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

Lecture - 10
Maximum Likelihood Estimation

Good morning. So, today we are going to discuss about the parameter estimation of the probability density functions, and the particular type of estimation technique that we will discuss is called maximum likelihood estimation. So, till now what we have discussed is we have talked about the discriminant functions for different classes and we have seen that, the type of discriminant function which is very popular is nothing but the logarithmic function of the probability density function. So, when we take any particular probability density function, the probability density function is specified by a number of parameters.

So, in particular what we have taken is the normal or Gaussian distribution and in case of normal or Gaussian distribution we have said that, there are two parameters which uniquely identify this PDF. One of them is the mean vector and the other one is the covariant matrix. So, these are the parameters that is mean vector and covariance matrix so this is given, you know what is the nature or structure of the probability density function, which is involved in that case.

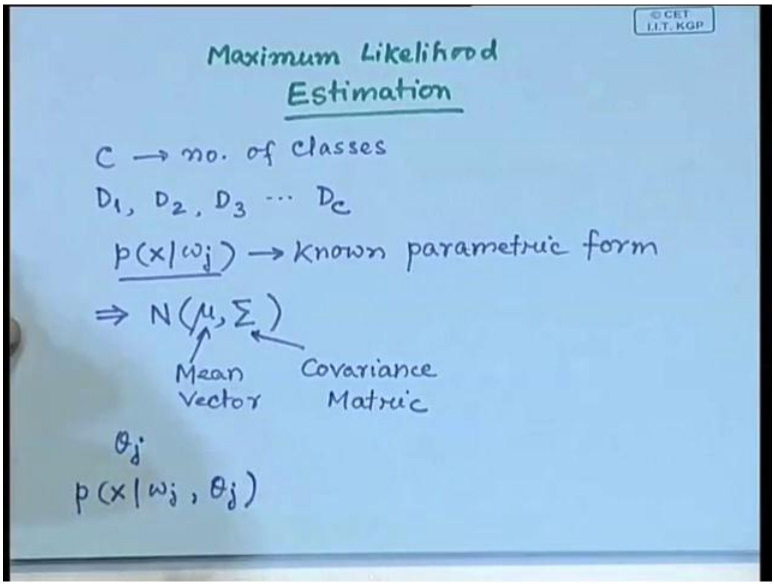
And so far what we have discussed is, we have taken the logarithmic of the probability density function or a combination of it, combined with the a priori probability of different classes to give us the discriminant function for different classes. And one you have taken the decision boundary between two different classes, and then we have seen that it is, nothing but the difference of the discriminant function.

And we have also seen that, depending upon the nature of this parameter vectors or the nature of the mean vector and the covariance matrix, we can have different types of boundary, the decision boundary between two different classes. You can have linear boundary, we can have a linear boundary which is and which is octagonal to the line joining the mean vectors. We have got linear boundary, but which is not octagonal to a line between two mean vectors and we have also seen that, we can also have in general the decision surfaces which are, nothing but quadrics in d dimension. Where d is the number of components in the feature of vector which, we use for classification.

So, effectively when we design such a kind of classifier or we want to find out the decision surface between two different classes, the nature of the probability density function is very very important. And as such, as the nature of the probability density function is defined by the parameters mu and sigma that is, mean vector and covariance matrix. We have to accurately estimate, the mean vector and covariance matrix given, a Gaussian distribution, we know the form of mean and covariance matrix. But if the probability density functions is other than that, other than the Gaussian or normal distribution, in that case we have to find out what is the parameter vector that identifies the probability density function.

So, today what we will be discussing about is a particular technique for parameter vector estimation, for a known parametric form of the probability density function. So, the type of estimation techniques that we will discuss about is the maximum likelihood estimation. So, the problem is something like this.

(Refer Slide Time: 04:11)



Maximum Likelihood Estimation

$C \rightarrow$ no. of classes

$D_1, D_2, D_3 \cdots D_c$

$\underline{p(x|\omega_j)} \rightarrow$ known parametric form

$\Rightarrow N(\mu, \Sigma)$

Mean Vector

Covariance Matric

θ_j

$p(x|\omega_j, \theta_j)$

Suppose we have got c number of classes so you have c number of classes and we are assuming that, the kind of classification, of the kind of learning that we are going to employ is supervised learning. That means for each of the classes, we have aesthete of samples for which I know that, to which of the classes this set of sample belongs.

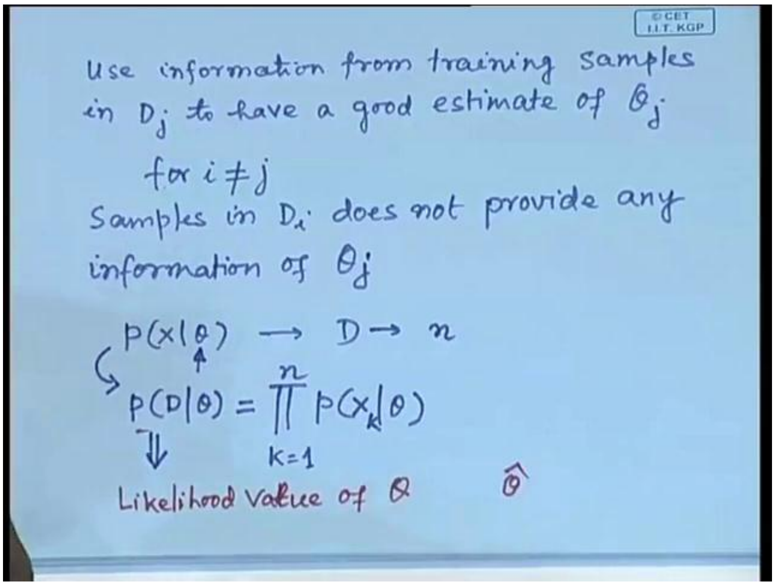
So, as we are considering that we have c number of classes so naturally we assume that we have c sets of feature vectors and let us represent that, c set of feature vectors as feature vectors D1, D2, D3 up to Dc, as we have c number of classes. So, you can assume that the set of samples in each of the classes is something like this, that these set of samples are drawn independently according to the probability law p x given omega j. So, the samples belonging to class Dj or belonging to set Dj within these set of samples is drawn independently according to the probability law p x given omega j. And we also assume that, this p x given omega j has a known parametric form.

So, in particular this known parametric form if it is a normal distribution or a Gaussian distribution then we have said that, this known parametric form is uniquely identified, is uniquely defined by the parameters, which are mu and sigma. Where, mu is the mean vector and sigma is the covariance matrix. So, these are the parameters which are, known for normal distribution or Gaussian distribution.

Similarly, I can have other types of distributions which are also parametric, but the parameters may be different from this. So, what I assumed is that the samples in set Dj is drawn, are drawn independently according to the probability law p x given omega j, but this p x given omega j has a known parametric form. So, it has a known parametric form so in general I can assume that for class omega j, the parameter vector is a vector given by theta j.

So, this is my parametric probability distribution function, probability density function which has a known parametric form or the parameter is given by, this parameter vector theta j. So, to write explicitly the dependence of this PDF on this parameter vector theta j, I can write this as p x given omega j theta j. So, this specifies that the probability density function has the parameter vector theta j. So, our problem in this case is that we have to use the information available from the samples in the set Dj, to estimate this parameter vector theta j. And the kind of technique that we are going to use, we are going to discuss for this estimation of parameter vector theta j is, what is known as maximum likelihood estimate.

(Refer Slide Time: 08:49)



Use information from training samples in D_j to have a good estimate of θ_j

for $i \neq j$

Samples in D_i does not provide any information of θ_j

$p(x|\underset{\uparrow}{\theta}) \longrightarrow D \rightarrow n$

$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$

$\Downarrow$

Likelihood Value of θ $\quad \hat{\theta}$

So, what you have to do is, we have to use information from training samples in set Dj, to have a good estimate of the parameter vector theta j. While doing so we also assume that the samples in set Di does not provide any information about the parameter vector theta j, for i not equal to j. So, for i not equal to j samples in Di does not provide any information of the parameter vector theta j whenever, i is not equal to j that means, the classes are or the samples belonging to different classes are unique independent. So, samples from one class do not provide any information of the parameter vector, of the probability density function of another class.

And as we have also assume that, the samples are drawn independently so I can make use of these individual samples to, estimate our probability density function, and because the samples in Di, does not provide any information about the parameter vector theta j. So, effectively what we have is, we have c number of independent problems so given a set of samples I have to estimate the parameter vector of the probability density function, which best estimates the data distribution within that particular set. So, I can simply get away with this subscripts i and j and all that, I can simply write p x given theta.

Because, I have independent c number of problems, I have a set of samples for that given set of samples, I have to estimate the parameter vector theta of the probability density function, which best describes the data distribution of that particular set. So, instead of using this subscripts I can simply write that, I have to estimate p x given theta or I have been given a set of samples. So, set of samples D and using the information available in these samples, I have to estimate this particular parameter vector theta. And as the samples are drawn independently and if I assume that D contains n number of samples.

Student: Sir, what is p x theta d given theta?

See, what I have said is p x given omega j theta j, what is the interpretation of this, this is the probability density function of the set of samples in set Dj. I know that, the samples in set Dj actually belong to class omega j and theta j is the, parameter vector which specifies this probability density function. And because of our assumption that, the samples in Di, does not provide any information of theta j. I have c number of such sets of samples, D1 to Dc as I am considering that have c number of classes and each of these set of samples, independently provide me the information of the parameter vector, of the corresponding class. D1 does not give me any information about theta2 similarly, D2 does not give me any information about theta1.

So, effectively what I have is, I have c number of independent problems given, the set of samples D, I want to estimate the parameter vector theta, of the probability density function which best describes the data distribution in set D. So, because the samples in one set is not giving me information about the parameter vector, of the other set corresponding to the other set. So, I can simply remove this subscript j or even I can remove this, cross conditional probability omega j because what I have is, given a set of samples, I want to estimate the corresponding theta.

Student: Sir ((Refer time: 14:38)) basically that are some descriptor of the.

It is the parameter, it is the parameter vector that describes the probability density function.

Student: Which can be like weight Sir? weight ((Refer time: 14:49))

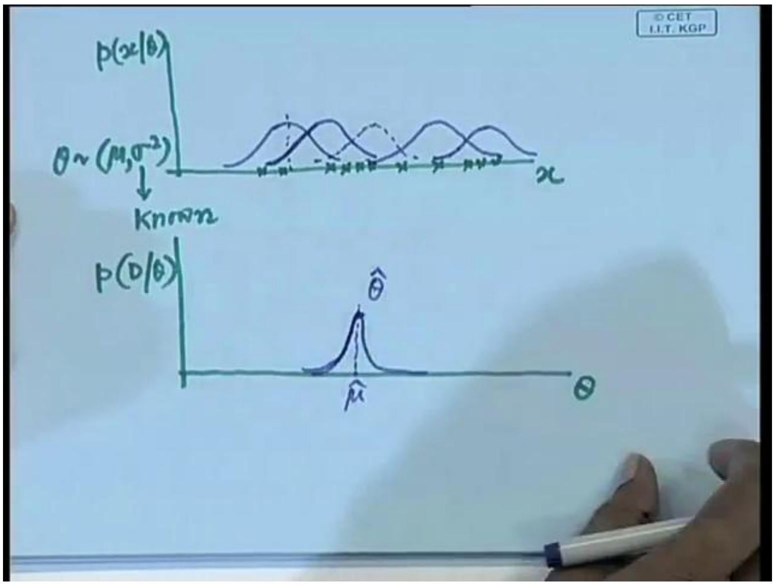
No, theta j simply tells the what is the shape of the probability density function like, in case of Gaussian distribution or normal distribution theta represents the components of mu and the covariance matrix involved. Similarly, for other probability density function the components of theta will be something else.

So, what I have is, I have a problem that, I have set of samples D, I am not bothered about which set is or from which class the samples has been drawn. I simply have a set of samples D, whose distribution is actually specified by the paramter vector theta. So, instead of this form the p x given omega j theta j because now omega j is no more important to me because I have separate separate problems. So, I am simply rewriting this in the form p x given theta, this theta specifies what is the probability, what is the nature or shape of the probability density function.

So, given this if I assume that D, the set contains n number of samples which are drawn independently. So, because they are drawn independently so I can also write from here what I have to do is, I have to make use of the information available in the set of samples to, estimate the value of theta. So, I can write this because there are n number of samples I can write that, p of D given theta in another form, theta specifies the parameter vector. So, I can write it in this form p of x given theta, product of this, let me put a subscript p of xk given theta, k varying from 1 to n. Because, I have n number of samples in this set D and the samples are drawn independently.

So, this can be written in the product form and this term p D given theta, this is what is known as likelihood value of theta. And the maximum likelihood estimate of theta is, the value set theta hat, which maximizes this function. When this function will be maximized and actual the set of samples which are drawn and the parameter vector theta, the estimate of the parameter vector which, best describes the probability density function of the distribution of the set of the samples xk. So, that time this value will be maximum and the value theta hat, which maximizes this likelihood value of theta, that is called maximum likelihood estimate of this parameter vector theta.

(Refer Slide Time: 18:58)



Now to be more specific, let us have a diagram, suppose I consider univariate case. This is x and on this side I plot p x theta and assume that because this is univariate case. I assume that, this theta is mu sigma square, where mu is the mean and sigma square is the variance. And suppose I have point distributions something like this, some arbitrate distribution something like this and this theta, the parameter vector is nothing but mu sigma square. Where mu is the mean and sigma square is a variance.

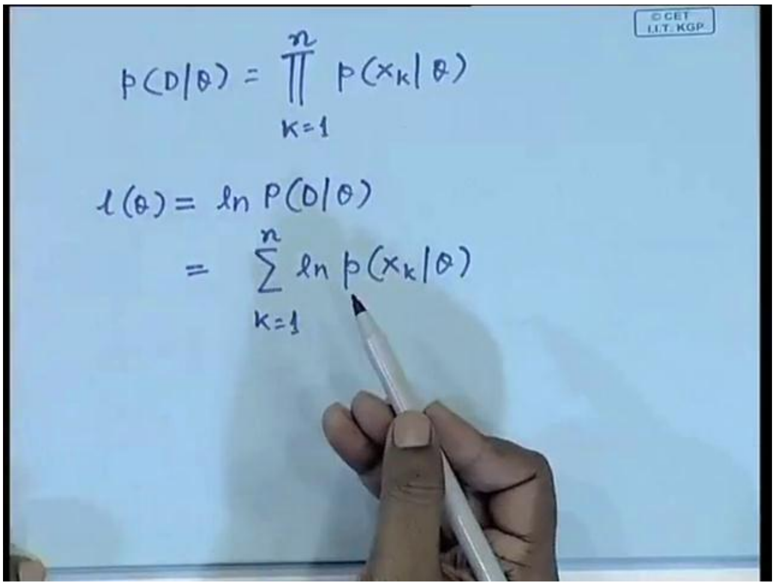
Now out of this, if I assume that this sigma square is known, I am just simplifying the case. If I assume, that sigma square is known then what is unknown is mu so from these point distribution, I have to estimate the value of mu. So, we are fine that for different values of mu, I can have different types of probability density functions like this, p x given theta. When I assume that, the value of mu is this for some other value of mu, the nature of probability density function will remain the same, but it will shift. Now, why this will remain the same because I have assumed that sigma square is known.

So, I can have this different types of distribution functions, but the variance value is same, but the mean value is different. Now, find that all these different distributions they do not really, all of them does not really represent the distribution of this data set. Only one of them will faithfully represent the distribution of this data set. So, for that this likelihood estimate, the likelihood value that p D given theta is very important.

So, if I plot this p D given theta you find that for different values of x this p of xk given theta, a theta is nothing but this mu we have different value. And here for the likelihood value of theta, what I am taking is, I am taking product of these values. So, if I plot similarly, now I plot with respect to thetap of D theta. You have find, you will find that this plot will be something like this and this will have a peak at theta is equal to theta hat.

And in this case, this will have a peak because theta is nothing but equal to mu assuming that, sigma square is known. So, it will have a peak at theta equal to mu hat so this is what is the maximum likelihood estimate of this p x given theta, which best represents the mean value of this data distribution so I have to find out this position of the mean, mu hat or in this case theta hat.

(Refer Slide Time: 23:30)



$$p(D/\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

$$l(\theta) = \ln P(D/\theta)$$

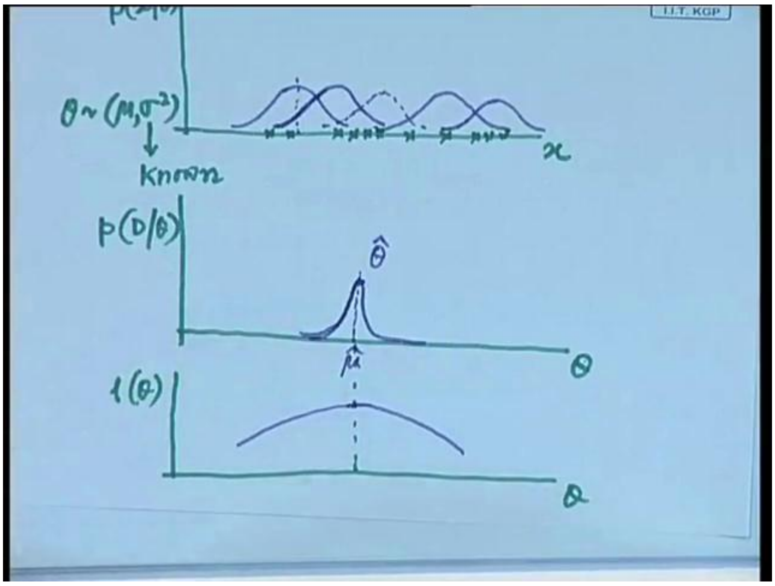
$$= \sum_{k=1}^{n} \ln p(x_k|\theta)$$

So, this maximum, the likelihood value of theta as we said is represented by p D theta, which is equal to p of xk given theta, take the product k is equal to 1 to n. Where n is the number of samples present in the set D. And you have also said earlier that, instead of taking this likelihood value, as we have said in case of our, that discriminating function, that instead of simply taking the probability density function as the discriminating function if you take the logarithmic of it, that helps us in analysis.

So, similarly here you find that this likelihood value is a product term so instead of taking this likelihood value directly, if we take the logarithm of this likelihood value, that is what is the log likelihood. So, I can define the log likelihood as l theta which is nothing but log of p D theta. And which will simply be p of D given theta is of this form, which is a product term, when I take the logarithm it will become a summation term.

So, this will become further to log of p xk given theta ,where k value is from 1 to n. So, this is the log likelihood, this is the likelihood value of theta, this is the log likelihood value of theta. And because we are trying to estimate the value of theta I simply represent it, represented these as l of theta. Assuming this log likelihood is a function of theta, which we are trying to maximize. And coming to this particular curve, instead of likelihood if I put log likelihood the position of the maximum will remain the same, but the curve will be smooth.

(Refer Slide Time: 26:01)



So, if I plot l of theta versus theta, the position of the maximum will remain the same, but the nature of the curve will be something like this. So, it will be a smooth curve while this is a sharp curve and here you find that, as we have more and more number of points this curve will be narrower. As we have more number of points, this curve will be broader so that clearly says that if I have a narrower curve, I can estimate the maximum more accurately. If it becomes a broader curve then, the maximum estimation may not be that accurate because around maximum I have a flat curve.

Over here or over here if it becomes a broad one, around maximum I will have a flat curve. So, the maximum estimation is not that accurate and if I have more number of points more and more number of samples, this curve will be more and more narrow

((Refer time: 27:11)) the confidence in estimating the maximum is much more. If I have very less number of points then, the confidence in estimating the maximum is less so that is what it means.

Student: Sir sir will this data had been changes or will remain same.

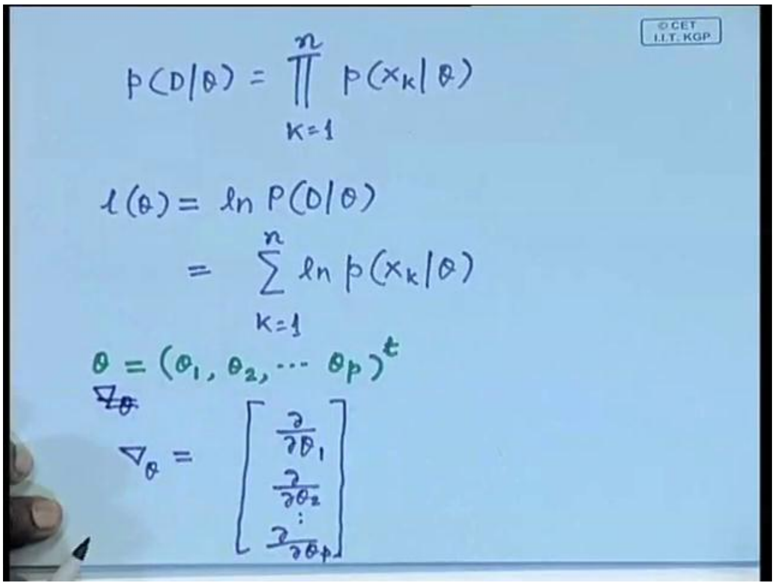
Pardon.

Student: This data had changes or it remains the.

That is what I said, that depends upon the number of values that I have. Our confidence in estimating theta hat, depends upon the number of samples I have. If I have more number of samples, my estimation of theta hat is more accurate, if I have less number of samples the accuracy is not that high.

So, this is what is log likelihood whether I use the likelihood value or the log likelihood value, the estimate of theta hat it will also, it will always give me the maximum likelihood estimate. Because, the logarithm is a monotonic function now to maximize this, we know the formula of our school level mathematics, that we have to use the concept of differential calculus. So, what I have to do is I have to differentiate this l theta and equate that differential to 0. Because, here theta is a vector so instead of simple differentiation I have to take the gradient.

(Refer Slide Time: 28:37)



$$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

$$l(\theta) = \ln P(D|\theta)$$

$$= \sum_{k=1}^{n} \ln p(x_k|\theta)$$

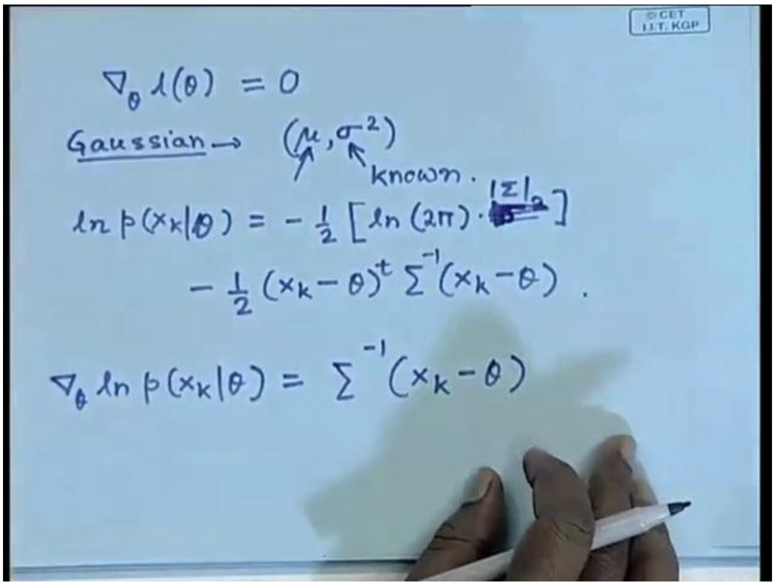
$$\theta = (\theta_1, \theta_2, \cdots \theta_p)^t$$

~~∇_θ~~

$$\nabla_\theta = \begin{bmatrix} \frac{\partial}{\partial\theta_1} \\ \frac{\partial}{\partial\theta_2} \\ \vdots \\ \frac{\partial}{\partial\theta_p} \end{bmatrix}$$

So, I have to take grad theta gradient with respect to theta and if this theta has got say, p number of components. So, if theta has got p number of componets which are defined as theta1, theta2 up to say thetap. I put it as transpose because theta is a vector so if it has got this p number of components then, grad theta operator will be simply del del theta1, del del theta2, del del thetap, this is the gradient operator. So, what I have to do is, I have to find out the gradient of the likelihood value of thetas.

(Refer Slide Time: 29:43)



$\nabla_\theta l(\theta) = 0$

Gaussian $\rightarrow (\mu, \sigma^2)$ known.

$\ln p(x_k|\theta) = -\frac{1}{2}\left[\ln (2\pi)\cdot |\Sigma|_2\right]$

$-\frac{1}{2}(x_k-\theta)^t \Sigma^{-1}(x_k-\theta)$.

$\nabla_\theta \ln p(x_k|\theta) = \Sigma^{-1}(x_k-\theta)$

That is, I have to find out gradient of l theta where, the gradient has to be with respect to parameter vector theta and I have to equate this to 0. So, when I equate the gradient equal to 0, I get a number of equations, simultaneous equations. I have to solve that set of simultaneous equations because this gradient has got p number of components because the gradient will have p number of components. And I equate this gradient to 0 that means, I will get p number of simultaneous equations and when I solve this p number of simultaneous equations, I get the components of the parameter vector theta. That is simply what you have done in our school level mathematics.

Now here the problem is, if I equate the gradient to 0 to get a number of simultaneous equations, I may lead to the actual mean or the global maximum or I may lead to local maximum not only that, simply equating gradient to 0 also may give me the minimum. So, again you know that if I take the second derivative, that is gradient of the gradient, the second derivative will be negative if it is the maximum, the second derivative will be positive if it is a minimum.

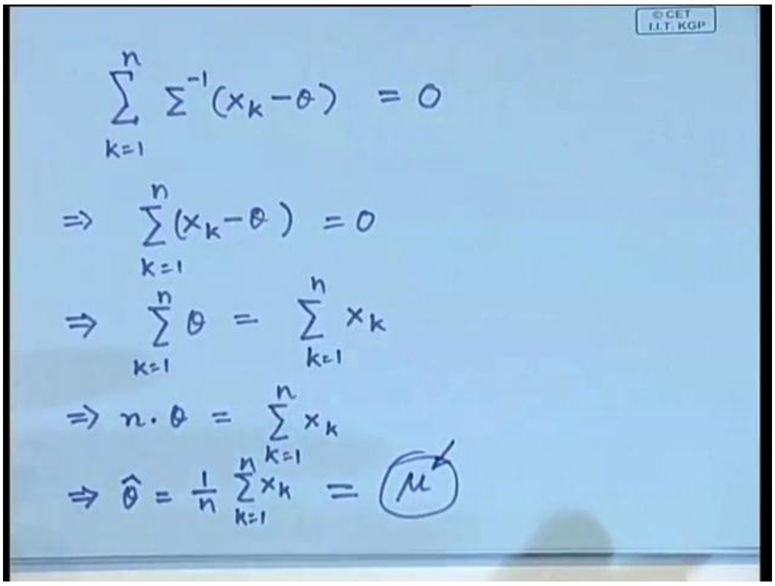
So that, where I can filter out which ones represent the maximum values and which ones represent the minimum values. I am interested only in the maximum values so I get a set of values of theta which represent, the maximum. Then for each of them I have to actually find out what is the functional value, to identify what is the global maximum. So, that is what I have to follow to find out the maximum likelihood estimate of a parameter vector.

Now let us see that, this one really gives us what we want that is, the validation of your procedure. To validate this, let us take some known cases and let us see that, for those known cases whether I get, really what I want. So, for validation I will take the same Gaussian case or the normal distribution, for which I know that the parameter vector which is given by mu. And for simplifiaction, instead of talking multivariate Gaussian, let me take the single variate case. So, the parameter vector is actually given by mu sigma square and initially for simplicity, let me assume that sigma square is known. So, what I have to find out is the mean mu.

So, let us see whether by using this maximum likelihood estimate procedure, I really get the mean value or not given a normal distribution with mean mu and a known variance sigma square. So, for this Gaussian distribution we know that, the log likelihood will be given by ln or the probability density function xk given, theta is nothing but minus half, this was nothing but sigma square minus half. Instead of sigma square, let me put it as this so this is my logarithm of the probability density function. So, what I want to do is, I want to differentiate this with respect to theta and in all case theta is equal to mu.

So, if I take theta ln p xk given theta, this will be simply, you find that here this term is constant and our variance sigma square is known. So, this term is also known so if I differentiate this, this term will be equal to 0. So, what I have to do is, I have to simply take the differentiation of this with respect to theta, which is a function of theta. So, if I differentiate this, it will simply become sigma inverse xk minus theta. And you find that, in the log likelihood I have to take the summation of this, for all values, for all the samples that is k equal to 1 to n and I have to equate that to 0.

(Refer Slide Time: 35:40)



$$\sum_{k=1}^{n} \Sigma^{-1}(x_k - \theta) = 0$$

$$\Rightarrow \sum_{k=1}^{n}(x_k - \theta) = 0$$

$$\Rightarrow \sum_{k=1}^{n} \theta = \sum_{k=1}^{n} x_k$$

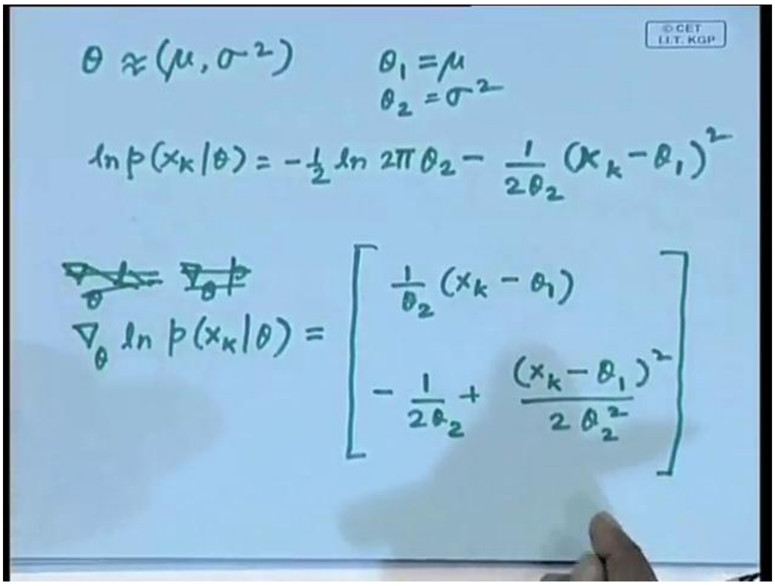
$$\Rightarrow n \cdot \theta = \sum_{k=1}^{n} x_k$$

$$\Rightarrow \hat{\theta} = \frac{1}{n}\sum_{k=1}^{n} x_k = \boxed{\mu}$$

So, effectively what I have to have is, I have to eqaute sum of so let me put this as sum sigma inverse xk minus theta, where k will vary from 1 to n, this should be equal to 0. So, from here you have find that, if I premultiply both sides by sigma that is covariance matrix then, what I simply get is summation xk minus theta, where k varies from 1 to n, this will be equal to 0. So, this simply gives us that summation of theta is equal to summation of xk, where here also k varies from 1 to n, here also k varies from 1 to n. What this means is, I have to sum theta n number of times so this is simply n multiplied by theta because it is the same theta which has been added n number of times.

That is equal to sum of xk, k varying from 1 to n, so this simply gives us the parameter theta, which is nothing but the maximum likelihood estimate theta hat is equal to 1 upon n sum of xk, k varying from 1 to n which is nothing but the mean mu. So, I started with this assuming the mean is unknown, we have followed the procedure of maximum likelihood estimate of the parameters. And we have really landed to, the unknown parameter which is the arithmetic mean of the set of samples that I have, which is nothing but mu. Now, let us take the more general case assuming that, neither we know mu nor we know sigma square, that is both mean and variance both of them are unknown.

(Refer Slide Time: 38:10)



©CET I.I.T. KGP

$\theta \approx (\mu, \sigma^2)$ $\quad \theta_1 = \mu$

$\theta_2 = \sigma^2$

$$\ln p(x_k|\theta) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

~~$\nabla_\theta l = \nabla_\theta p$~~

$$\nabla_\theta \ln p(x_k|\theta) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

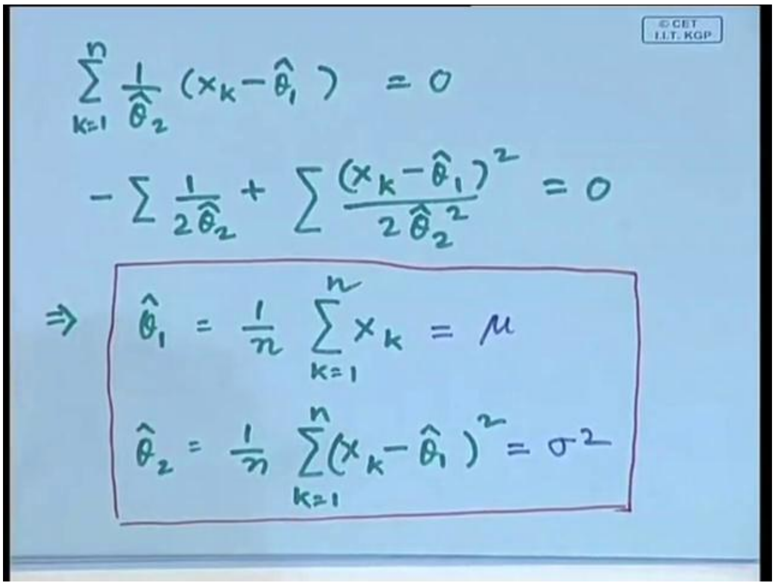
So, over here our parameter vector theta is nothing but mu and sigma square, it has two components, for simplicity I am concentrating the univariate case, for multivariate case the number of operations will be more, that is all, but the procedure will remain the same. So, here this parameter vector theta has got two components, one is mu and other one is sigma square so assume that the first component theta1 is mu and the second component theta2 is sigma square.

Now over here, the same ln p xk given theta is nothing but half of ln 2 pi theta2, theta2 is sigma square minus 1 upon 2 theta2 in to, xk minus theta1 square. Because, it is exponential of minus x minus mu square upon 2 sigma square so it simply becomes 1 upon 2 theta2 into xk minus theta1 square, in the logarithmic form. So, again if I take the derivative of this with respect to theta so grad theta of the logarithimic form l. So, it is nothing but grad theta, I am not yet come to l, because l contains the summation term.

So, let me put it as grad theta in to p ln, p xk given theta so this is nothing but I have two components, one is derivative of this with respect to theta1, other one is derivative of this with respect to theta2. So, when I take the derivative of this with respect to theta1, you find that, this, the first component is independent of theta1, this will be equal to 0. So, what I have is the derivative of this, with respect to theta1 and when I take the derivative of this with respect to theta1, it simply becomes 1 upon theta2 in to xk minus theta1.

And when I take the derivative of this, with respect to theta2, you find that the derivatives corresponding to both the terms will exist because in both the places I have theta1. And that will simply become, 1 upon 2 theta 2 plus xk minus theta1 square upon 2 theta2 square. So, I took take the summation of this and summation of this over, all the samples that is where k varying from 0 to n and equate that to 0, to give me the maximum likelihood estimate.

(Refer Slide Time: 41:58)



$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0$$

$$-\sum \frac{1}{2\hat{\theta}_2} + \sum \frac{(x_k - \hat{\theta}_1)^2}{2\hat{\theta}_2^2} = 0$$

$$\Rightarrow \boxed{\begin{array}{l} \hat{\theta}_1 = \frac{1}{n} \sum_{k=1}^{n} x_k = \mu \\ \hat{\theta}_2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{\theta}_1)^2 = \sigma^2 \end{array}}$$

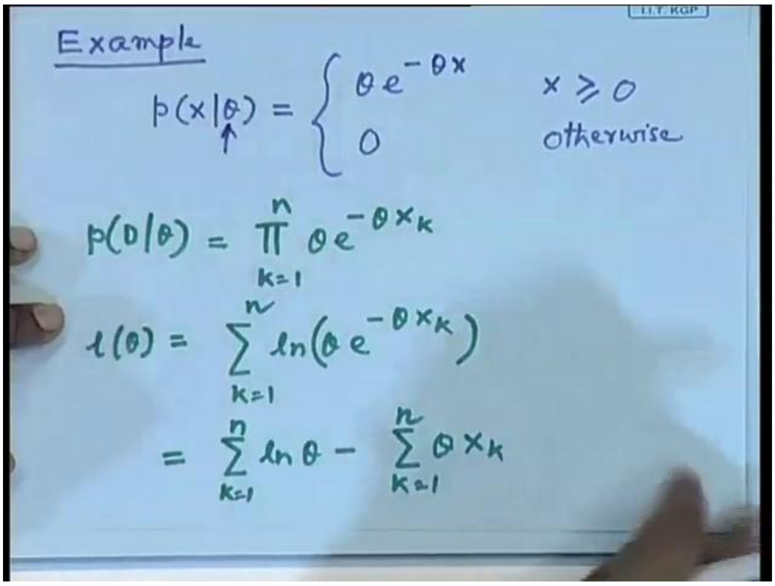
So, what I have is summation of 1 upon theta2 hat in to xk minus theta1 hat, k varying from 1 to n that should be is equal to 0, one equation. And the second equation will be minus summation of 1 upon 2 theta2 hat plus, summation of xk minus theta1 hat square upon 2 theta2 hat square, this should be equal to 0. So, I have to solve these two simultaneous equations and when you solve these two simultaneous equations, you will find that you will get answer of the form theta1 hat is nothing but 1 upon n summation of xk, k varying from 1 to n. And theta 2 hat will be nothing but 1 upon n summation of xk minus theta1 hat sqaure , k varies from 1 to n. So, simply by solving these two equations I will get these two solutions.

Now you find that, this is nothing but mu and this is nothing but sigma square. So, that validates that when I follow this maximum likelihood estimate then, I really get the parameter vectors, which truly represent the set of samples, that I have then of course, as I said that regarding the accuracy of this parameter vectors that I have, I have to depend upon the number of samples. The more the number of samples I have, more accurate our estimation will be, if I have less number of samples our estimation will be less accurate.

So, as far accuracy I have to depend upon the number of samples, I have to have more number of samples, but in most practical cases I do not get sufficient number of samples. So, as we cannot get sufficient number of samples our parameter estimation may not be sufficient, may not be accurate not only that, what kind of distribution it really follows, that is also difficult to predict.

I mean whether, it follows Gaussian distribution or it follows exponential distribution, what kind of distribution it follows that is also sometimes difficult to predict because we have insufficient number of samples. So, in such cases instead of trying for parametric probability distribution, we have a method called a non parametric estimation. We will discuss that later on, but before that let me just take another example to illustrate how this parameter estimation can be done for arbitrary probability estimation, probability density function.

(Refer Slide Time: 45:41)



Example

$$p(x|\underset{\uparrow}{\theta}) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$p(D/\theta) = \prod_{k=1}^{n} \theta e^{-\theta x_k}$$

$$l(\theta) = \sum_{k=1}^{n} \ln(\theta e^{-\theta x_k})$$

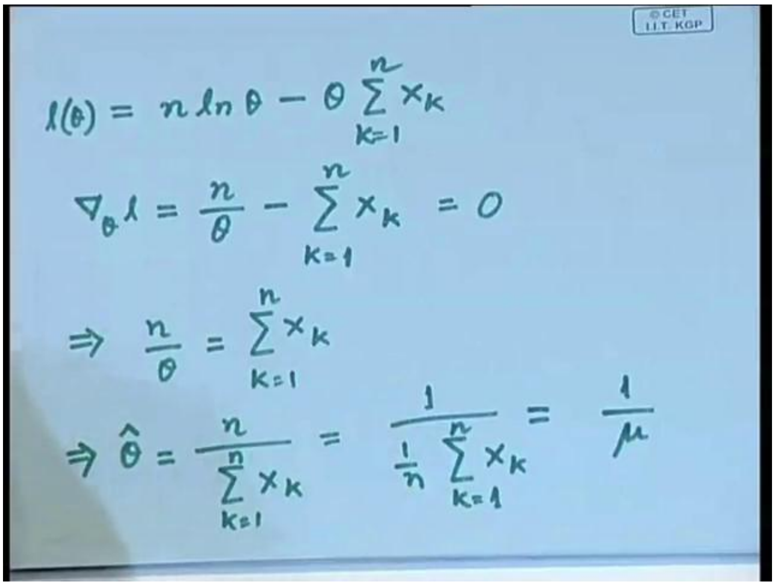
$$= \sum_{k=1}^{n} \ln \theta - \sum_{k=1}^{n} \theta x_k$$

Suppose, we have a probability density which is given by so I will take an example. So, I have p of x given theta, which is given by theta e to the power minus theta x, for x greater than or equal to 0 and this is equal to 0 otherwise. And suppose again we have k number of samples so our problem is to have the best estimate, of the maximum likelihood estimate of this parameter theta.

So, I follow the same procedure that is, I found out the likelihood value p D given theta, which is nothing but product of theta e to the power minus theta xk. Where k varies from 1 to n, once I have this likelihood, I found out what is the log likelihood l theta, which is nothing but logarithm of this. So, logarithm of this will simply be, logarithm of this term so the product term will be converted to summation term.

So, what I will have is sum of ln theta e to the power minus theta xk ,where k varies from 1 to n, this will simply be sum of ln theta because here again a product term that will be converted into two summation terms so sum of ln theta where k varies from 1 to n. And as before, it is ln theta that will be added n number of times because it is a constant term that I am adding n number of times. So, this minus sum of theta xk, where k varies from 1 to n so this I can rewrite in the form. So, here you find that it is theta in to xk, sum is taken over 1 to n so I can take out theta outside summation. So, it simply becomes sum of xk theta in to sum of xk,k varying from 1 to n.

(Refer Slide Time: 48:39)



$$l(\theta) = n \ln \theta - \theta \sum_{k=1}^{n} x_k$$

$$\nabla_\theta l = \frac{n}{\theta} - \sum_{k=1}^{n} x_k = 0$$

$$\Rightarrow \frac{n}{\theta} = \sum_{k=1}^{n} x_k$$

$$\Rightarrow \hat{\theta} = \frac{n}{\sum_{k=1}^{n} x_k} = \frac{1}{\frac{1}{n} \sum_{k=1}^{n} x_k} = \frac{1}{\mu}$$

So, it will simply be n in to ln theta minus theta in to sum of xk ,where k varies from 1 to n so that is my log likelihood l theta. Next what I have to do is, I have to take the gradient of this. So, I have to find out the gradient of l and equate that to 0. So, if I take the gradient of l what do I get, gradient of this term with respect to theta, is nothing but n by theta because it is log of theta if I take it, if I differentiate this with respect to theta, I get 1 upon theta. So, this is simply n by theta minus, if I differentiate this term with respect to theta, this theta goes away so I simply have summation of xk. So, it becomes sum of xk, where k varies from 1 to n and I have to equate that to 0.

So, when equate this to 0, you simply get n upon theta is equal to xk, k varying from 1 to n. So, this simply gives you the maximum likelihood estimate of theta,which is theta hat is, nothing but n by sum of xk, k varying from 1 to n, which is nothing but 1 upon, 1 by n sum of xk, k varying from 1 to n, which is nothing but 1 upon mu, that is the arithmetic mean. So, this is what my maximum likelihood estimate, if the probability density function is of this form.

So, what you have seen today is that, if the probability density function has a known parametric form so far our assumption is that, the pdf has a parametric form or a unknown parametric form. That means, I know that what are the parameters, that is what are the parameters which represent that particular probability density function. Now given that, another set of samples that is in case of supervised learning, another set of samples which are drawn according to that probability law, described by that set of parameter vectors.

So, using the information available in this set of samples, how I can best estimate the parameter vector and for doing that, what we have done is, we have used simply our differential calculus, that we have, that you have learned during your school level. Take the gradient equate the gradient equal to 0 and the value that you get is, what is called the maximum likelihood estimate, that is, this is the set of paramters which maximally supports the set of samples, which are drawn independently according to the given probability law.

And of course, as I said that the accuracy of this estiamtion will depend upon the number of samples that you have. Naturally, if I just have 5 samples I cannot say that these 5 samples will follow Gaussian distribution or normal distribution of a given mean vector and given variance. But if I have say, 5000 samples I can say that these 5000 samples, these 5000 samples, I can accurately estimate. Of course, within some error, I can accurately estimate the value of mean and the value of variance. So, the accuracy always depends upon the number of samples that I have, more the number of samples more accurate my estimation will be and the less the number of samples, less accurate my estimation will be.

In some cases, the sample size is so small that I cannot go for parameter estimation at all or even I do not know what is the parametric form, I cannot even predict what is the parametric form. So, in such cases what we have to go for is, the non parametric method I do not use any parametric probability density function, but given the set of samples you estimate what is the probability, that a sample has some values xy. Given, a distribution of a set of ((Refer time: 53:53)). So, I will stop here today, more in the next class.

Thank you.