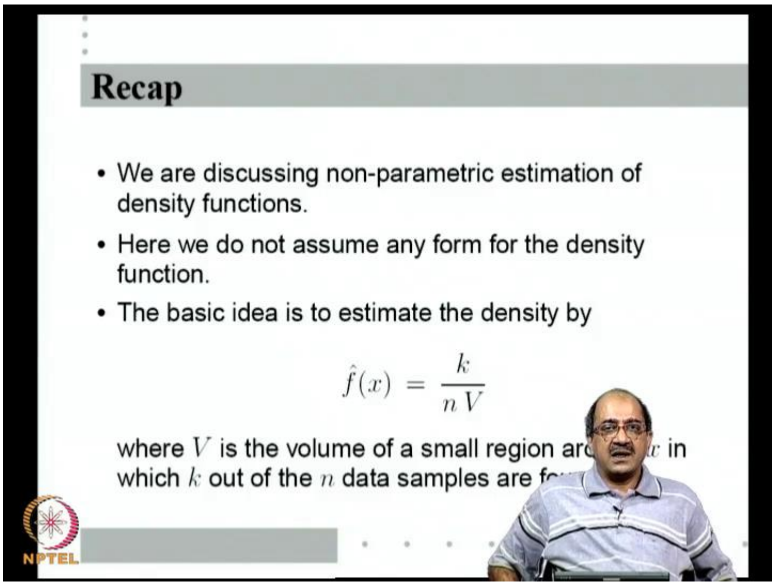
Pattern Recognition
Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 13
Nonparametric estimation, Parzen Windows, nearest neighbour methods

Hello and welcome to the next lecture in this course on pattern recognition. We have been discussing the nonparametric density estimation techniques.

(Refer Slide Time: 00:40)



Recap

- We are discussing non-parametric estimation of density functions.
- Here we do not assume any form for the density function.
- The basic idea is to estimate the density by

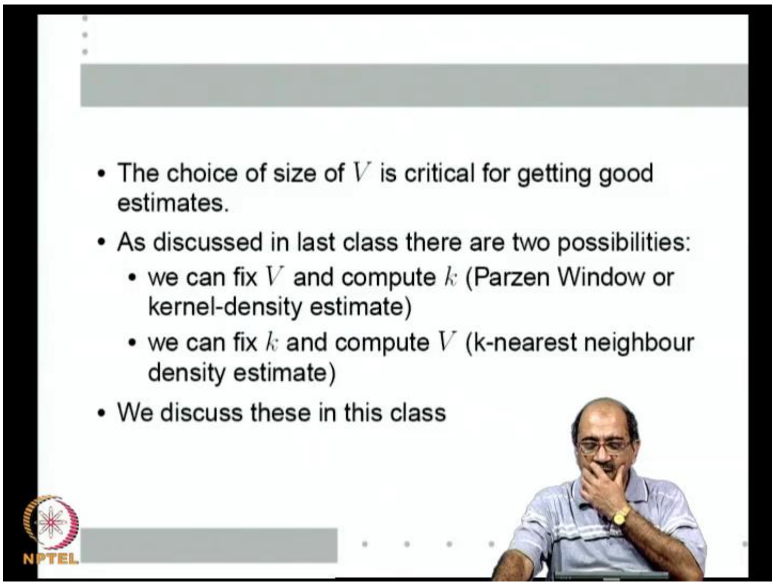
$$\hat{f}(x) = \frac{k}{n\,V}$$

where V is the volume of a small region ar x in which k out of the n data samples are f

As I said last class in the nonparametric method of density estimation, we do not assume any functional form for the density. In the parametric density estimation, we assume a particular form for the density function and then estimate the parameters. Whereas, in the nonparametric form we do not assume any form for the density function; and still we need to have density estimate.

As we discussed last class the basic idea is generalising a simple histogram for the density estimation. So the basic idea is to estimate the density value at a particular value x as represented as f hat x to be equal to k by n times V, where V is a small volume element a small element of volume V put around x in which out of the n data samples we found k data samples. As we discussed last class this is a reasonable estimate if out of k out of n samples are found in this volume essentially, if you are assuming around this small volume density is constant, then this is a good estimate with the (( )).

(Refer Slide Time: 01:47)



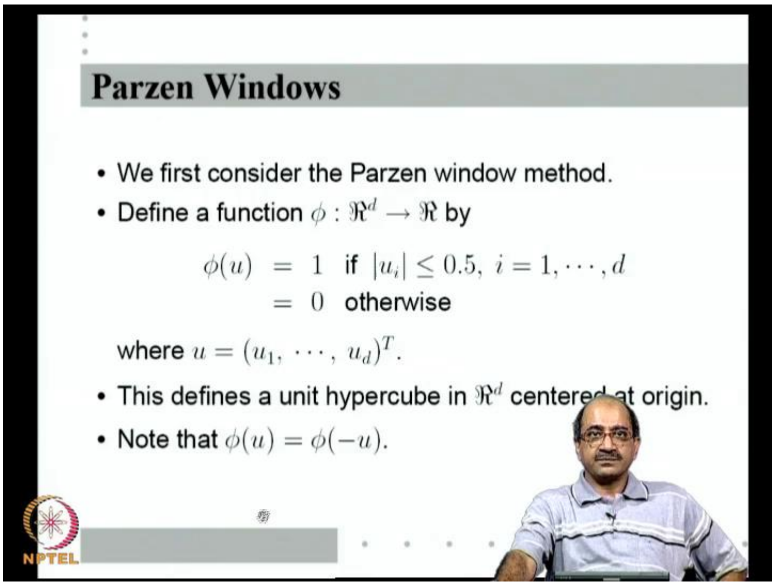
- The choice of size of V is critical for getting good estimates.
- As discussed in last class there are two possibilities:
  - we can fix V and compute k (Parzen Window or kernel-density estimate)
  - we can fix k and compute V (k-nearest neighbour density estimate)
- We discuss these in this class

So, the (( )) estimation our final estimate is k by n V, where V is the volume of the small region around x at, which we choose and k is the number of data samples that are found in that region, while n is the total number of data samples. As we already said the choice of the volume element is critical for getting good estimates, if it is too large, then we get very heavily smoothed out density estimate, which is not accurate and if it is too small not many data samples will fall there, so most of the places the density estimate may become 0.

So, the choice of V is quite critical in nonparametric density estimation, and we have discussed in last class that there are basically, two possibilities in getting this kind of density estimate by this kind I mean, f hat x is k by n V kind of estimates. The two possibilities are one is at each x we fix V, so at each x we take a region of volume V around x and then compute k the number of sample that fall in the t shell these are called kernel-density estimates or Parzen windows. And the other approach would be to fix k and compute the needed region to enclose the k samples and compute its volume.

This I call a k nearest neighbour density estimates in nonparametric estimates mostly one uses only kernel-density estimate most of the time only kernel-density estimate when we want an explicit density estimate, but k nearest neighbour density estimates were also used in a sense to justify nearest neighbour classification rule and also they are used in certain regression estimates we will consider both of them in this class.

(Refer Slide Time: 03:21)



Parzen Windows

- We first consider the Parzen window method.
- Define a function $\phi : \Re^d \rightarrow \Re$ by

$$\begin{aligned} \phi(u) &= 1 \text{ if } |u_i| \leq 0.5,\ i = 1, \cdots, d \\ &= 0 \text{ otherwise} \end{aligned}$$

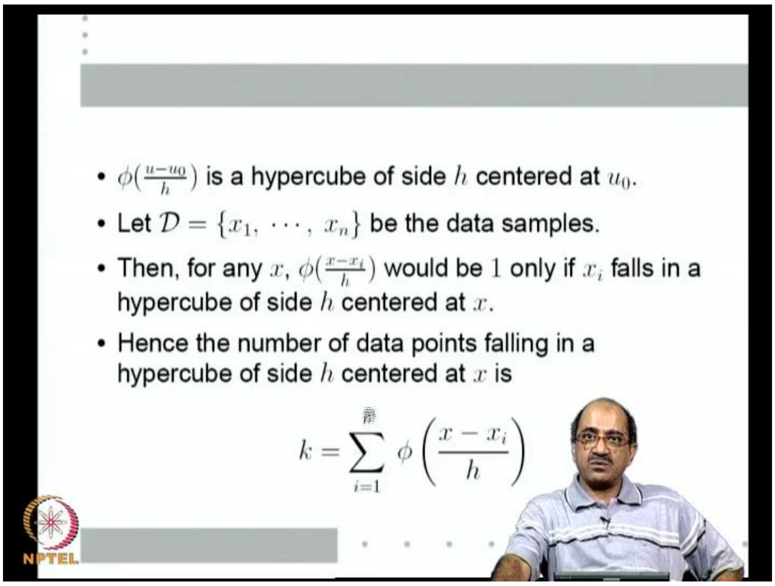
where $u = (u_1, \cdots, u_d)^T$.

- This defines a unit hypercube in $\Re^d$ centered at origin.
- Note that $\phi(u) = \phi(-u)$.

Now, let us first look at the Parzen window technique, so for this first we will first define what is known as a window function. So, let us define a function phi, which we think of as a window function it is a function from r d to r where d is the dimension of the feature vector or feature vectors are in r d, so we define a function r d to r by phi of u is 1 if u i is competent of u is its absolute value is less than 0.05 is 0 otherwise. So, u is the vector u 1 to u d as I already said we do not put any bold notation vectors.

So, from context, so this u is a vector and its components are u i and phi u is 1 only if each component is in absolute value less than 0.05. So, what does this give us? This defines a hypercube in r d centred at origin, because it is a unit hypercube side of side 1 that is why for u i between minus 0.05 to plus 0.05 if all i if for all i u i is between minus 0.05 to plus 0.05 then phi u is 1 otherwise phi u is 0.

(Refer Slide Time: 04:58)



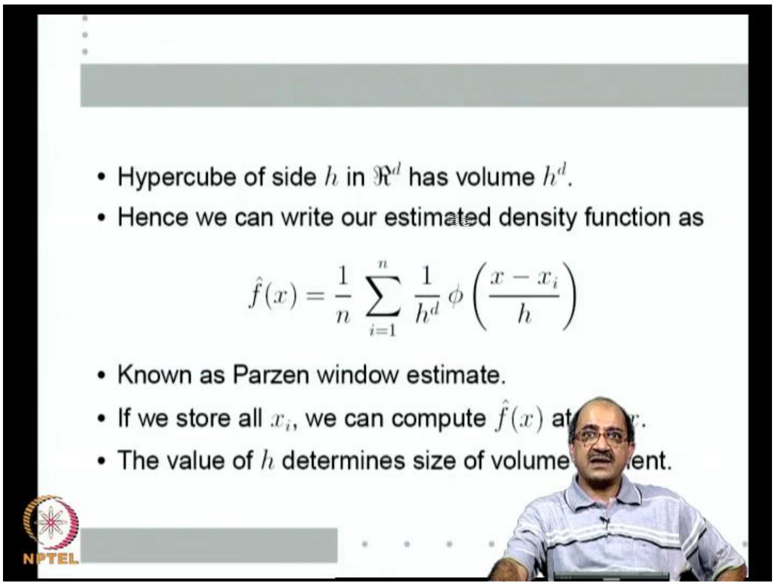
- $\phi(\frac{u-u_0}{h})$ is a hypercube of side h centered at u_0.
- Let $\mathcal{D} = \{x_1, \cdots, x_n\}$ be the data samples.
- Then, for any x, $\phi(\frac{x-x_i}{h})$ would be 1 only if x_i falls in a hypercube of side h centered at x.
- Hence the number of data points falling in a hypercube of side h centered at x is

$$k = \sum_{i=1}^{n} \phi\left(\frac{x-x_i}{h}\right)$$

So, this phi u is essentially a indicator function of the unit hypercube in r d centred at origin that is to say if the point u is inside the unit hypercube in r d centred at origin then phi will be 1 otherwise phi will be 0, okay? We also note that, because they are centred at origin and the weight is defined phi u is symmetric phi u is phi f minus u, so if I do a translation and scaling of phi then phi u minus u 0 by a translation would be a unit hyperbolic cube centred at u 0, so phi u minus u 0 by h will be hypercube of side h centred at u 0.

Now, let us say as usual x 1 to x n are the data samples then for any given x that is any point in the feature space phi of x minus x i by h would be 1 only if x i falls inside an hypercube of side h centred at x we just seen that phi is symmetric. So when you consider phi x minus x i by h i can think of it as a hypercube centred at x i or centred at x depending on, which I am fixing, so for any particular value of x phi of x minus x i by h can be thought of as a hypercube of side h centred at x and hence, it will be one only if x i is inside this hypercube. So, this function essentially allows me to ask, which are the samples are within a hypercube of side h centred at x.

(Refer Slide Time: 06:46)



- Hypercube of side h in $\Re^d$ has volume h^d.
- Hence we can write our estimated density function as

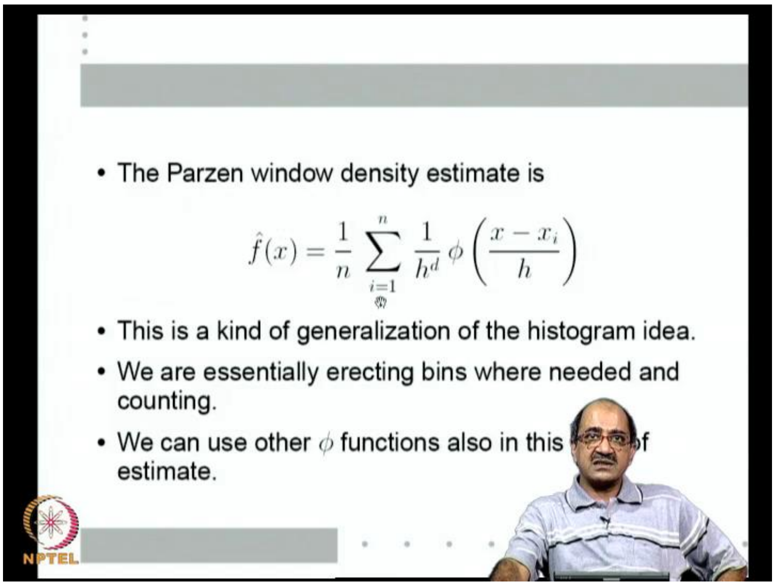
$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} \phi \left(\frac{x - x_i}{h} \right)$$

- Known as Parzen window estimate.
- If we store all x_i, we can compute $\hat{f}(x)$ at .
- The value of h determines size of volume ent.

So, if we if I take the volume element to be hypercube of side h centred at x then this is how I can count how many of my samples fall in this volume element. So, that means the number of data points falling in a hypercube of side h centred at x is given by some more i is equal to 1 to n phi of x minus x i by h, where each i phi of x minus x i by h is 1 if x i is inside this hypercube of side h centred at x, so this sum will give me the number of points in this hypercube. And in r d a hypercube of side h has volume h power d, now we know the volume element h power d we know the number of points that fall in this volume element and n is the total number at samples. So, we can write our estimate f hat x as sum over this phi x minus x i by h is k, k by n V is what i need k by n and V is h power d we will write this, because it is a more convenient form.

So, it is 1 by n summation is equal to 1 to n 1 by h d phi x minus x i by h a matter of fact this should remind you of the mixture density model that we considered couple of classes ago h it become it is a mixture density and I will come back to that later. But given this window function phi if I wanted to do my estimate as k by n V that we discussed earlier this will be the expression, so I can use this phi function to actually find the density value at any x, so this is known as the Parzen window estimate. Essentially, what it means is that? When I want density estimate what do I need density estimate for to implement the Bayes classifier, so to implement the Bayes classifier when I need see a new feature vector x at that x I have to compute the value of the class conditional densities.

(Refer Slide Time: 08:43)



- The Parzen window density estimate is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} \phi \left(\frac{x - x_i}{h} \right)$$

- This is a kind of generalization of the histogram idea.
- We are essentially erecting bins where needed and counting.
- We can use other ϕ functions also in this of estimate.

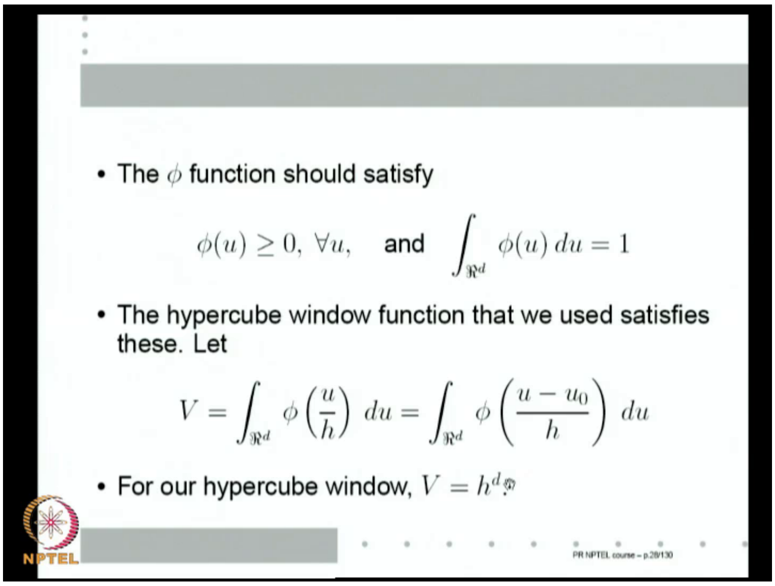
So, if I store all my samples x i given any x I can compute phi hat x using this oaky. The value of h determines the size of the volume element we said we have to take some volume of appropriate size, so that choice is now translates to choice of it, so I choose some h and then given any x I compute this that will give me the density estimate at x. So, Parzen window density estimate is given by this I hope one can see there is a kind of generalization of the histogram idea we take a bin and count a volume bin and count how many data points fall in that volume bin.

Only thing what we are doing is we are erecting bins where we need, so I need to estimate the value of density at a particular given feature vector value x, so I only need f hat x at that x value, so as and when need it I compute this by you know keeping all the x i with me and this size of the volume element is controlled by the choice of h. Essentially, the reason for writing like this is to say that the counting can be done by a function phi we have chosen of course, the unit hypercube for this function, but many other functions that are feasible.

So, I can choose many different volume elements to say and actually we can even generalise this to say that it is not actually a, a, a, a volume element with very discrete boundaries. So, we will look at if I want something like this to be a density estimate what kind of property should phi have and then by choosing any such phi we can get a similar estimate all such estimates are called kernel-density estimates and sometimes also called

Parzen window estimates. Originally only the hypercube estimate is known as Parzen windows, but generally any permissible phi when you use it is generally called as a Parzen window estimate or a kernel-density estimate.

(Refer Slide Time: 10:26)



- The ϕ function should satisfy

$$\phi(u) \geq 0, \ \forall u, \quad \text{and} \quad \int_{\Re^d} \phi(u)\, du = 1$$

- The hypercube window function that we used satisfies these. Let

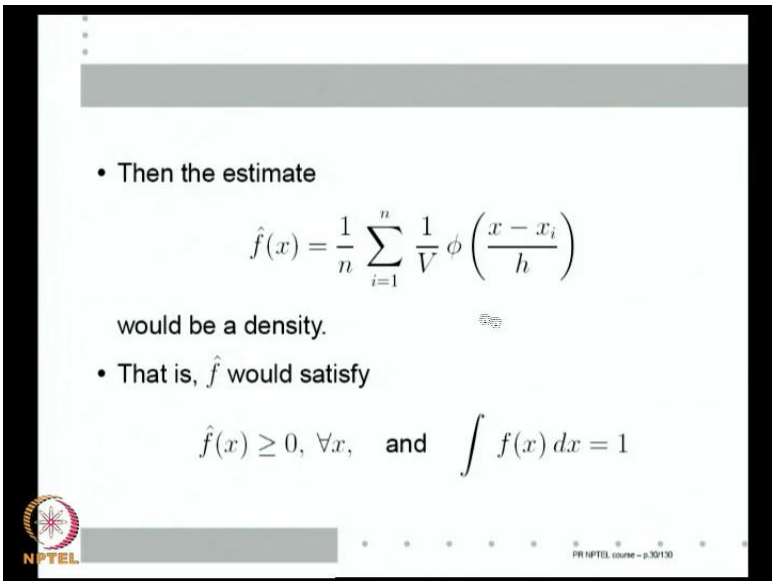
$$V = \int_{\Re^d} \phi\left(\frac{u}{h}\right) du = \int_{\Re^d} \phi\left(\frac{u - u_0}{h}\right) du$$

- For our hypercube window, $V = h^d$.

PR NPTEL course – p.28/130

So, let us ask what are the properties that phi should satisfy? Essentially we are only using for this to be a density estimate we only need two properties of phi; one is that phi should be positive and two integrate over r d phi should integrate to 1 that is to say phi should itself be a density. The reason is the following if phi satisfies this and we choose the volume that is that V in the denominator of my f hat to be this some over integral over d phi of u by h, because this integral is over all r d a translation does not make any difference, so integral phi u by h d u is same as integral phi u minus u by h d u by we can simply put u minus u 0 is equal to u prime then d u will be d d u d u prime integral is still be over r d.

(Refer Slide Time: 11:57)



- Then the estimate

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{V}\,\phi\left(\frac{x - x_i}{h}\right)$$

would be a density.

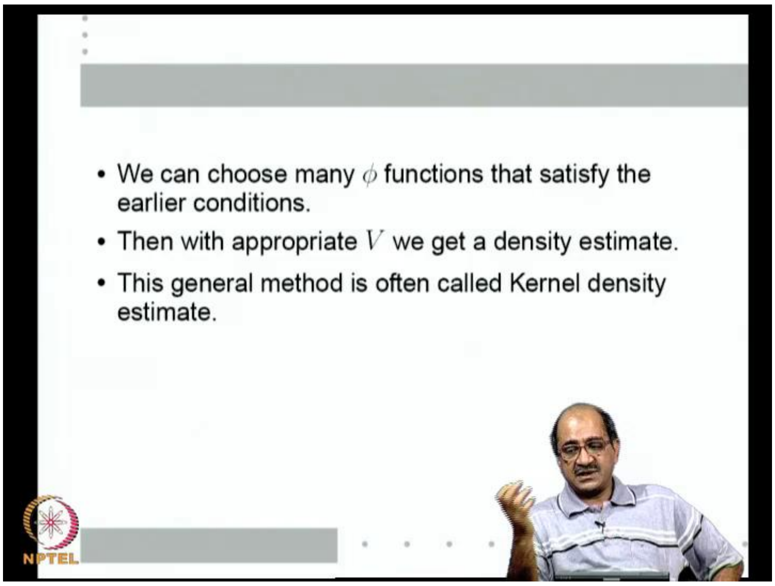
- That is, $\hat{f}$ would satisfy

$$\hat{f}(x) \geq 0,\ \forall x, \quad \text{and} \quad \int f(x)\,dx = 1$$

PR NPTEL course – p.30/130

So, if a function phi satisfies this and phi u h by d u integral of r d is V. Then this density estimate with this 1 by h d replaced by 1 by V would be a would be a proper density for our for our particular hypercube window function V a stand out to be h power d that is why we got that particular form for the Parzen window estimate. But in general as long as I define V by this quantity then this will be a density why this a density? Because by this definition of V and the properties of phi. This 1 by V phi x minus x i by h will integrate to 1 over x and I am summing n of them, so it will become n and divide by 1 by n it becomes 1.

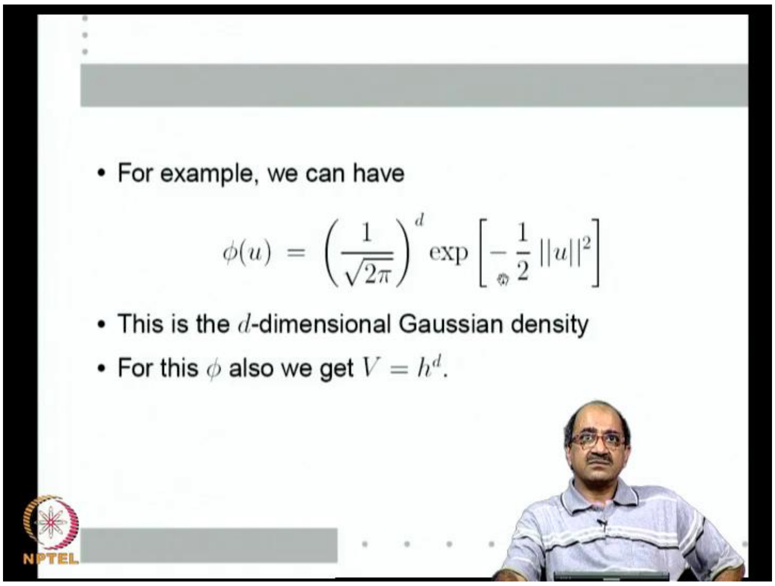
(Refer Slide Time: 12:51)



- We can choose many ϕ functions that satisfy the earlier conditions.
- Then with appropriate V we get a density estimate.
- This general method is often called Kernel density estimate.

So, essentially if I use this then f hat would be a density that is f hat will always be positive because phi is positive and f hat f hat x d x will be 1, because the way I defined V and the properties of phi would ensure that each term here integrates to 1. We can choose any, now that we understand that this is what we want out of our phi we can choose many other phi functions and correspondingly calculate the V and make different Parzen window estimates. We can choose from many function that satisfy these conditions then with appropriate V we will get a density estimate and all such a methods all such estimates are called kernel density estimates and the particular phi function we have chosen is called the Kernel function for that particular density estimate.

(Refer Slide Time: 13:29)



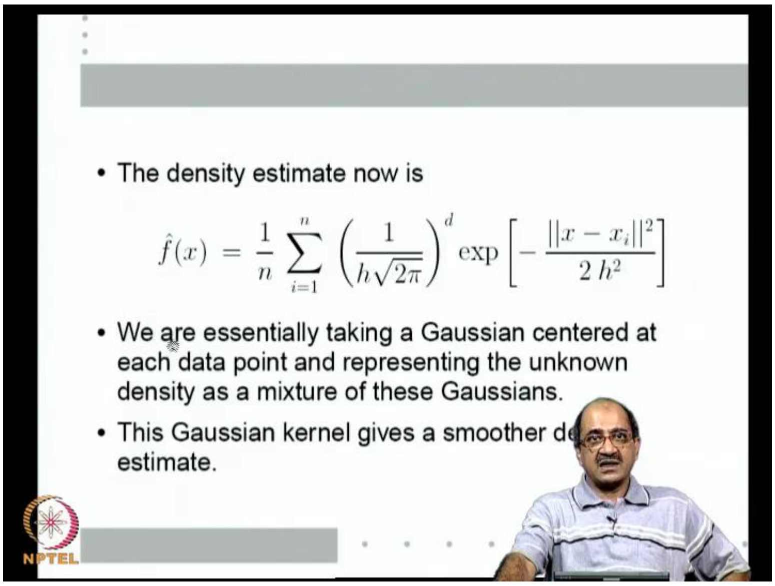
- For example, we can have

$$\phi(u) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{1}{2}||u||^2\right]$$

- This is the d-dimensional Gaussian density
- For this ϕ also we get $V = h^d$.

So, this is the general method of Kernel density estimates we essentially choose a phi function that satisfies these conditions and we choose V to be this integral. Then this estimate is such a density estimate this will be a density and such a density estimate is called the Kernel density estimate with the Kernel function phi for example; we can see at least one other phi function. Essentially, we wanted phi to be a density, so for example; I can choose phi to be gaussian density this is the d dimensional gaussian density. 1 by root 2 pi to the power d exponential minus half u (( )) mu square is essentially product of d 1 dimensional gaussian densities.

(Refer Slide Time: 14:56)



- The density estimate now is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{h\sqrt{2\pi}} \right)^{d} \exp \left[- \frac{||x - x_i||^2}{2\, h^2} \right]$$

- We are essentially taking a Gaussian centered at each data point and representing the unknown density as a mixture of these Gaussians.
- This Gaussian kernel gives a smoother de estimate.

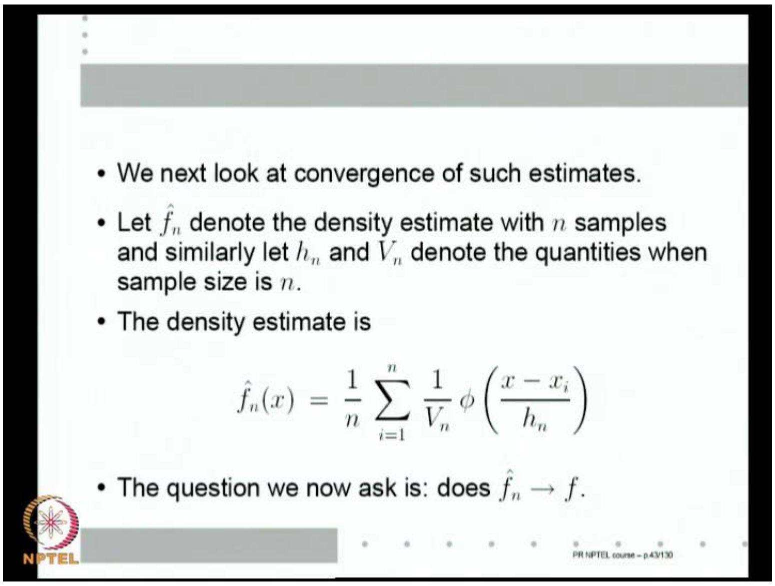
So, this is a gaussian Kernel or a gaussian window by simply integral finding that other integral we can show that for this function also the appropriate volume will be h power d. Essentially, I scale each u by h here, so then it will become u i square by h square, so h will become the variance, h square will become variance so here I will get h to 2 pi, so I will get a h power d term extra, so that is how V will become h power d. So, with V is equal to h power d this will be my density estimate f hat x is 1 by n i is equal to 1 to n 1 by h root 2 pi to the power d exponential x minus x i whole square by 2 h square.

As you can see this actually product of n 1 dimensional Gaussians each with variance h square, so what we call the size of the volume? And now is actually the variance of each of the x's. As you can see this is actually a mixture density. We are essentially taking Gaussians centred at e x i that is each of our data points and summing them. So, this is if I take 1 by n inside this is sum of n the n terms each one is 1 by n times a density, so it is a convex combination of densities, so this is a mixture density.

We have considered such Gaussian mixtures earlier; so instead of just choosing some Gaussian mixture and estimating it in an (( )) method to find the mixing coefficients and the parameters of the individual Gaussians. Here we are choosing exactly n Gaussians where n is the number of data points and centring each Gaussian at a data point, and keeping the Gaussians who have diagonal covariance matrix and the same variance and keeping variance of all of them to be h, which is our control parameter for the size of the volume element. So, with this phi function our Kernel density estimate is essentially a mixture of Gaussians by erecting 1 Gaussian itself.

Now, you as you can see it is no longer have the strictness of a unit hypercube window, because we are not really saying how many k actually fall inside this volume element because, now the volume element does not have a finite boundary, because Gaussian will go all the way up to infinite. But essentially we are erecting a Gaussian at each sample point and then representing our unknown density as a mixture of these Gaussians, so this also comes under the same generalised histogram kind of idea. And this is another form for Kernel density estimates very often one uses this, because this gives us a much smoother density estimate our original Parzen window estimate, which used at the unit hypercube has discontinuities at the boundaries of the hypercube.

(Refer Slide Time: 18:01)



- We next look at convergence of such estimates.
- Let $\hat{f}_n$ denote the density estimate with n samples and similarly let h_n and V_n denote the quantities when sample size is n.
- The density estimate is

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V_n} \phi \left(\frac{x - x_i}{h_n} \right)$$

- The question we now ask is: does $\hat{f}_n \rightarrow f$.

PR NPTEL course – p.43/130

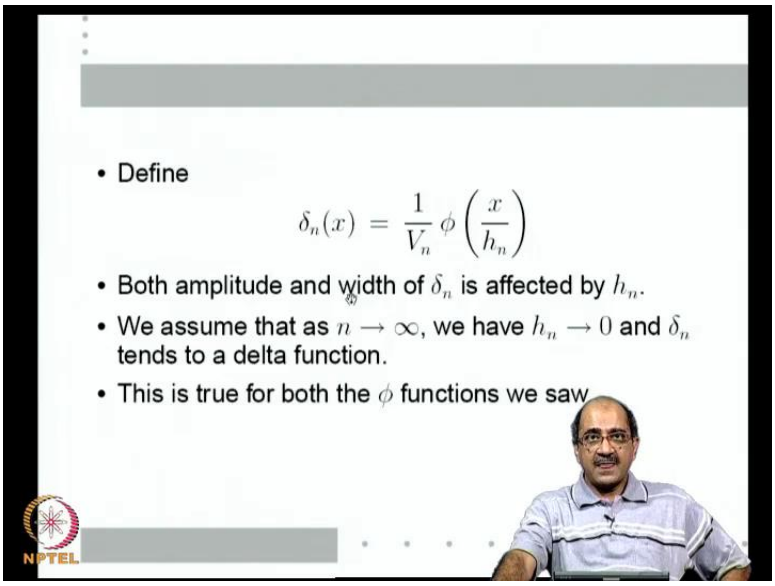
So, that kind of intrigues artificial discontinuities in my estimated density function whereas, this being sum of versions gives me very smooth density estimate, so for example, very often the choice of kernel function for a kernel density estimates is Gaussian and I and one uses this kind of a nonparametric or a kernel density estimate. Now, let us look at how such things works see, now this is we have moved quite a bit away from the simple intuitive idea of histogram we started with the histogram, then defined the window function to actually capture the histogram, then looked at the role that the window function is actually playing. We looked at what are the properties that function phi should satisfy? When we realised that by choosing different phi we can get very nice smooth estimates even though these are really counting estimates they are quite smooth.

Now, we can ask you know do such estimates converge by converge what we mean, is if I get very large number of data samples as number of data samples goes to infinity does this a estimate converge to the two density estimate this is like the question of consistency of estimate that we considered in the parametric case. For example; we we stated that maximum likelihood estimate is consistent in the sense as the number of data goes to infinity I get back the true parameter values.

In the same way I can ask if the number of our data samples goes to infinity do the Kernel density estimates actually give me the true density. So, let us look at this question next, so we will be looking at the convergence of these estimates. So, because we are looking at the convergence of these estimates we have to ask what will be the estimate when I have n data samples. So, let f hat n denote the density estimate with n data samples and similarly. Essentially, if I want convergence as n increases how to change the size of my volume element? Because if h remains constant, but number of data sample goes to infinity it would not work last time we seen what we need as n tends to infinity to get the true density estimate as n tends to infinity how to shrink the volume to 0.

So, the h should shrink to 0, so at different n values different h s, so let h n and V n denote these quantities when the sample size is n. So, which means I am looking at a sequence of density estimates f hat n x this is what I will get if I have n samples is 1 by n i is equal to 1 to n 1 by V n phi of x minus x i by h m. So, given that these are the sequence of density estimates I am getting the question that we want to ask, now is does this f hat n converts to f of course; this question converge is a little difficult here, because f hat as you can see is a function of x i and x i is are i i d sample they are random, so f hat n is random.

(Refer Slide Time: 20:58)



- Define

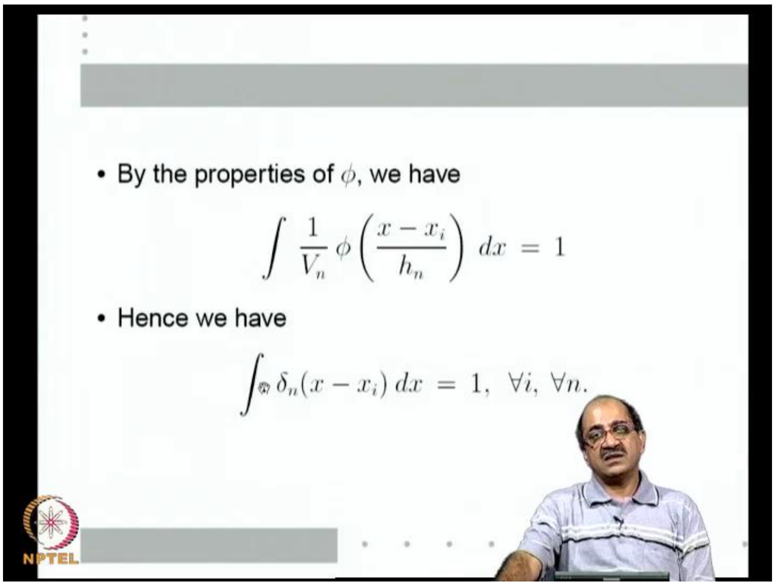
$$\delta_n(x) = \frac{1}{V_n}\phi\left(\frac{x}{h_n}\right)$$

- Both amplitude and width of δ_n is affected by h_n.
- We assume that as $n \to \infty$, we have $h_n \to 0$ and δ_n tends to a delta function.
- This is true for both the ϕ functions we saw

So, this is a sequence of random variables f hat n x, so I have to specify in what sense is this convergence to be understood and like in the consistency definition earlier we want to think of this as convergence in probability. So, we need a little more notation let delta n x represent one by V n phi of x by h n. Essentially, if I think of this as a function its amplitude is 1 by V n and its width is determined by h n, so as a change h n the width changes and also V n changes, so the amplitude changes.

Essentially, for both the functions both our hypercube window function as well as the Gaussian function as h n becomes smaller and smaller the width reduces whereas, 1 by V n increases the amplitude goes up and the width reduces, so both the amplitude and width of delta n are affected by h n and for both these functions as the width shrinks to 0 the amplitude goes to infinity.

(Refer Slide Time: 22:41)



- By the properties of ϕ, we have

$$\int \frac{1}{V_n} \phi\left(\frac{x - x_i}{h_n}\right) dx = 1$$

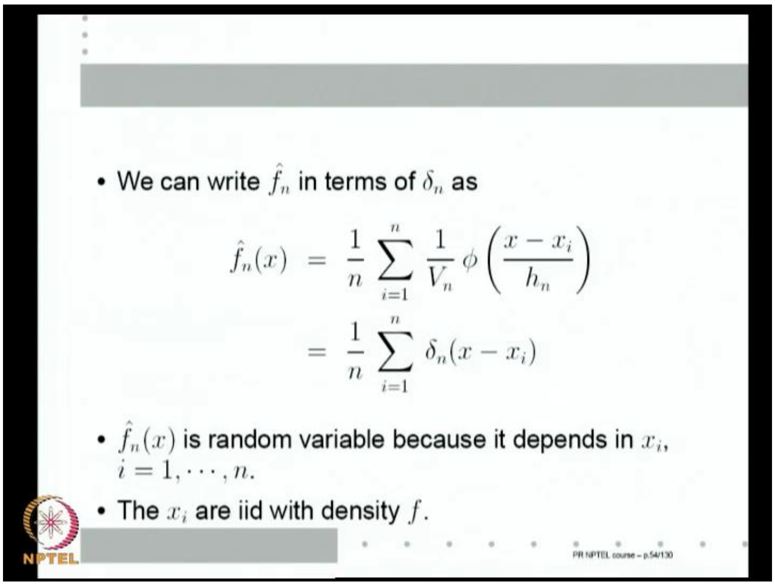
- Hence we have

$$\int \delta_n(x - x_i)\, dx = 1, \ \forall i, \ \forall n.$$

So, we assume that as n tends to infinity we have our h n goes to 0 in such a way that delta n tends to the delta function. We said as n tends to infinity you have to let h n go to 0, so you are saying h n should go to 0 in an appropriate fashion, so this function delta n tends to a proper direct delta function for both the phi. We both the hypercube function as well as the Gaussian function this is true, so if we simply let h n go to 0 with n then as h n goes to 0 as n tends to infinity delta n tends to a proper delta function.

Now, having defined delta n we will write our estimate f hat n in terms of the delta function. Now, one more property of the delta function that we are getting is we have we have defined V n to be this integral of phi x minus x n by x i by h, so 1 by V n phi x minus x i by h n d x is always 1, which means delta n x minus x i d x is 1 for all i all m. So, this is another a property of delta function, which ensures that as n tends to infinity this delta n converges to a true direct delta function.

(Refer Slide Time: 23:21)



- We can write $\hat{f}_n$ in terms of δ_n as

$$\begin{aligned}\hat{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{V_n}\phi\left(\frac{x-x_i}{h_n}\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\delta_n(x-x_i)\end{aligned}$$

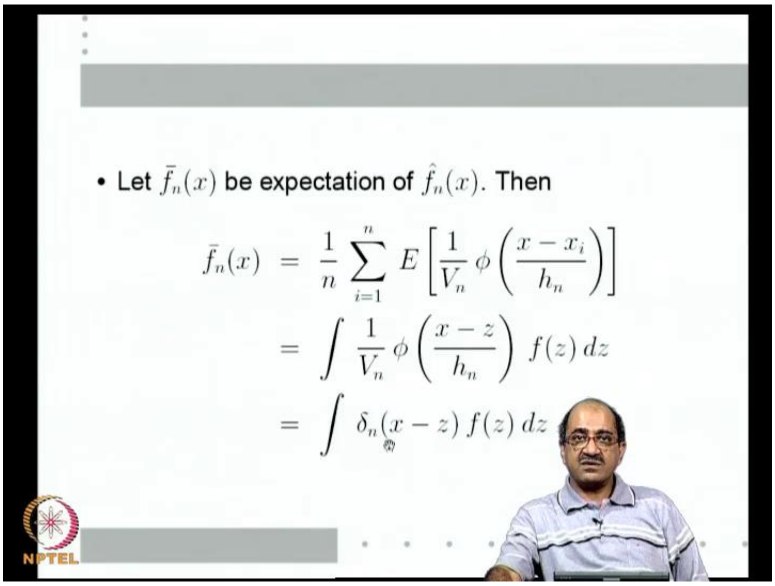
- $\hat{f}_n(x)$ is random variable because it depends in x_i, $i = 1, \cdots, n$.
- The x_i are iid with density f.

PR NPTEL course – p.54/130

So, we can write f hat in terms of delta n as f hat n a layer is 1 by n 1 by v n phi x minus x i by h n as I said earlier the reason I kept one by n outside as a purpose that is because, now I can put this as delta i. I can write this as 1 by n delta n x minus x i, so this is my estimate in terms of delta n. Now, as we already discussed f hat n is random variable f hat n x is a random variable, because it depends on this i i d a data sample, which are random x i, so this is a random variable there is a function of this x i and let us also remember that x i are i i d and each x i has density f the unknown density.

The f is what we do not know and we are estimating about the f hat and each of these x i’s are i i d according to f both these facts are useful to us in ultimately obtaining the convergence to obtain. The convergence, we have to ask is f hat n x converges in probability to f x for that what we look at is what happens to the mean and variance of f hat n x tends to infinity.

(Refer Slide Time: 24:34)



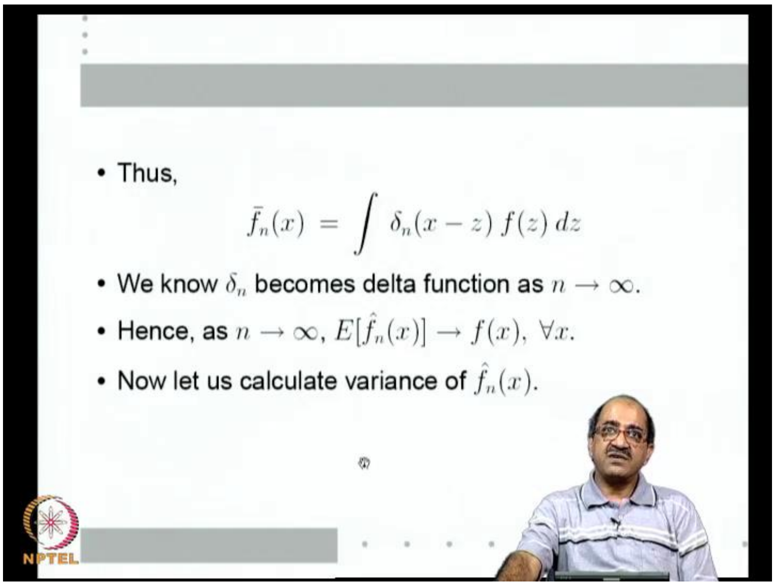
- Let $\bar{f}_n(x)$ be expectation of $\hat{f}_n(x)$. Then

$$\begin{aligned}\bar{f}_n(x) &= \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{1}{V_n}\phi\left(\frac{x-x_i}{h_n}\right)\right]\\ &= \int \frac{1}{V_n}\phi\left(\frac{x-z}{h_n}\right) f(z)\, dz\\ &= \int \delta_n(x-z)\, f(z)\, dz\end{aligned}$$

Let f hat and x be the mean the bar f bar n x be the mean of f hat n x, so f hat n x is expectation of f bar n x is expectation of f hat n, so expectation goes inside the sum I put it there. What is inside this expectation? Is a function of the random variable x i and x i's are i i d, so this expectation of this same value for each i and, so I am summing them and dividing by n, so this entire thing will have the same value as expectation of any one of these terms.

Now, expectation of any one of these terms is easy to write, because in this expectation what is random is this x i and I know the density of x i no many f i I know the symbol for the density of x i x i are x i have density f, so I can write this entire expression as integral one by V n phi of x minus z by h n f z d z, because this x i is the random variable which has density f.

(Refer Slide Time: 26:25)



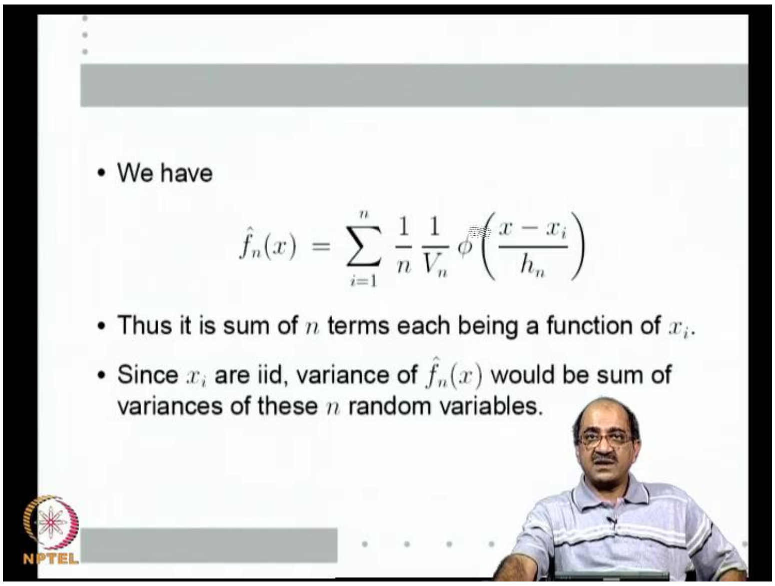
- Thus,

$$\bar{f}_n(x) = \int \delta_n(x-z)\, f(z)\, dz$$

- We know δ_n becomes delta function as $n \to \infty$.
- Hence, as $n \to \infty$, $E[\hat{f}_n(x)] \to f(x),\ \forall x$.
- Now let us calculate variance of $\hat{f}_n(x)$.

So, the expectation inside is given by this integral as you said x i r i i d, so each of these expectations are the same value and then I divide by 1 by n, so both 1 by n and the summation will go and the f bar n x will become this. Now, we already have a symbol for this, this is delta n, so I can write this as delta n x minus z f z d z. So, I know the that f bar n x, which is the expected value of f hat n x is given by integral delta n x minus z f z d z for all n. Now, to ask what happens to the mean as n tends to infinity. We have to ask what happens to this as n tends to infinity. So, we have this and as n tends to infinity we know that delta n becomes the delta function, when delta n becomes the delta function integral delta x minus z f z d z will be f x that is the property of the delta function.

(Refer Slide Time: 27:20)



- We have

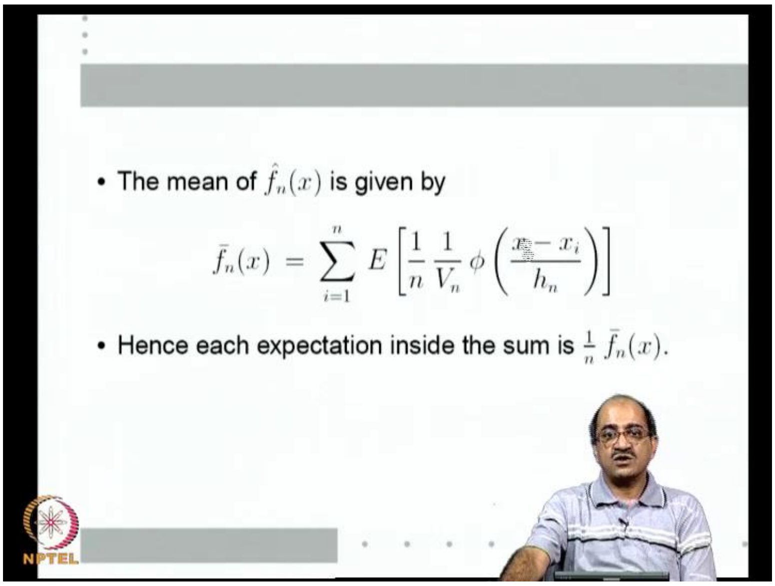
$$\hat{f}_n(x) = \sum_{i=1}^{n} \frac{1}{n} \frac{1}{V_n} \phi\left(\frac{x - x_i}{h_n}\right)$$

- Thus it is sum of n terms each being a function of x_i.
- Since x_i are iid, variance of $\hat{f}_n(x)$ would be sum of variances of these n random variables.

So, because we know as n tends to infinity delta n becomes the delta function this equation tells us that as n tends to infinity f bar n x, which is expected value of f hat n x goes to f x. So, as n tends to infinity the mean of f hat n x goes to f x, so if I can show that the variance goes to 0 then this means f hat n x converge in probability to f x, now that we know that the mean converges to f x. Let us look at what happens to the variance of f hat n x.

The f hat n x is given by i is equal to 1 to n 1 by n 1 by v n phi x minus x i by h n i just put 1 by n inside for a purpose, so f hat n is the sum of some n terms each one is a random variable, it is a sum of n terms each of function of x i. Now, x i is are independent, so this random variable is sum of some n random variables they are independent to each other each of these are independent, because for different i these are functions of different x i and, hence they are also independent of each other, so f hat n x can be represented as sum of n independent random variables. Since x i are independent variance of f hat n would be sum of variances of this n random variables, so essentially to find f hat n x I have to find variance of each of these and sum them up.

(Refer Slide Time: 28:58)



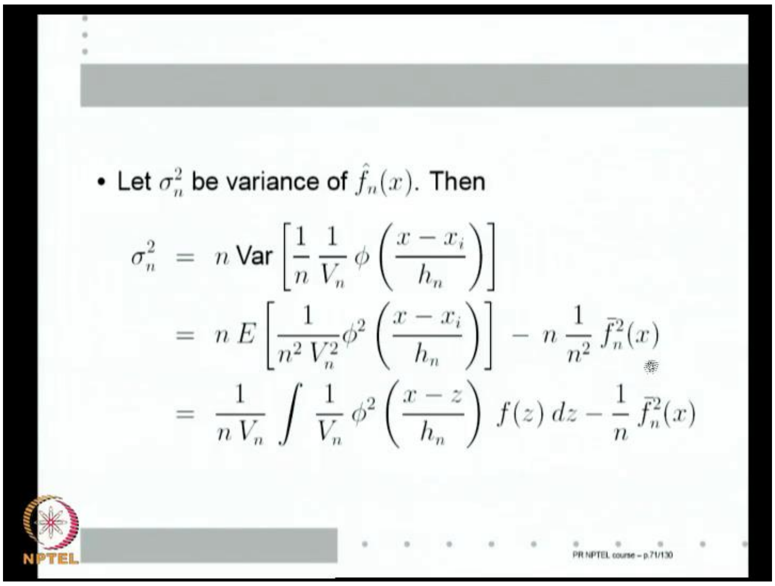
- The mean of $\hat{f}_n(x)$ is given by

$$\bar{f}_n(x) = \sum_{i=1}^{n} E\left[\frac{1}{n}\frac{1}{V_n}\phi\left(\frac{x-x_i}{h_n}\right)\right]$$

- Hence each expectation inside the sum is $\frac{1}{n}\bar{f}_n(x)$.

Now, each of these are i i d this each term is dependent on the random variable x i and the same function they are all the same functions of x i, x i, r i, i d,so the variance of each of them would be same. So, variance of f hat n will be n times variance of any one of them to find variance I have to find variance of any random variable z is expected value is z square minus expected by z whole square, so I need to find the expectation of each of these and then also expectation of square of each of these. First let us look at the expectation the mean of f hat n which which we called f bar n x is i is equal to 1 to n expectation 1 by n of this.

(Refer Slide Time: 29:42)



- Let σ_n^2 be variance of $\hat{f}_n(x)$. Then

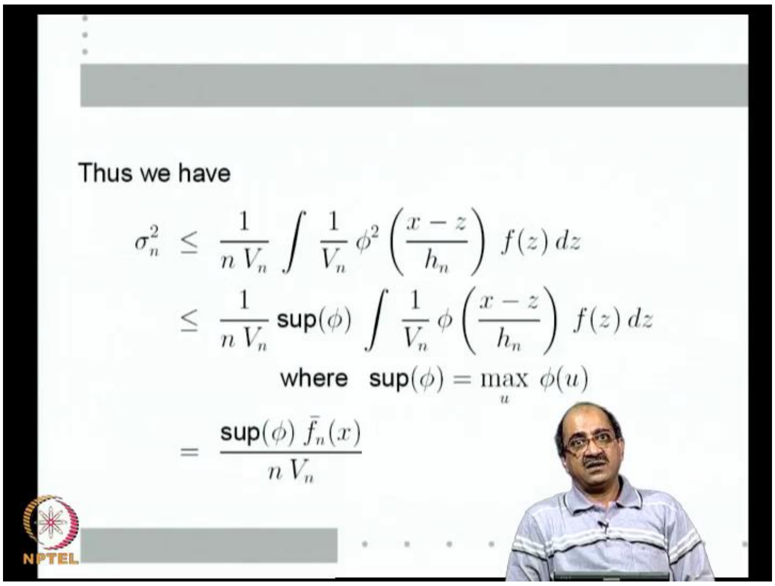
$$\begin{aligned}\sigma_n^2 &= n\,\mathrm{Var}\left[\frac{1}{n}\frac{1}{V_n}\phi\left(\frac{x-x_i}{h_n}\right)\right]\\ &= n\,E\left[\frac{1}{n^2 V_n^2}\phi^2\left(\frac{x-x_i}{h_n}\right)\right] - n\,\frac{1}{n^2}\,\bar{f}_n^2(x)\\ &= \frac{1}{n\,V_n}\int \frac{1}{V_n}\phi^2\left(\frac{x-z}{h_n}\right) f(z)\,dz - \frac{1}{n}\,\bar{f}_n^2(x)\end{aligned}$$

PR NPTEL course – p.71/130

Now, we know that this is equal to f bar, now each expectation is same and sum of n of them is equal to f bar, so each one of them is 1 by n f bar so, hence the each expectation inside the sum is 1 by n f bar n x, which means each of these random variables how mean 1 by n f bar n x and we need to find the variances of each of them and sum them up. And of course, we do not have to sum them up each of them have the same variance, so we take n times variance of any one of them. So, if we think of sigma square n as the variance of f hat n x then we can write sigma square n as n times variance of this. As I said f hat n x is sum of these n by things each of them have the same variance, so variance of f hat n will be n times variance of each of these.

So, sigma square n will be n times variance of 1 by n 1 by v n phi x minus x i by h n where x n is the random variable with respect to, which i am finding variance of this term. So, what is variance expectation of this square minus expectation of this whole square expectation of this. We already know is 1 by n f bar n, so I can write this as n times expectation of this square of this 1 by n square 1 by V n square phi square x minus x i by h n minus expectation of this whole square once again n times n times expectation of each of them is f n by n, so square is f bar n square by n square, so that is sigma square n.

(Refer Slide Time: 32:06)



Thus we have

$$\sigma_n^2 \leq \frac{1}{n\,V_n} \int \frac{1}{V_n} \phi^2\left(\frac{x-z}{h_n}\right) f(z)\,dz$$

$$\leq \frac{1}{n\,V_n} \text{sup}(\phi) \int \frac{1}{V_n} \phi\left(\frac{x-z}{h_n}\right) f(z)\,dz$$

where $\text{sup}(\phi) = \max_u \phi(u)$

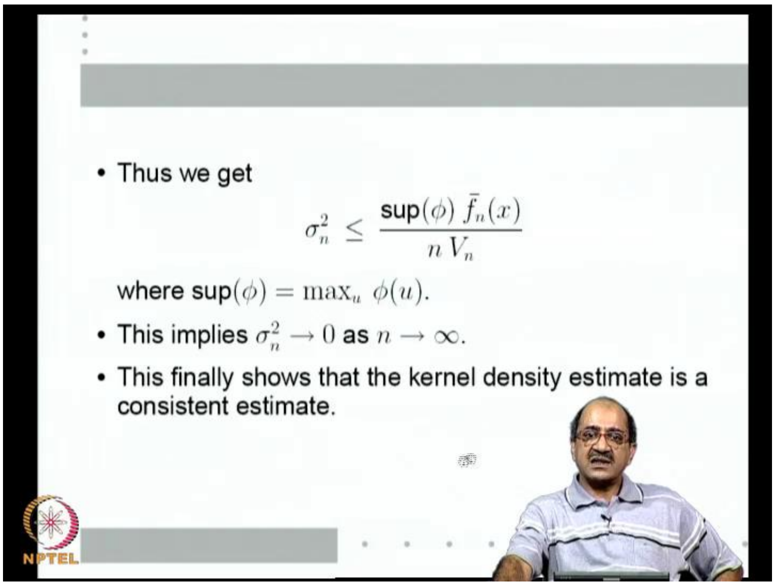
$$= \frac{\text{sup}(\phi)\,\bar{f}_n(x)}{n\,V_n}$$

Now, this term is always positive n into 1 by n square into f bar n whole square, so sigma n square is this minus this, so if I drop this term all, sigma n square I get the before I drop the term. Let me write as well first get this expectation out this is expectation of this where x i is the random, so I can write this as this n will cancel with one of these n's, so I can take 1 n V n out I get 1 by V n phi square x minus z by h n f z d z, because each of the x i's are i i d with density f, so expectation of this is given by this expectation integral the integral is over r of d minus this is 1 by n f bar square n x.

Now, given this as I said sigma square n is this minus this and this is always a positive term, so if I drop this I can write sigma square n is less than or equal to 1 by n times V n integral 1 by V n phi square x minus z by h n into f z d z. It is the same thing if first term alone. The reason for writing it like this is that 1 by V n phi x minus z h n f z d z is something we already know that is f bar n, so what I can do is I can take 1 phi out, so this is phi square x minus z if I want to take 1 phi out of the integral I can substitute it with the maximum possible value of I can have. Let us call it supremum of phi supremum of phi is max over u phi u for all the window functions we have considered this is finite.

So, if I take that out, so 1 phi I have taken out I left 1 by V n phi x minus z by h n, f z ,d z and we have already seen this integral, so that will give me f bar n, that is how the mean of f hat n is defined. So, what I get, finally is supremum of phi f bar n x n into v n f bar n x is finite this is the mean of these random variables supremum of phi is finite as we have seen the way we lead this sizes go to 0 in the previous class we said we have to have n v n should go to infinity.

(Refer Slide Time: 33:55)



- Thus we get

$$\sigma_n^2 \le \frac{\sup(\phi)\,\bar{f}_n(x)}{n\,V_n}$$

where sup$(\phi) = \max_u \phi(u)$.

- This implies $\sigma_n^2 \to 0$ as $n \to \infty$.
- This finally shows that the kernel density estimate is a consistent estimate.

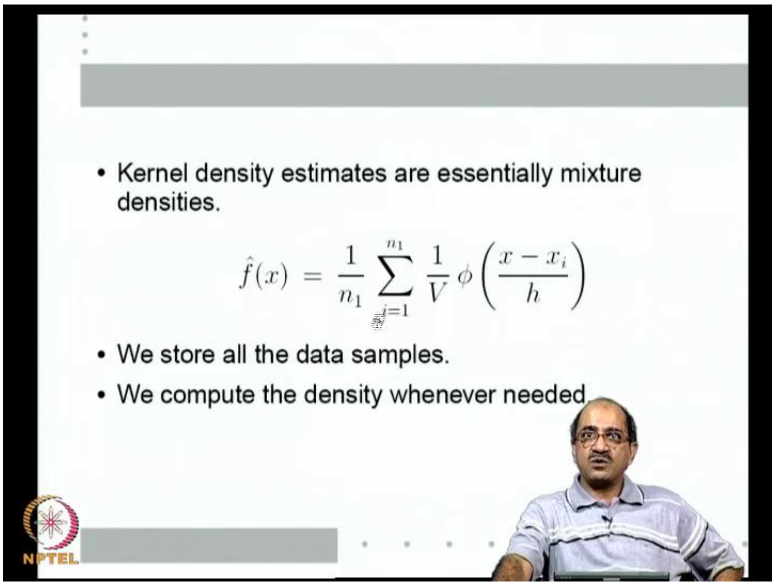
So, as n tends to infinity sigma square n becomes 0, so we have sigma square n less than this and this implies sigma square n goes to 0 as n tends to infinity. So, what we have shown is as n tends to infinity the mean of f hat n f hat n x at any given x becomes f of x and its variance goes to 0, so as we get large samples and we choose over volume element h n such that h n goes to 0 as n tends to infinity is such a way that this n V n goes to infinity see as h n goes to 0 the volume elements shrinks to 0, but we should shrink it in such a way that n times V n should go to infinity.

So, if we let out h n go to infinity in the in the way that means we choose our h correctly as n tends to infinity. Then we can show that f hat n x are such that as n tends to infinity its mean, becomes f of x and its variance goes to 0 that is showing that f hat n x converges in probability to f x and that tells us that kernel density estimate is a constituent estimate. So, to sum up we came up with Kernel density estimates by starting with this simple intuitive idea of a histogram you simply cut your space into bins and count how many points fall in it out of n.

Then using binomial theorem and assuming that the density is constant over each at the bins we can get a very simple formula for estimate that is f hat n is k by n V n where V is the volume element, n is the total number of samples and k is the number of sample that fall in V. Then we have seen we can write this using a proper window function we first started with a rectangular window function with the Parzen window, so called Parzen window that exactly counts like this.

And then written it out written out our f hat n x and that showed us that you know we if you have that phi function and all the x i with us we can simply compute f hat x at any x that we have it also showed us that window function. We can chose any other window function instead of a simple hypercube window function and essentially as long as we keep phi as a density function it will work. Then we generalise it to a Kernel density, which is essentially a mixture density form where we erect one density centred at each of the sample point.

(Refer Slide Time: 37:24)



- Kernel density estimates are essentially mixture densities.

$$\hat{f}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{V}\phi\left(\frac{x-x_i}{h}\right)$$

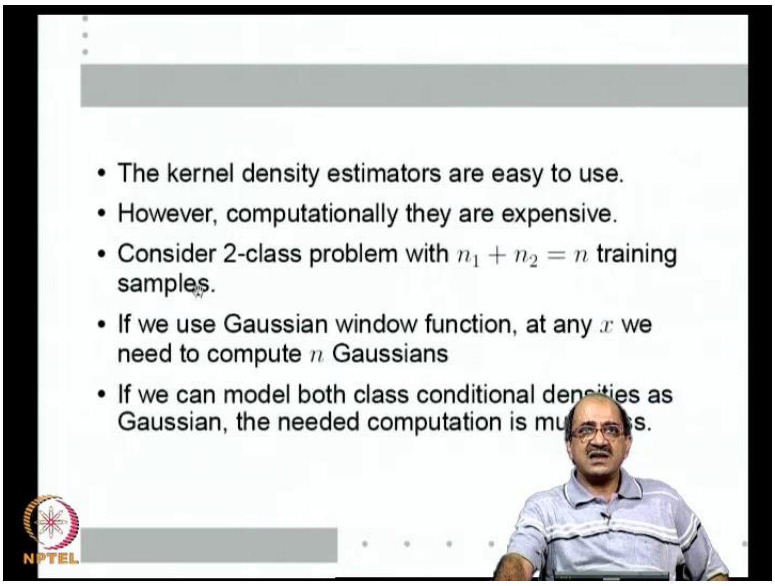
- We store all the data samples.
- We compute the density whenever needed.

For example; when we use the Gaussian function at the kernel when with the Gaussian Kernel the Gaussian Kernel density function, Kernel density estimate is the is such that at each of the sample points x i we erect a Gaussian. And then take a uniform convex combination of them as our density estimate that is the kernel density estimate, and for that kernel density estimate we showed just like our other parameter density estimates that the density estimate is consistent that is as the number of samples goes to infinity we get the true estimate of course, at any finite samples we we have to know what has to choose, but at least it is nice to know that asymptotically it is consistent density estimate.

So, Kernel density estimates are essentially mixture density estimates we store all the data samples and then compute the density wherever needed. Essentially, depending on how many data samples we have for example; if I have n one data samples, so we erect n 1 such densities each density centred at one of the data samples and then do this mixture density estimate. In general one particular form for Kernel density estimate that is often preferred is the Gaussian Kernel and the idea is that we do not have to store any density estimate formula.

We only know the formula for phi and we store all the data samples x i and given any x, whenever you need to compute f hat we compute it using this formula for any number of samples we have that is why I change it n to n 1 to say that in practice we may have some specific number of samples, so we cannot use the asymptotic estimate that is why we used a fixed V n fixed h.

(Refer Slide Time: 38:41)



- The kernel density estimators are easy to use.
- However, computationally they are expensive.
- Consider 2-class problem with $n_1 + n_2 = n$ training samples.
- If we use Gaussian window function, at any x we need to compute n Gaussians
- If we can model both class conditional densities as Gaussian, the needed computation is mu s.

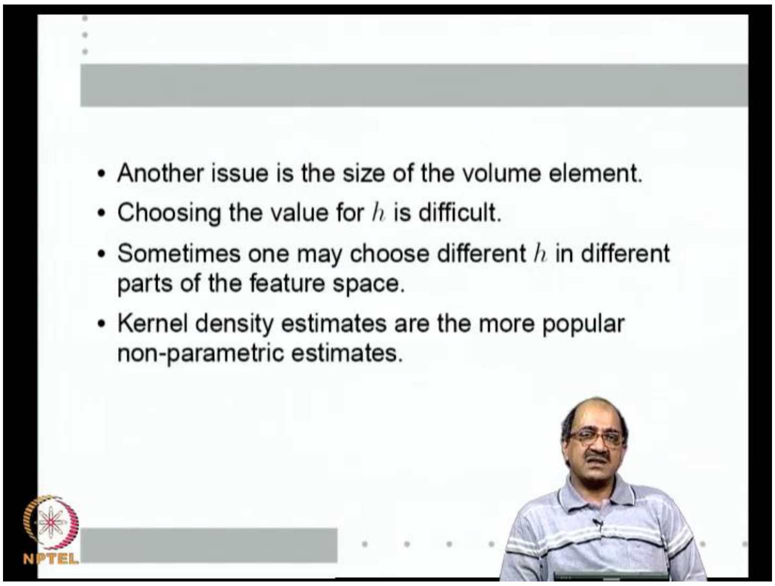
Now, Kernel density estimators are easy to use, so you simply store all the x i and you can compute it at any x that you need, however computations could be expensive when I want to compute f hat x I need to compute all these phi’s. So, if I have n one that samples in one class and a feature of the phi’s are Gaussian or given any particular x if I have to compute f hat x I have to actually compute n 1 Gaussian functions, because Gaussian function has exponential in it that is the most expensive computation.

So, I have to actually compute n 1 Gaussians if I have n 1 data samples data samples could be in hundreds, so this is a non trivial amount of extra computation. So, computationally they are expensive, so as I said in particularly for two class problem with n 1 data samples from one class and n 2 data samples from the other class, so to compute the first class one class conditional density I need to do n one phi computations and similarly, for the class two density I need to do n two phi computations.

So, if we use a Gaussian window function at any given x we need to compute n Gaussians. On the other hand if we can if it is feasible to use a density model if it is feasible to assume that the class conditional densities are Gaussian then the computation is very simple, if we can model both the class conditional densities are Gaussian we need computation much less we need to compute only two Gaussians. So, Kernel density estimates are expensive obviously we will go for such expensive estimates only if we feel that no simple density model will work for this class conditional densities.

So, if the class conditional densities are such that we think we can model them with Gaussian or 1 or mixture of 1 or 2 Gaussians it is always good to model them like that. Only if we think that the class conditional densities can vary, very much they have very heavily multi model there is when we use Kernel density and then obviously we cannot complain about the extra expense, because the problem is difficult.

(Refer Slide Time: 41:09)



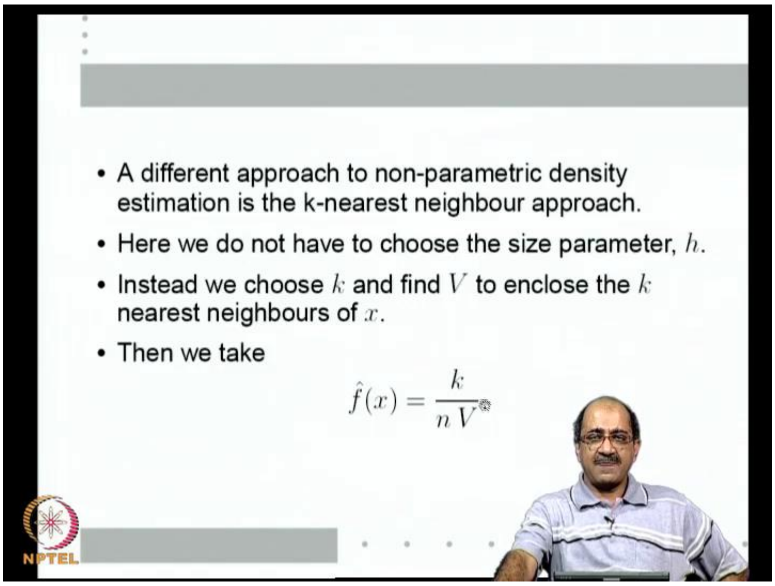
- Another issue is the size of the volume element.
- Choosing the value for h is difficult.
- Sometimes one may choose different h in different parts of the feature space.
- Kernel density estimates are the more popular non-parametric estimates.

Another issue with Kernel density estimate is the size of the volume element. As I said the choice of h is critical at any given time we have only finitely many samples, so you have to choose a particular h and there are no simple rules to say, which h is reasonable for a given sample size. Also the same value of h may not be good at all points in the in the feature space, because the the class conditional density is obviously not uniform over the featured space at some places the density will have a high value some place will have low value.

So, in our sample set also we get many more samples in certain regions of the feature space and very few samples in some other regions of feature space. So, using this same h throughout the feature space may not also be a good idea, but on the other hand we do not know how to vary it. One can choose different h in different parts of the feature space. But once again what rule do I use to adopt h like that one method is to say that if if I want the if I want the density at a particular point x if there are seem to be too many samples around that x then I can use a small h otherwise I can use a big h.

This this goes a little bit like the k nearest neighbour estimate there are few (( )) to do such a (( )) adapt the choice of h, but there are no no well grounded theoretically well grounded rules for this. In spite of all these problems in spite of its computational cost, in spite of the difficulty in choosing h Kernel density estimate are still the most popular nonparametric estimates, because if I want to do a nonparametric estimate I will go for a nonparametric estimate only when the problem is difficult and I cannot easily capture the density model in, which case kernel density estimates are about as good as one can get.

(Refer Slide Time: 43:22)

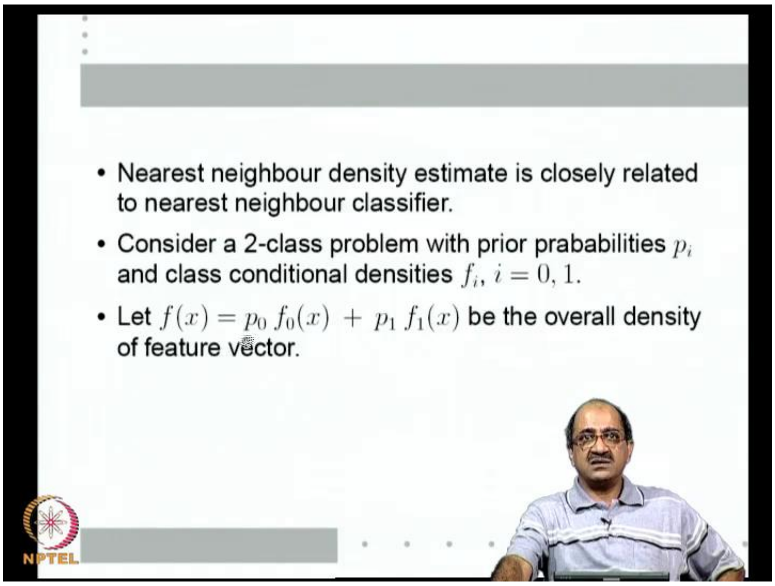


- A different approach to non-parametric density estimation is the k-nearest neighbour approach.
- Here we do not have to choose the size parameter, h.
- Instead we choose k and find V to enclose the k nearest neighbours of x.
- Then we take

$$\hat{f}(x) = \frac{k}{n\,V}$$

But anyway we have another method as we said the k nearest neighbour approach, so let us briefly consider this also generally k nearest neighbour approach is not so much preferred for not so much preferred for a density estimate very often it is preferred more for regression problems where they want to do a function estimate. But any case it is nice thing is that we do not have to choose h instead we have to choose k, which at least intuitively looks much easier I want to know how many nearest neighbour samples should I rely should I have one should I have, three should I have, five that might be little more easier to guess.

(Refer Slide Time: 44:49)



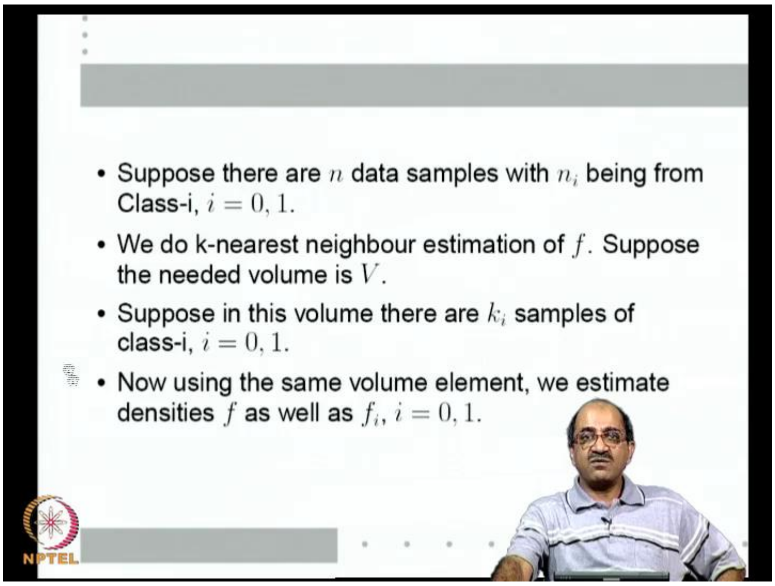
- Nearest neighbour density estimate is closely related to nearest neighbour classifier.
- Consider a 2-class problem with prior prababilities p_i and class conditional densities f_i, $i = 0, 1$.
- Let $f(x) = p_0\, f_0(x) + p_1\, f_1(x)$ be the overall density of feature vector.

So, we choose a k and find a volume to enclose k and that is how we get the nearest neighbour estimate. Once again the estimate is same k by n V, so we first choose k, so given an x I will ask where is the k th nearest neighbour and I draw a sphere a hyper sphere to just include the k th nearest neighbour its radius will be the distance from x to the k th nearest neighbour that will be the value of V. And then this will be my nearest neighbour k nearest neighbour density estimate what we would like to consider in this is one interesting relationship between such a nearest neighbour density estimate and the nearest neighbour classifier.

If you remember in our second class we considered a simple classifier, which we have called the nearest neighbour classifier. What does the nearest neighbour classifier do? It just stores all the x i all the training samples then give it any nu x from the training samples it finds what is the nearest training sample to x? And if the nearest training sample to x is in class one I will put x also in class one if the nearest training sample is in class two I will put x also in class two and so on, so that is the nearest neighbour.

The k-nearest neighbour classifier is instead of finding one nearest neighbour I find the k-nearest neighbours of a of a sample. And then ask the majority class I i ask, which are the let us say out of the k nearest neighbours k i are in class i I am asking for, which i k i is largest. So, if of the five nearest neighbours if two are in one class and three are in the other three are in class two then I will put x one also in class two that is what k nearest neighbour is.

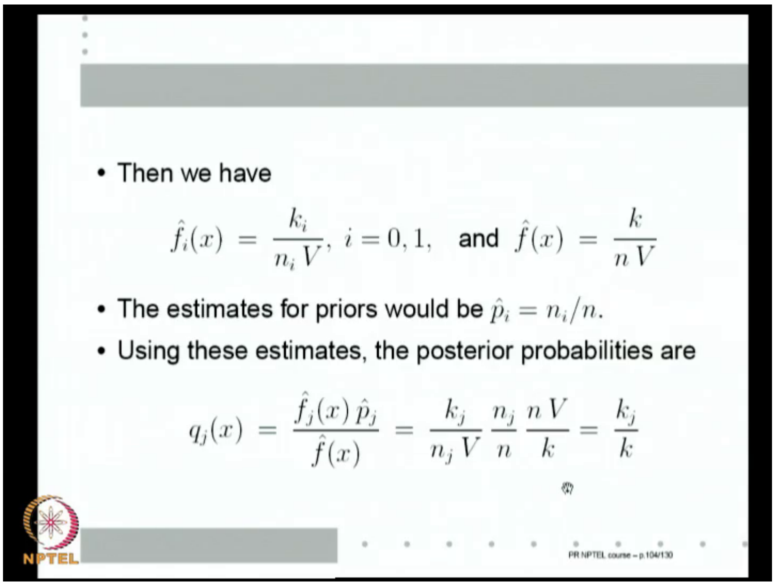
(Refer Slide Time: 46:36)



- Suppose there are n data samples with n_i being from Class-i, $i = 0, 1$.
- We do k-nearest neighbour estimation of f. Suppose the needed volume is V.
- Suppose in this volume there are k_i samples of class-i, $i = 0, 1$.
- Now using the same volume element, we estimate densities f as well as f_i, $i = 0, 1$.

So, the nearest neighbour density estimate is closely related to the nearest neighbour classifier. Let us say we have a two class problem with prior probabilities p i and class conditional probabilities f i i is equal to 0 1 1 that is p 0 p 1 are the 2 prior probabilities f 0 f 1 are the 2 class conditional densities. And then f x is p 0 f 0 plus p 1 f 1 is the overall density of the feature vector x, right? This is the throughout the class label this is the overall density of the f x, suppose there are n data samples of both classes of out of which n 0 are from class 0 and n 1 are from class 1. Let us say we do k nearest neighbour estimate of f the overall density say in the overall density we do not have to consider class labels. So, I have got n data samples and I want to estimate the overall f and suppose at a particular x for the k-nearest neighbour estimate the needed volume element is V.

(Refer Slide Time: 47:41)



- Then we have

$$\hat{f}_i(x) = \frac{k_i}{n_i V}, \; i = 0, 1, \quad \text{and} \quad \hat{f}(x) = \frac{k}{n V}$$

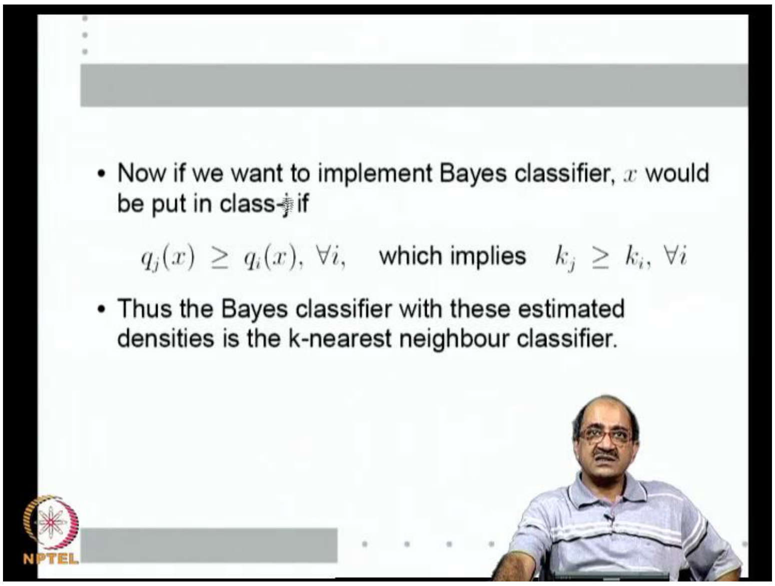
- The estimates for priors would be $\hat{p}_i = n_i / n$.
- Using these estimates, the posterior probabilities are

$$q_j(x) = \frac{\hat{f}_j(x)\, \hat{p}_j}{\hat{f}(x)} = \frac{k_j}{n_j V} \frac{n_j}{n} \frac{n V}{k} = \frac{k_j}{k}$$

PR NPTEL course – p.104/130

Suppose, in this volume in, which I know there are k neighbours of x k i of the neighbours are of class i, so k 0 from class 0 and k 1 from class 1. Now, let us say using this same V I can now estimate f as well as f i, because for f i I have n i samples of, which k are in this volume and for f I have n samples of, which k are in this volume, so using the same volume I can simultaneously estimate f 1 as well as f 0. Let us, suppose we do this estimate then what will be f i hat k i out of n i are in this volume V, so it is k i by n i V and f hat is k by n V, because k out of n are in this volume. What will be my prior estimates? p i hat will be simply n i by n thus p 0 hat will be n 0 by n and p 1 hat is n 1 by n this is the simple Bernoulli estimate for priors.

(Refer Slide Time: 48:52)



- Now if we want to implement Bayes classifier, x would be put in class j if

$$q_j(x) \geq q_i(x), \forall i, \quad \text{which implies} \quad k_j \geq k_i, \forall i$$

- Thus the Bayes classifier with these estimated densities is the k-nearest neighbour classifier.

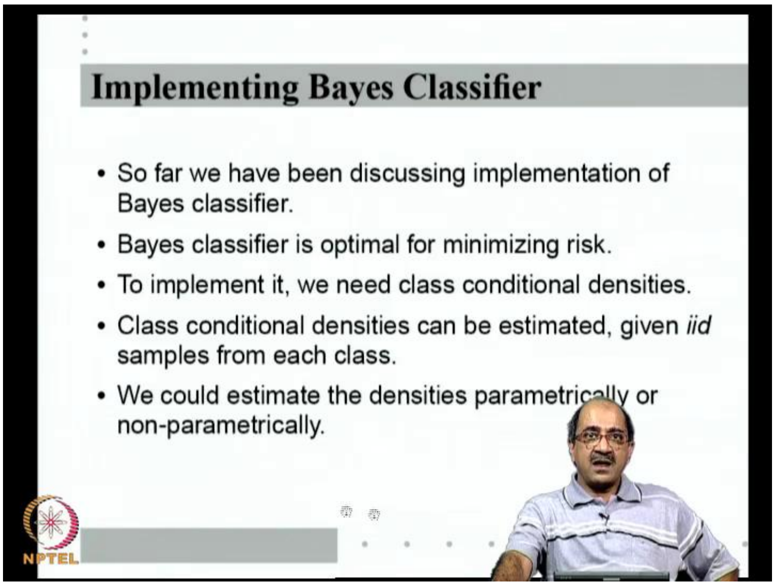
Now, let us say using these estimated priors and class conditional densities we want to calculate the posterior probabilities. So, posterior probability q j using the estimated quantities this is f hat j p hat j by f hat f hat j x p hat j x by f f hat x, so substitute for this f hat j will be k j by n j V p hat j is n j by n and 1 by f hat x will be n V by k. Now, these n j is will cancel V is will cancel n will cancel this is simply become k j by k. So, what does this mean? So, if we want to implement Bayes classifier x would be put in class j if q j x is greater than q x. Let us take 0 1 loss function for simplicity, so if the posterior probability of class j is greater than posterior probability of class i I put x in class j.

Now, q j x greater than q a x same as k j greater than k. What are k j and k i? Out of the k nearest neighbours of x k i are in class i k j are in class j, so if class j neighbours are more than class i neighbours out of k, so out of k neighbours if the j th class is the majority class that is what k j greater than k i means then Bayes classifier says put it in q j. So, if I use a nearest neighbour density estimate and use that estimate to implement Bayes classifier. Then what I get is the nearest neighbour classifier. As the matter of fact we do not have a complete the nearest neighbour density estimate also we can study asymptotically. If we study asymptotically like that one can actually show that the nearest neighbour classifier has a very interesting property that its error rate will never be more than twice that of the Bayes rate.

So, if Bayes error is p star then the error rate of a nearest neighbour classifier p will be less than twice p star asymptotically that is as the number of samples goes to infinity the nearest neighbour classifier, so worst case error rate is always bounded above by twice the Bayes rate that comes, because essentially the nearest neighbour classifier is the Bayes classifier implemented through this kind of density estimates. That is a very interesting thing.

Because if I think optimal Bayes error is let us say 05.00, so if I had actually estimated my class conditional density is exactly and implement based classifier I expect to get 95 percent correct and essentially doing nothing just doing a nearest neighbour classifier I can hope to get 90 percent accuracy. So, this is another reason why a nearest neighbour classifier is always used as a benchmark to see whether a complicated classification method is needed in a given application.

(Refer Slide Time: 51:39)



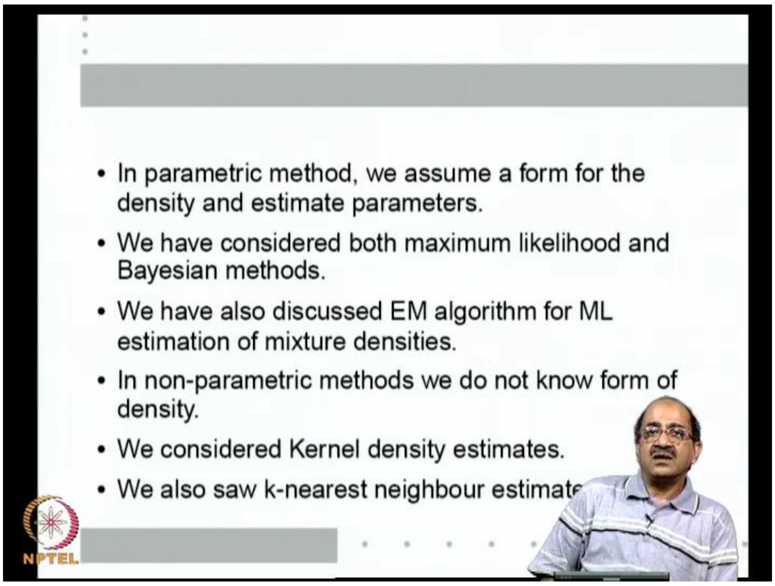
Implementing Bayes Classifier

- So far we have been discussing implementation of Bayes classifier.
- Bayes classifier is optimal for minimizing risk.
- To implement it, we need class conditional densities.
- Class conditional densities can be estimated, given *iid* samples from each class.
- We could estimate the densities parametrically or non-parametrically.

So, but anyway move to the point in our current context essentially the k nearest neighbour density estimate is such that the if you use such density estimates and implement the Bayes classifier based on such a density estimate what we get is the k nearest neighbour classifier. Now, let us sum up almost by the last seven eight classes we have been considering issues involved in implementing the Bayes classifier. So, all these last seven eight classes are a little more we have been discussing essentially implementation of Bayes classifier.

The idea is that Bayes classifier is optimised optimal for minimizing risk if you give me the class conditional densities and prior probabilities under any cross function Bayes classifier is the best risk. So, that is the reason why one would like to implement based classifier, but to implement it we need the class conditional densities, because we do not know the class conditional densities the idea is that class conditional densities can be estimated if I have got IID samples from each class the idea is for each class I have taken the IID samples.

(Refer Slide Time: 52:54)



- In parametric method, we assume a form for the density and estimate parameters.
- We have considered both maximum likelihood and Bayesian methods.
- We have also discussed EM algorithm for ML estimation of mixture densities.
- In non-parametric methods we do not know form of density.
- We considered Kernel density estimates.
- We also saw k-nearest neighbour estimate

So, now the problem reduces to given some density f, which is unknown, but I have n i i d sample from the density can I estimate the density, so the idea with given the density estimates we can implement Bayes classifier and as we said we can estimate densities either parametrically or non-parametrically. So, we have looked at both methods in parametric method we assumed a form for the density and estimate the parameters that is we assumed density to be let us say Gaussian, but we do not know the mean and covariance matrix we have assumed the density to be in the discrete case say geometric, but we do not know the parameter p and so on.

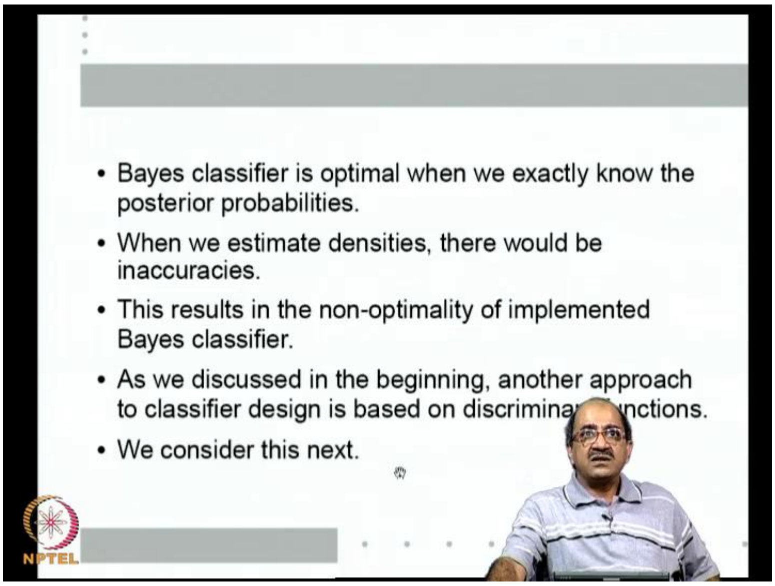
Then the idea is given the IID samples you estimate the parameters we looked at two methods of parametric estimation maximum likelihood and Bayesian methods. We have seen that both of them are give us concession to estimates and essentially Bayesian methods allow us to use any extra information. We have about the unknown parameters at the cost of little more complicated analysis for deriving the estimates using this, so called conjugate priors also they allow us to get better estimates at small sample size at large sample size both maximum likelihood and Bayesian methods are same.

We have also considered estimation we considered maximum likelihood and Bayesian estimation techniques for all the standard densities essentially at the exponential family of densities. And we also looked at mixture density models, which are more general, which can model, which can capture multi model data distributions and we have looked at an e m algorithm a specialised algorithm for m l estimation of any mixture density model.

Then we also looked at the nonparametric method, where we do not know the form of the density we are not willing to assume any form for the density, but we still need a density estimate and in last class and this class we have seen some ways of looking at such density estimate mainly a the so called Kernel density estimates. And we have also looked at the nearest neighbour density estimates and seen their relationship with k nearest neighbour.

So, this kind of completes one aspect of our pattern recognition journey, so we started with this statistical pattern recognition model whereby we said the variability is in feature vectors belonging to the same class are modelled as densities. Then we derived the Bayes classifier we have seen how we can rate different classifiers using minimisation of risk through a los function. And we showed that if we know the class conditional densities and prior Bayes classifier will give you the minimum risk and then we also looked at a few other classifier structures.

(Refer Slide Time: 55:51)



- Bayes classifier is optimal when we exactly know the posterior probabilities.
- When we estimate densities, there would be inaccuracies.
- This results in the non-optimality of implemented Bayes classifier.
- As we discussed in the beginning, another approach to classifier design is based on discrimina nctions.
- We consider this next.

Then we comeback to implementing Bayes classifiers and for that you how to use the training samples for estimating the density functions. So, with the estimating density functions we can now the, with estimating density functions we can implement the Bayes classifier. The issue with Bayes classifier is (( )) optimal only when we exactly know the posterior probabilities when we estimate densities of course, there will be inaccuracies and, because of the inaccuracies the implemented Bayes classifier will be non-optimal.

So, even though the Bayes classifier is best any implemented Bayes classifier will not be optimal anyway, because there will be inaccuracies in density estimates. So, as we have discussed now in one of our beginning lectures the second lecture there are other approaches to classifier design, which we called for example; the one based on discriminant functions. So, beginning next class we will look at the discriminant function based approach, so we kind of now completed the implementation of Bayes classifier and all in the discussion based on that. And now we move on to other ways of looking it, so we first look at implementing linear classifier essentially, linear discriminant function. Based classifiers and the associated regression function estimates both for linear models first and once again using a risk minimisation statistic that is what we will do starting from next class.

Thank you.