Indian Institute of Technology Madras
Presents

NPTEL
National Programme on Technology Enhanced Learning

Pattern Recognition

Module 06

Lecture 04

Probability Density Estimation

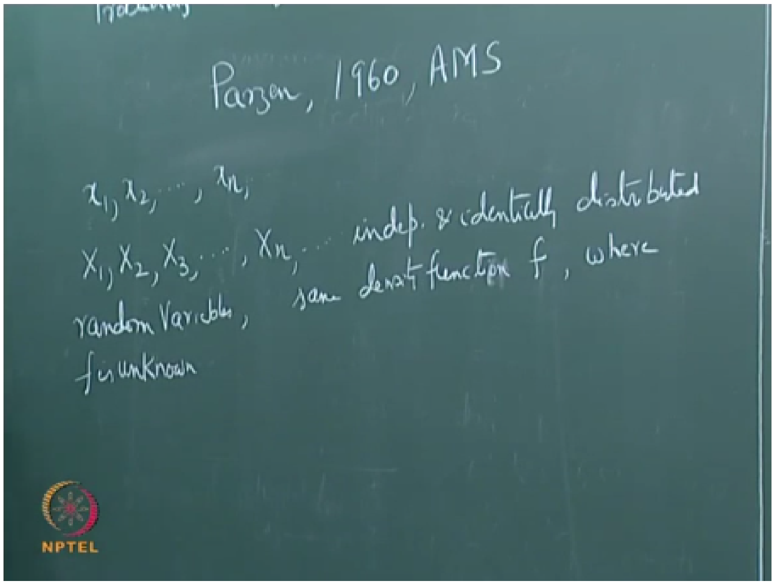
Prof. C. A. Murthy
Machine Intelligent Unit,
India Statistical Institute, Kolkata

Today we shall be discussing about probability density estimation you have already seen in these lectures that whenever we discussed classification classifiers many times we assumed the form of the density function either it is normal density or some other density function we assume that the density function is so and so in reality how do you know what density function the dataset follows this is the problem which I am going to tackle it today there are various approaches for finding densities given a data set one of the approaches which we generally learn in 2nd and 3rd year of any statistics course is fitting a distribution given a data set.

You assume that the data set follows let us just say binomial distribution and then estimate the parameters of the binomial distribution and check whether it satisfies certain tests like the fit is good or not a chi-square test or some other such test and the same is true for whether you want to fit the for a Poisson distribution or a normal distribution or any other distribution, so basically this approach deals with you assume a functional form of the distribution and estimate the parameters of the distribution but it is not necessarily true that every data set follows a standard distribution.

So the data set follows something that is not one of the usual distributions then how do you actually get the distribution in those cases we cannot assume the functional form of the density function, so the question is how does one estimate the functional form of the density function automatically from the data set that is the basic question that needs to be answered as you can see it is a fundamental problem and it is a fundamental question even for statisticians.

(Refer Slide Time: 03:01)



Parzen, 1960, AMS

$x_1, x_2, \ldots, x_n,$

$X_1, X_2, X_3, \ldots, X_n, \ldots$ indep. & identically distributed

random variables, same density function f, where

f is unknown

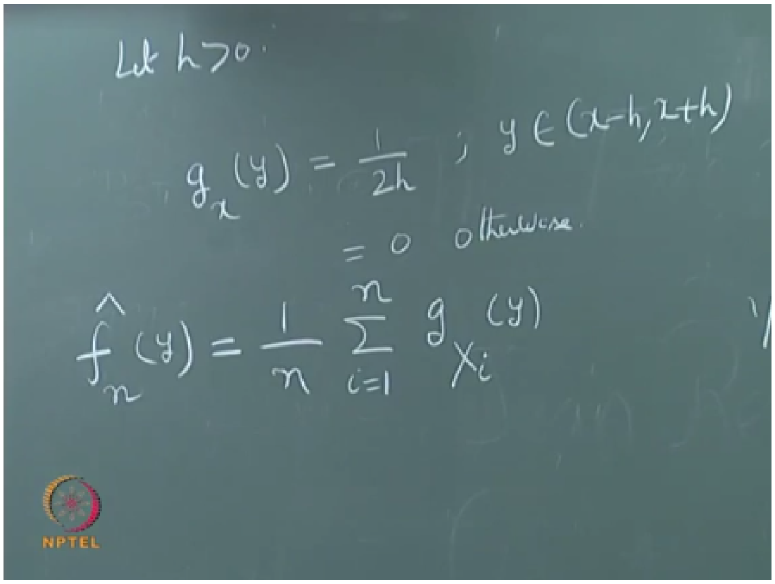
So one of the first works on this was done by someone named parson in the year 1960 and the paper was published in AMS annals of mathematical statistics this is usually known as parson density estimation this is usually known as parson density estimation now let me just explain this thing to you suppose you have any observations these are drawn randomly from an unknown distribution these are drawn randomly from an unknown distribution.

Our, I should say independent and identically distributed these are the set of observations basically what you are the set of that one always likes to write is X1, X2, X3… Xn are independent and identically distributed random variables independent and identically distributed random variables I have already explained to you the meaning of independent and identically distributed so I would not go I would not explain it once again these are independent identically distributed random variables.

So they have the same density function f same density function f density what I mean is probability density what I mean is probability density they have the same density function probability density function f where f is unknown where f is unknown, so this is the basic problem you have independent identically distributed random variables they have since they have the same distribution that means they have the same density function f but then f is not known.

So another way of saying it is that you have a you have a data set which is drawn independent and identically which is drawn independently and naturally from the same distribution and the data set is represented by X1, X2, Xn and if you have more than you have more points so this is the way usually it is formulated now the question is how does one find f the question how does one find f now let us see I will do it in the following way.

(Refer Slide Time: 06:42)



Let h > 0.

$$g_x(y) = \frac{1}{2h} \; ; \; y \in (x-h, x+h)$$

$$= 0 \quad \text{otherwise.}$$

$$\hat{f}_n(y) = \frac{1}{n}\sum_{i=1}^{n} g_{X_i}(y)$$

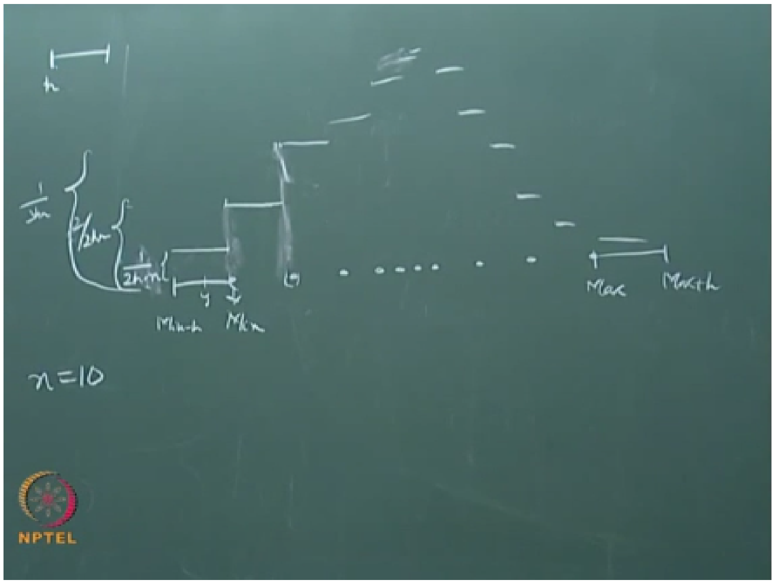
Let me first define g we shall take a positive quantity h let h be let h > 0 this is the positive quantity how to take hi will come to it later now for every x belonging to R you can define this function g (x) g(y) which is dependent on x this function is taking the value 1 / 2h if y lies between x –h + x+ h and if y does not belong to this interval it takes 0 it takes the value 0 now what is the meaning of this meaning is suppose this is x this is x- h and this is x+ h the height is the height is the same height and this is 1/ 2h this height is 1/ 2h.

Now note that this length is 2h and the height is /2h so the area under this curve is 1/2h this height is 1/2h so area under this curve is 1 so after we define g now we define estimate of the function estimate of the function f using end points and since I am calling it as an estimate let me put the sign hat estimate of the function at the value y is equal to because it is an estimate I am using the word ^ I am using the symbol ^ it is based on n random variables I am using the symbol n I am finding the value of the estimate at the point y.

So it is fn ^ y this is one by small n summation I is equal to 1 p.m. I have already defined gxi(y) y = 1 let me first explain some of the mathematical properties now I will explain intuitively what is happening hereafter that note that ∫ over gx = 1 for all x right because in the interval x- 2h, x+ 2h the value is 1 and other tests I mean - ∞ - ∞ x -2h x - 2h x + 2 h x + 2 h to ∞ you can divide the whole range into those 3 parts and the middle portion the value the ∫ value is1 and there are other portions since the function is taking the value 0 it is 0.

So totally it is 1 and for all X this is true now what about the integration of this is the average of this each ∫ here is 1 so what is the meaning of that what is the prop what are the properties of densities a function f is said to be a density function when f ≥ 0 and ∫ is 1 and the ∫ here is 1 okay it does not matter what exercise you're going to take what exercise you take the ∫ is 1 actually a density function.

(Refer Slide Time: 12:21)



Now let me tell you what exactly is happening for that I will what I will do is say this is our data set g is actually window actually we know and the ∫ is one constraint I will be using the words windows and other things slightly later say this is our given data set how many points are here 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 points so n = 10 there are 10 points here now we are supposed to estimate density function.

Now where do you think the density would be high given these 10 points can you say given these 10 points where do you think the density may be really high somewhere here right and then the density is decreasing okay now suppose h suppose this is h this much length is say h okay suppose h is this much now so what is happening from this point say this is minimum this is a minimum - h this point right anything below this anything less than this the density value will be 0 according to the estimate according to the estimate because x anything less than x - h g 0 and we are taking the minimum - h is this.

So the value of the density would be 0 estimate of the density would be 0 now this side this is your maximum and this is + h greater than this the estimate of the density would be 0 rather than this the estimate of the density would be 0 now let us see what is going to happen when you take a y here suppose you take a y here now what will happen with respect to this point you will get the value 1/2h but then with respect to this you would not get any 1/2h with respect to this you would not get any 1/2h so for all these points let us say this height is 1/2h x n this height is 1/2h x n are you understanding I will repeat it if you take anyway here then it is in the 2h interval of this point but it is not in the 2h interval of this point or this point or any other point.

So it is in the 2h interval of this point so the value is 1/2h and there is an L here so 1/2h x n for all the points here okay now let us look at this point let us say for this point that h this h is coming let us just say it is coming say somewhere here it is coming say somewhere say here then from this point onwards the height will go to1/2h n + 1/2h n it will be increased from this point onwards it will be increased okay.

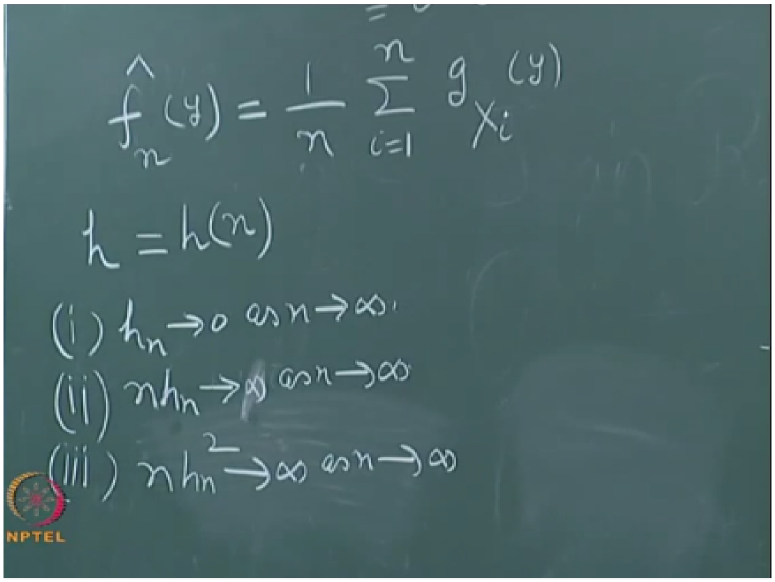
And it will be say this is 2/ 2hn this height is 2/2hn okay maybe the influence of this point it will come up to this but then at the influence of this point in the interval suppose say it is starting say some where here then from this point onwards the height will be 1/3hn because this point is influenced by this it is in that h interval of this is in the h interval of this is in the h interval of this so this will be 1/3hm and somewhere when you come to this it will be more okay actually it should one should not write these lines one should write only things like this.

And somewhere here you are going to have some maximum maybe so this is basically a step function that you are saying this is basically a step function that you are seeing.

And it basically goes up like this for some datasets maybe depending on the datasets the shapes will change this is fn ^ y this is fn ^ y right now somehow this fn ^y given the data set it is somehow whatever our feeling is there that where it the value it should be the density should be maximum it is actually giving that now the question is incredibly yes we are getting a step function we are not getting a continuous function it is actually discontinuous at many points as you can see another question is does it have any theoretical properties.

Now pardon what he did was he had shown that under some conditions this density estimate it will go to the actual density estimate actual density whatever may be the functional form under some conditions now let us see what those conditions are the first condition is regarding this h the first condition is regarding this h needs to be a function of n.

(Refer Slide Time: 20:54)



$$\hat{f}_n(y) = \frac{1}{n}\sum_{i=1}^{n} g_{X_i}(y)$$

$h = h(n)$

(i) $h_n \to 0$ as $n \to \infty$.

(ii) $nh_n \to \infty$ as $n \to \infty$

(iii) $nh_n^2 \to \infty$ as $n \to \infty$

h needs to be a function of n so let me just write it as h as hn is the number of points. Now as the number of points n goes to∞ hm should go towards 0 that is the first condition this is the first condition as n goes towards ∞ hn goes towards 0 why the reason is this suppose hn is not going to 0 hn is going to let us just say some finite quantum quantity greater than 0 then what happens is that suppose the original density function say its values are it takes values from say 0 to 1 okay f o(x) where x takes values in the interval 0 to 1 now since this f is and this h is not going towards 0 it goes to say some quantity $\Delta > 0$.

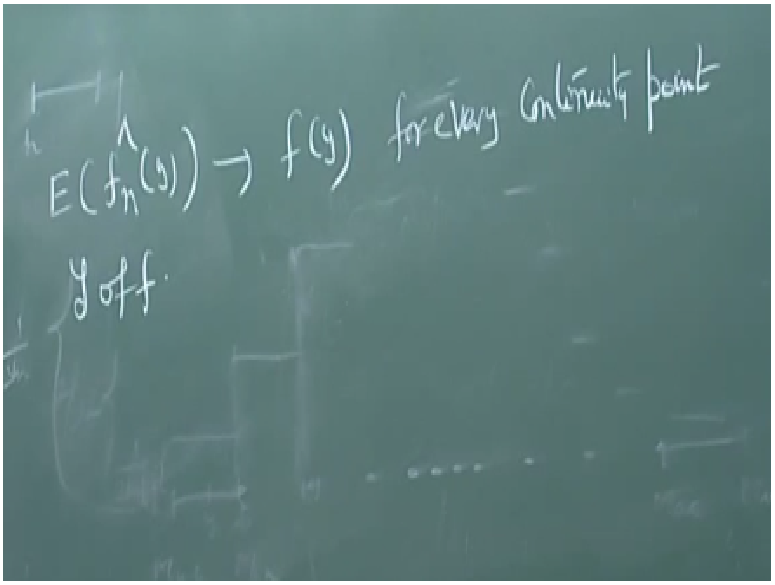
Then this fn^ y it will be - Δ to 1+ Δ because always you are taking this interval you are taking an open set around each one of the points that is given to you are taking a neighborhood around it and the neighborhood size you are decreasing but you are not decreasing it to 0 you are decreasing it to some quantity greater than 0 so whatever points lying in those interval in those intervals they are going to have positive value right then in that way it will not go to the original density so if you want to get original density hn should go towards 0 as n goes towards ∞ there is a second condition.

Suppose hn goes towards 0 very fast suppose hn goes towards 0 very fast what is the meaning of that it goes to 0 so fast that around these points we are considering disks or open sets ultimately it will go to only those points it does not include any other points then in that case you will only get at each point some height so that should not happen you want to include some points there some more points other than the given points so it should go to 0 it should go to 0 slowly so the second condition is it should go to 0 but it is going to 0 slowly that is the end time section sorry.

It goes to ∞ as n goes to ∞ n times hn it goes to ∞ as n goes to ∞ so what is the meaning of this I mean there are many sequences actually which satisfy these two conditions I will tell you a few such sequences one sequence is let us look at this1 / $n^{1/3}$ it goes to 0 as n goes to ∞ but the n time section goes to ∞ as n goes to ∞ 1/ log n you are computer scientists we like to okay look at log n is 1 / log n goes to 0 but n times 1 / log n goes to ∞ right so there are actually many sequences which satisfies these two conditions then these are two very small conditions I mean there is nothing really intuitively we understand how these conditions are coming now by using these conditions in fact he has used 1 another condition.

The other condition is this then he had shown that this density estimate is asymptotically unbiased and consistent what is the meaning of asymptotically unbiased asymptotically unbiased means first we find the expectation of this expectation is a function of naturally and this expectation it goes to the actual expectation value I am an actual value of the function I will write it clearly I will write this statement clearly.

(Refer Slide Time: 27:29)

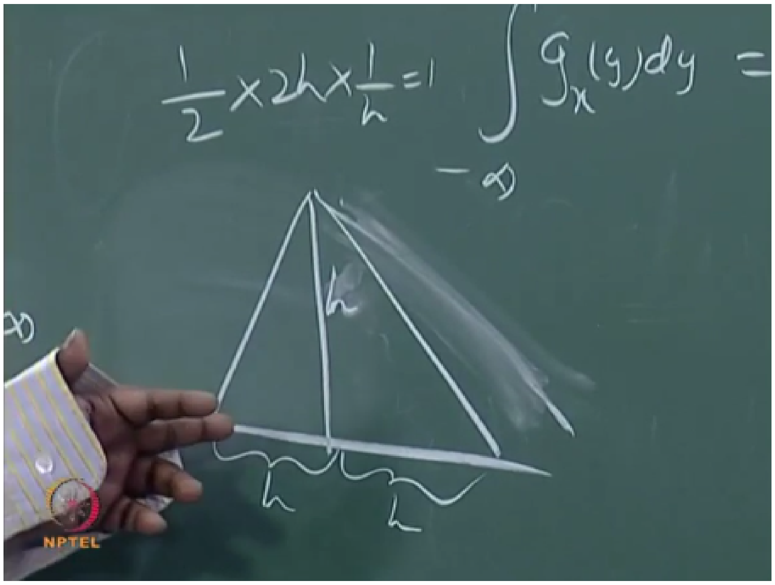


Expected value of f^n why it goes to f of y for every continuity point y of f and the squared error that goes to 0 as n goes to ∞ okay. The squared error that goes to 0 as n goes to ∞ for that he I think he used this one note that if this is satisfied and surely this is also satisfied is it correct when this is satisfied this is satisfied because HN is going towards 0 but this is satisfied it does not mean that this is satisfied so if we take a sequence hn goes to 0 n times action square goes to ∞ then naturally n times hn also will go towards ∞.

And he showed that this estimate is asymptotically unbiased that is this and consistent means basically the error goes to 0 the error goes to 0 so this is what we had shown now I will come to the wordings used by professor das use the word window what exactly are we doing we are taking a window of length h around every point and in that window we are considering some value and otherwise we are taking here 0 and that window actually we are moving on the data set we are taking a window of length h around each point and that actually we are going or the entire data set so this is also known as Parsons window techniques for estimating densities.

So you also see Parsons windows okay you will also see the word Parsons windows in the literature now let us look at this function this is uniform distribution as you know it you might ask me a question suppose I do not take uniform I take some other distribution then what is going to happen yes you need not have to take uniform distribution you can take.

(Refer Slide Time: 30:16)


$\frac{1}{2} \times 2h \times \frac{1}{h} = 1$
$\int_{-\infty} g_x(y)dy =$
h
h
h
NPTEL

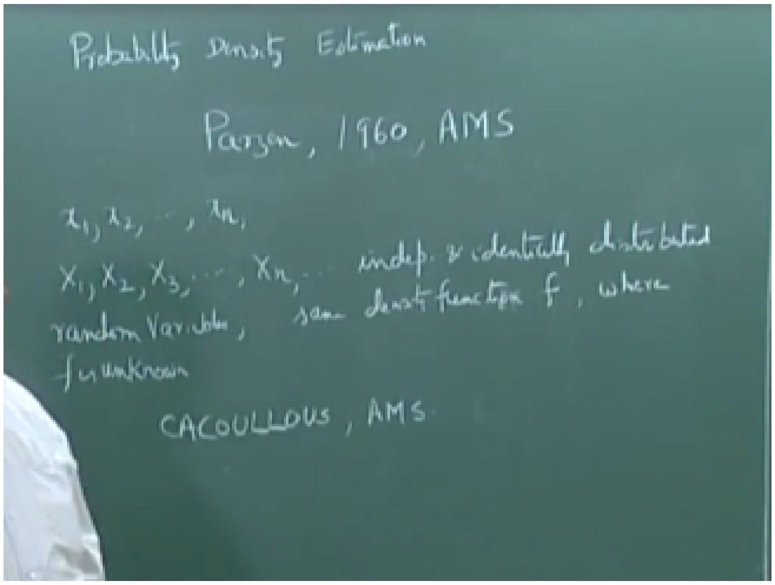
Say for example a triangular distribution a triangle a symmetric one in fact you can take it say this is h and this is h and the height will be 2h 1/2 times base into height that must be equal to 1 so this must be h 1/2 times base is 2h so height must be 1/h that is 1 so this is h so you can take this then also you will get a similar result in fact he had shown it for a class of functions this 1 / h sorry this is 1/h in fact he had shown it for a class of functions and he called them as kernel functions one of the first places where you see the word kernel is by parson.,

So this is also called kernel density estimation this is also called kernel density estimation, so he had given properties of kernels there you will see it in his paper you will see it in many other books has book many other pattern-recognition book you will see the properties of the kernels so whichever kernel function you take and if you take a chance satisfying these properties then all of them will give you that result that is asymptotically unbiased and consistent this is a very strong statement that is why pardons density estimation.

This was done in 1960 even now we are talking about it because he had used wide varieties of functions and the conditions are simple conditions are simple conditions he had used here wide varieties of functions and for this all those wide varieties of functions he could prove the same research naturally normally is also included here so if you take normal distribution for g then what are you going to write then the h is going to be for the variance okay h will be for the variance so normal distribution with some mean the mean is x and the variance is H that is what you are going to write here.

Are you understanding me normal distribution with the mean X and the variance h that is what you will write here so now this has been proved for univariate case what is the meaning of univariate case note that we are assuming random variables not random vectors so instead of random variables suppose you have random vectors that is you have observations in let us just say two dimensional or three dimensional are in general in small m dimensional space and you would like to estimate densities.

(Refer Slide Time: 34:26)



Probability Density Estimation

Parzen, 1960, AMS

$x_1, x_2, \ldots, x_n,$

$X_1, X_2, X_3, \ldots, X_n, \ldots$ indep. & identically distributed

random Variables, same density function f, where

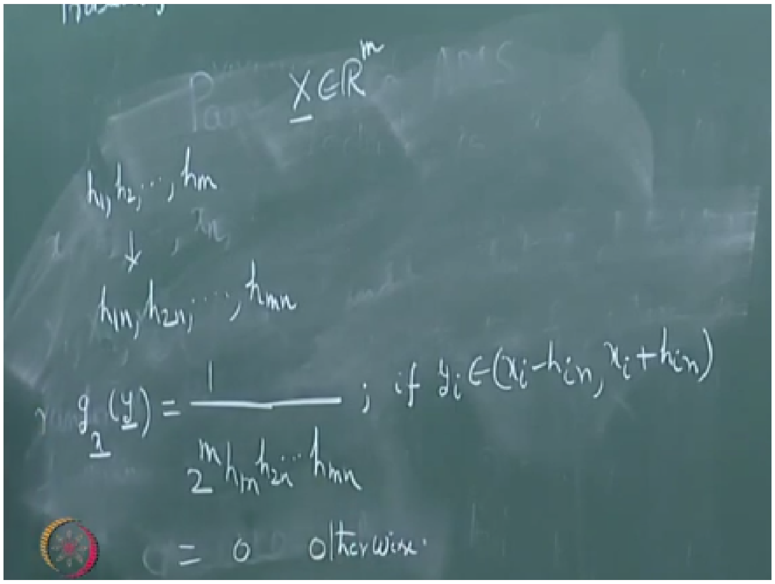
f is unknown

CACOULLOUS, AMS

Then this was generalized to multiple dimensions and the generalization was done by someone named calculus the spilling could be slightly wrong calculus and this was again a paper in annals of mathematical statistics is again a paper in mathematics of mathematical statistics either 1962 or 63 and the generalization is fairly simple generalization fairly simple generalization virtually writing the same things many times for example for two dimensions here instead of 1h you might be having 2, h1 and h2.

So if you are in the m dimensional space you might have h1, h2, hm m such values each one is dependent on n you have M such values each one is dependent on n so each of them you have to assume that it goes to 0 and all the things corresponding conditions then actually if you look at it I will just write down this denominator portion in that the denominator portion for m dimensional case can you tell me what it would actually raise this portion.

(Refer Slide Time: 36:18)



$\underline{X} \in \mathbb{R}^m$

$h_1, h_2, \ldots, h_m$

$\downarrow$

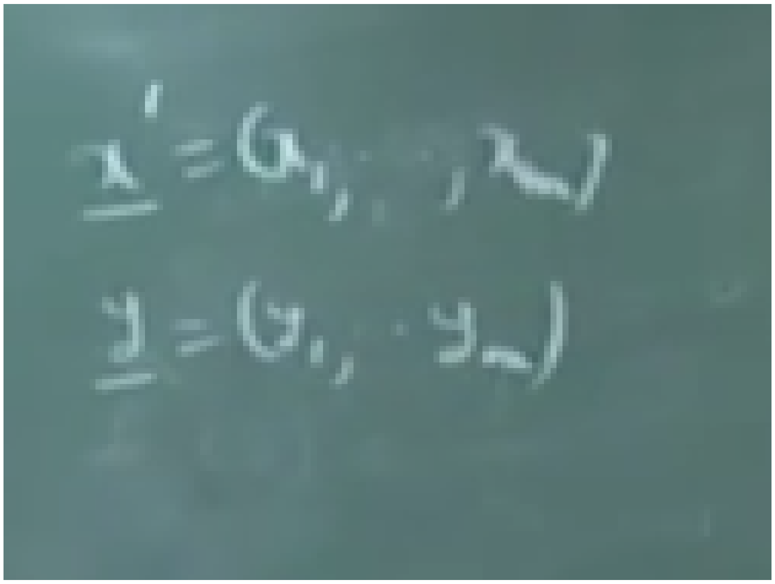
$h_{1n}, h_{2n}, \ldots, h_{mn}$

$$g_n(\underline{y}) = \frac{1}{2^m h_{1n} h_{2n} \cdots h_{mn}} \; ; \text{ if } y_i \in (x_i - h_{in}, x_i + h_{in})$$

$= 0$ otherwise.

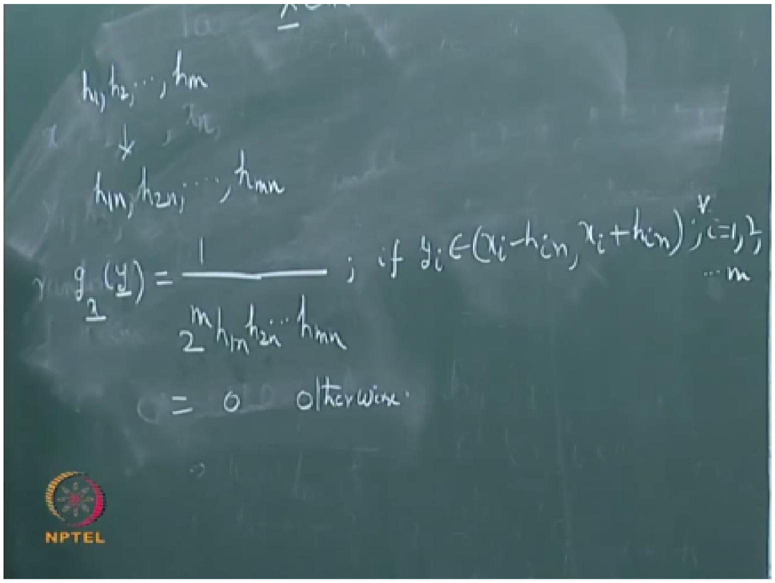
So you have h1, h2, hm or random variables they take values in m dimensional space you have h1, h2, hm each one of them is again dependent on n so it is basically you have h1 and h2n and hmm okay and for each one of them if you have something like that then you have here and x okay and here it is y this will be 1/ 2^m h1n,h2n, hmn if let me write yi belongs to xi- hin 2x i + hin 0 otherwise okay. What are these xs and ys.

(Refer Slide Time: 38:03)


$\underline{x}' = (x_1, \ldots, x_m)$
$\underline{y} = (y_1, \ldots, y_m)$

Your vector x is x1, to xm vector y is y1 to ym I = 1, 2 up to you now look at this X instead of just a single x here you have a vector X instead of single Y you have a vector yg xy is the variable so this vector y is y1 to ym so look at the ith component yi this yi if it is belonging to xi -han 2xi + hain for all I then this is the one otherwise it is 0 so here also you are considering a window and what is this is the volume right this is the volume this is volume of m dimensional rectangle where the sides are the first side is to h1n the second side is to 2h2n and the mh side is 2 hm this is the volume of m dimensional rectangular.

(Refer Slide Time: 38:48)



$h_1, h_2, \ldots, h_m$

$\downarrow$

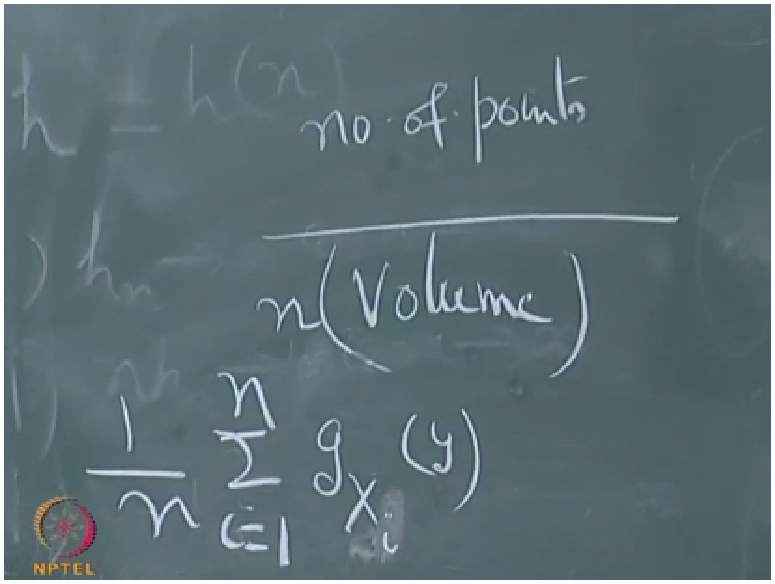
$h_{1n}, h_{2n}, \ldots, h_{mn}$

$$g_{\underline{1}}(\underline{y}) = \frac{1}{2^m h_{1n} h_{2n} \cdots h_{mn}} \; ; \text{ if } y_i \in (x_i - h_{in}, x_i + h_{in}) ; \forall i = 1, 2, \ldots m$$

$= 0$ otherwise.

And if you write down actually the estimated density function with respect to this then at each point Y what you are going to see is that how many of these capital Xi are lying in that particular volume how many of these capital Xi are lying in that particular volume just that number divided by small n into this volume that is what you are going to see I will repeat it in the denominator you will see a volume you will also see m and in the numerator the number of points that are lying in the rectangle the number of points that are lying in the rectangle around this point y it is basically some number of points.

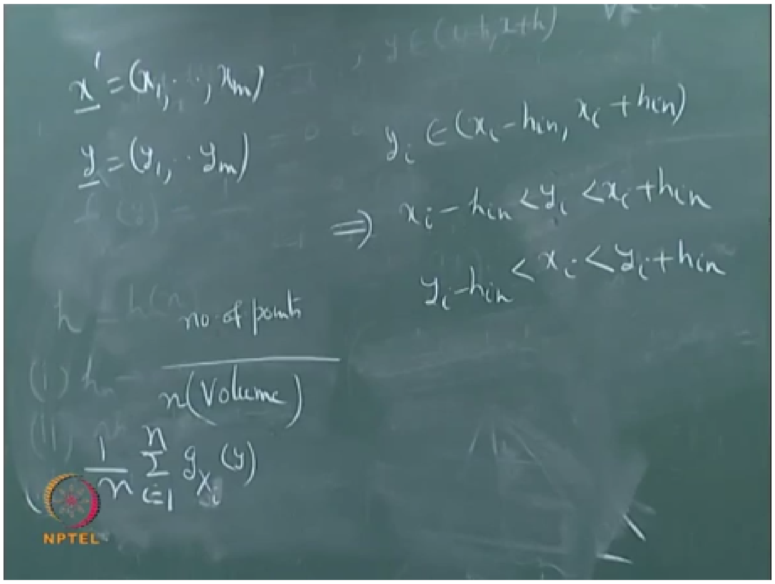
(Refer Slide Time: 40:54)



This if you assume uniform distribution now let me just ask you a simple question when did you first come across the word density probably in class 8 or 9 and what did we learn about the meaning of the word density mass by volume right and here you see volume here you see volume instead of mass here you have number of points and this is out of the small n points you have this so this looks like I mean instead of mass actually if you replace that thing by this number of points it is basically like that word density that we have used and there is a volume here around y and just arm if you okay let me this is for the single dimension.

What does this thing mean you take y you take the h interval the number of points like number of X is lying in that interval you take y you take the h interval find the number of points if there is no point then the estimated density is 0 okay and this is generalized to higher dimensions and you have volume here and this is number of points in rectangle of this volume number of points in the rectangle of this volume around y there I mean the number of point number of points lying number of Xi lying in this volume around this point y among these Xi show many points are lying in this volume around y that is it that is this number of points. The basic expression it shows around X the expression so you see I want to do it for each I okay forget about for all of i.

(Refer Slide Time: 44:34)

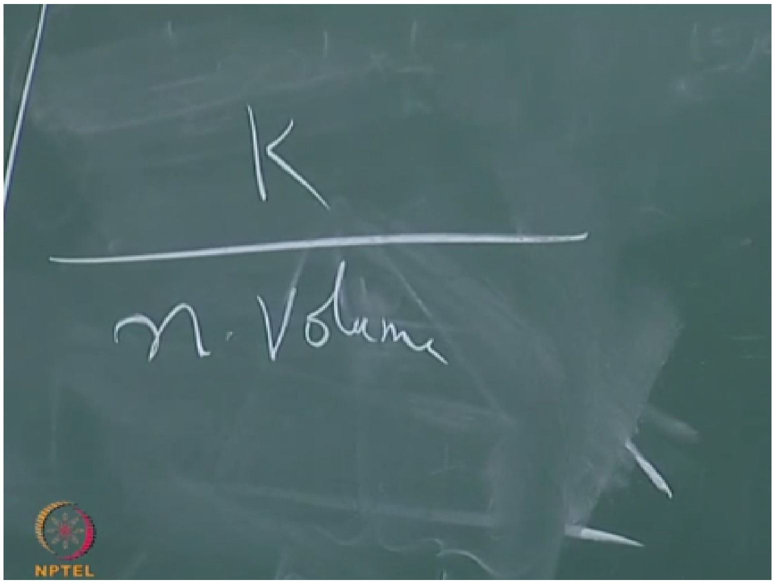

$\underline{x}' = (x_1, \cdot\cdot, x_m)$
$\underline{y} = (y_1, \cdot\, y_m)$
$y_i \in (x_i - h_{in}, x_i + h_{in})$
$\Rightarrow x_i - h_{in} < y_i < x_i + h_{in}$
$y_i - h_{in} < x_i < y_i + h_{in}$
$\frac{\text{no of points}}{n(\text{Volume})}$
$\frac{1}{n}\sum_{i=1}^{n} g_{x_i}(y)$
NPTEL

Let us just look at only 1 yi why I belongs to x i- h AI n 2x i + hIm this means what x i- h Ai n > y a less than xi + hai right right you bring this hsn on this side then x i will be less than y i + him bringing this hin here bring this hin here so either you write this are you write this the same y is your variable at which you want to get your density y is the one at which you want to get your density these Xi are the things are given to us according yes of course this is those kernel functions that is what they do you can use any kernel function he is just one of them uniform distribution is one of them you can use it by spears you can use it you can use any arbitrary shapes no problem you can use arbitrary shapes no problem.

So looking at this laughed Scott and gave his wound density estimate note that here we are actually fixing the volume we are counting the number of points and the last garden in 1966 what he did was he fixed the points he varied the volume I will repeat it he fixed the number of points and he varied the volume so how do you fix the number of points say for a particular point y you want to find it is estimated density so what you are going to do you find it is yet the nearest neighbor you find it is k^{th} nearest neighbor okay.

And so find the distance between this y and the K^{th} nearest neighbor distance between this Y and the K nearest neighbor now that distance use the distance as radius find the volume use the distance as radius and find the volume then his estimated density is actually m and then that particular volume with that radius.

(Refer Slide Time: 48:18)



And he has fixed the number of points K he has fixed the number of points K this particular density is known as density estimation procedure is known as K nearest neighbor density estimation procedure lofts garden 1966 annals of mathematical statistics loftsgaarden1966 annals of mathematical statistics now using this K nearest neighbor density estimation procedure you can derive K nearest neighbor decision rule you can derive the K nearest neighbor decision rule stop it.

End of Module 01 – Lecture 02

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

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