samples which represent different sample values of the variable x. When I tried to estimate the probability density function along the access, I put all those different sample values of x and along the vertical dimension I will put the estimate value of the probability density function. And for doing this what I have done is out of all different intervals, I have taken by j th interval whose width I have assumed to be W j.

So, what I need to do is I here all the samples and tried to find out that out of all those samples, suppose I have total of capital N number of samples. So, out of this capital N numbers of samples are to find out that how many samples actually fall within this j th interval. Now, while doing so it may so happen that the sample with the following the boundaries, either the j minus first interval or j th interval or the boundary between the j th interval and the j th first classified. So, in such situations obviously there is an that whether that is sample should be considered to belong to j th interval or the sample should be considered to j minus first interval.

Similarly, if the sample falls on the boundary between the j th interval and the j th first interval, that is an ambiguity that whether the sample should be considered to belong to j th interval or it should be considered to j th first interval. So, to avoid such ambiguity conventionally what it goes is, whenever I have sample falling on the boundary conventionally the sample is considered to belong to the to the interval which is right of it.

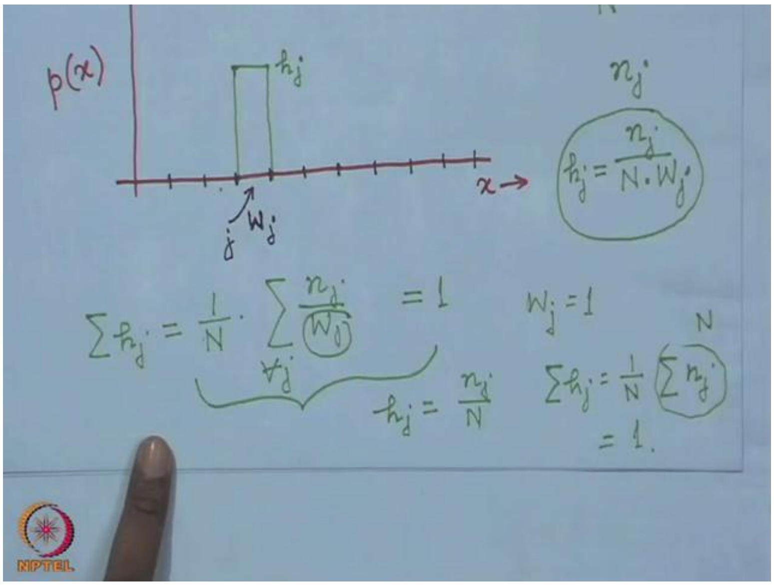
So, if I have a sample for you are here I will consider that the sample belongs to the class of j interval. If I have a sample following over here, I will consider that sample to belong to j th interval. So, by considering this if I find that I have in a j number of samples which falls within the j th interval, then I will have a bard of representing the probability function. But the height of the bar will be given by in j by N into W j, at this W j is width of the j th interval, capital number is the total number of samples that I have and n j is the number of samples. Out of these capital number of samples the number of samples which falls in the j th interval and this is what will give me the j th, the height of the bar representing this probability density function.

So, I will have a situation something like this, I have a bar, rectangular bar just over this where this height of rectangular bar in this j. So, which actually represents that what is the probability that x will have a sample value within this j th interval value. Now, this particular computation has significance. The significance is we assume that this histogram, so collection of bars is nothing but histogram.

So, if this histogram has to represent or approximate a continuous probability density function, then they are not continuous distribution function. When area under this probability distribution function has to be equal to or in other words, I must have some of h j which is nothing but 1 upon n into some of h j by W j, for all j that must be equal to and you find if I compute this h j in this form this condition will always be satisfied.

Now, in a particular case that if assume that this W j is equal to that is every interval has an unit with in that case h j, that is the probability the density estimate in this interval w in this j th interval will be simply be equal to n j upon capital n. So, it will be something like this in j upon capital n, and in such case this is quite obvious sum of h j will be simply 1 upon capital N into sum of n j.

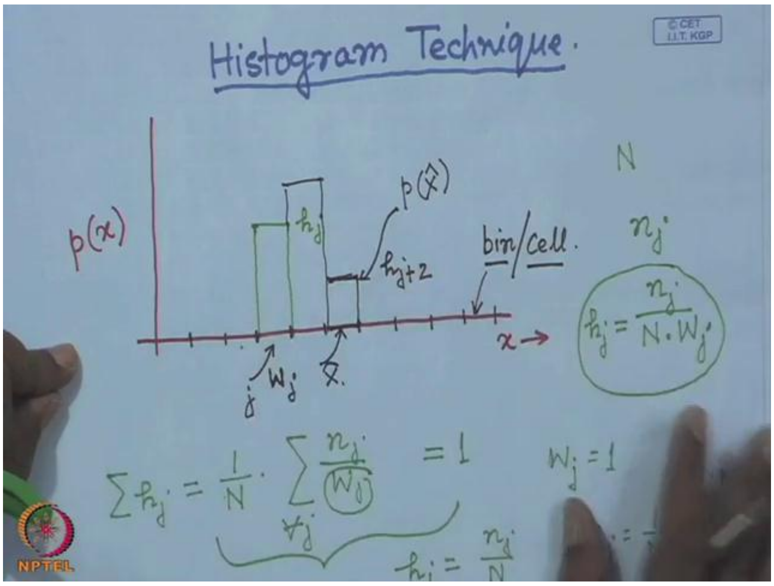
(Refer Slide Time: 53:19)



The sum of n j is nothing but capital n. So, this will be obviously equal to which is also equal to this particular test and that can be easily proved. So, given this what I have to do is in summary I have a large number of samples of the variable x, the x access of the probability of density function. When I bought the probability density function the x access which represents my variable x, it is divided broken into a number of intervals and for breaking this number of intervals, I have to have an idea of what is the minimum all the value of x and what is the maximum value of x or in other words what is the range of x.

So, once I have this knowledge of the range of x, that range I can always divide into a number of intervals. Then for this given number of samples, out of this total number of samples I had to find out, what is the fraction of samples which falls within the particular interval. The fraction of samples divided by the width of the interval tells me that what is the probability density function estimate of the probability density, who have that particular interval.

(Refer Slide Time: 55:08)



So, once I have this kind of estimate, so this h j is the priority density for this j th interval. This the probability of j of first interval, this may be probability density function, the second interval and so on. So, if I have any unknown value of x say x over here, then immediately I will say that value this is x hat a sample value x hat, immediately I will say that p of x hat is nothing but this, which is nothing but h j plus that is the height of the bar in j plus second meter.

So, obviously it is here, assumed that within an interval the priority density is constant and the probability density will be different is likely to be different in different intervals. So, when I this x access or the variable into a number of interval something like this, each of these interval is usually called a beam or aspire or I can either call it a beam. So, this is how the histogram technique works, histogram probability density function works. So, in the next class will explain for the help of some examples, and then will extend this to multi dimension when x instead of being a scalar variable it is affected.

Thank you.