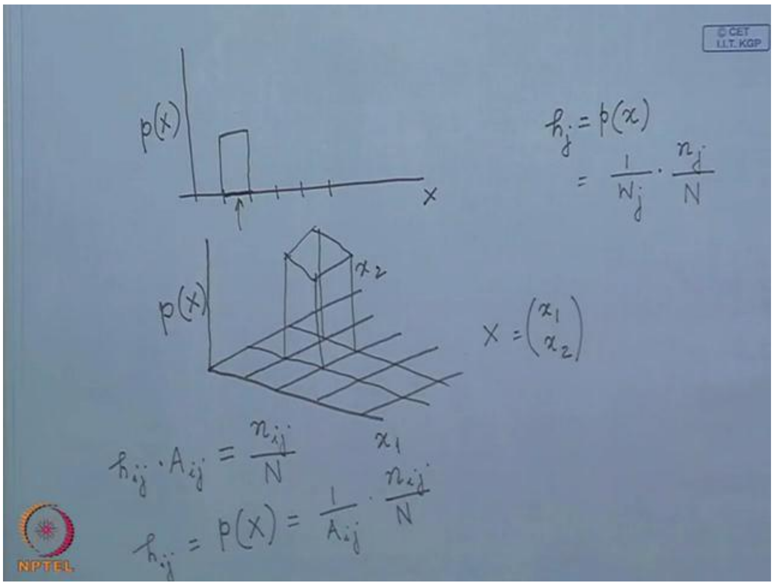
Pattern Recognition and Application
Prof. P. K. Biswas
Department of Electrical and Electronics Communication Engineering
Indian Institute of Technology Kharagpur

Lecture - 14
Probability Density Estimation (Contd.)

Hello, so we were discussing about the histogram technique for the probability density estimation. We have considered different cases in one dimension, in two dimensions and we started our discussion in multi dimension and the dimension is more than two. So, we have seen that in one dimensional case, what we have is a probability density curve, in two dimensional case, what we have is a probability density surface.

And in both the cases our constant was that, in case of one dimension the total area under the probability density curve has to be equal to 1. In case of two dimension the total volume under the probability density surface that has to be equal to 1 and accordingly we have computed that what is the probability density value on every beam whether it is in one dimension density or in two dimensions. So, in case of one dimension beams are actually the line segments like this.

(Refer Slide Time: 01:33)



So, if I have variable X, and on this axis I plot the probability density function p X, then what we have done is the X axis is divided into number of beams and number of cells. And within a cell we have assumed that the probability function is constant, so it is something like this, where the height of this bar which is nothing but the value of the probability density within this cell, that is determined by the fraction of the training samples which falls under the beam and simultaneously what is the width of this particular beam.

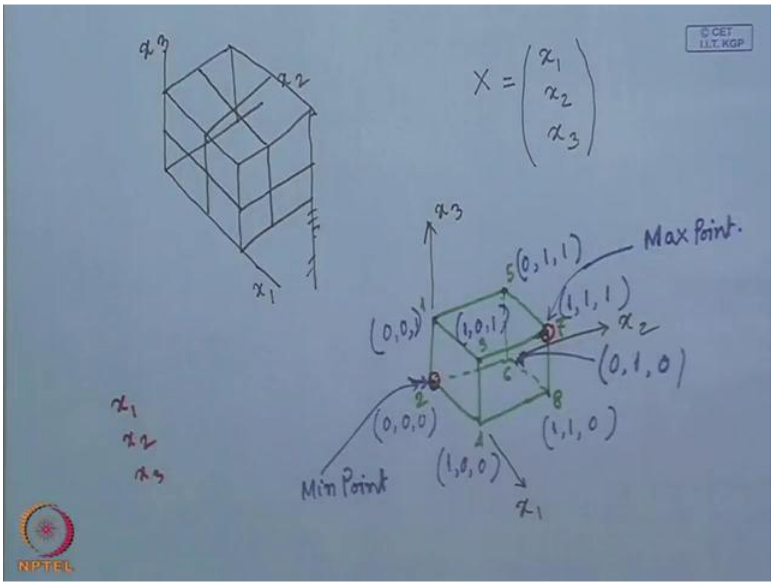
So, this h i or the h j for j th beam which is nothing but p of X, where X lies within this j th beam which is given by 1 upon w j or w j is the width of the j th beam times in j by capital N, at this n j is the number of training samples, which falls under the j th beam and capital N is the total number of the training samples. So, this is what we had in one dimensional case or when a variable x is a scalar variable.

In two dimensional cases, we have extended this concept, so we have vector variable X whose components are x 1 and x 2, so I have this plane x 1 and x 2 where this is the x 1 axis, this is the x 2 axis and axis perpendicular to this, this indicates what is the probability x, where x is a vector. So, following the same concept we defended this plane x 1 and x 2 plane into a number of beams like this where each of this beams is now a rectangle. The probability density value over a beam will be represented by bar something like this, where the bar is having some volume is given by if this beam i j th beam.

So, h i j times A i j or A i j this is the area of this beam and you say that this has to be equal to n i j upon N, where n i j is the number of samples which falls under this i j th beam and N is the total number of samples. So, from this we can compute h i j which is nothing but p of X where this X is a vector. So, what is the probability that this vector x which falls under the i j th beam, so which is given by 1 upon A i j into n i j upon capital N and here again it is assume that is within a beam the probability density is constant, but the probability value varies.

So, we have extended this into three dimensions or obviously in case of one dimension as the beams are line segments. In case of two dimensions the beams are area segments which are nothing but rectangle. In case of three dimensions the beam will be volume segments, so it is a cuboids or parallelograms thing like that.

(Refer Slide Time: 06:06)



So, in the last class we have taken a particular scenario, so my vectors x will be three dimensional vector. So, vector X is equal to x 1, x 2, x 3, so it has got three components, so accordingly I said this is my vector x 1 axis this is my x 2 axis and this is my x 3 axis. So, accordingly I define a three dimensional space and every vector is appoint in that three dimensional space. So, to estimate probability density function what you have to is this three dimensional space has to be divide into a number of volume elements which are nothing but the beams.

So, in the last class we had taken volume element something like this, so here you find out that there are eight volume elements, four on this side and four on this side. Once I divide this three dimensional space into such volume elements every volume element will have eight vertices. So, if I consider a single volume out of this, so these volume elements have got eight different vertices the vertices are one, two, three, four, five, six, seven and eight. So, these eight vertices uniquely defined this particular volume element.

Now, we take an advantage is that because these volume element is are rectangle parallelogram, I need not specify all the eight different vertices rather if I specify only two vertices then also my volume element is completely defined. So, one of them if I say that this is my X axis or x 1 axis and this is x 2 axis and this is x 3 axis considering this, you find that if I specify the coordinates of these volume element.

That is coordinates in these volume element that uniquely specify particular rectangular parallelogram, where the coordinates of this vertices the x 1 component will be minimum of x 1 component of all the eight volume. Elements of all the eight vertices x 2 component will be the minimum of x 2 component of all these vertices, x 3 component will also be the minimum of x 3 components of all these vertices. Similarly, for these vertices you will find that x 1 component is the maximum of x 1 component of all the eight vertices. x 2 component will also be maximum of components of x 2 vertices of all the eight vertices, x 3 component will also be the maximum of x 3 component of all the eight vertices.

To make it more clear, let us assume that this is the unique cube and if you find unique cube and this vertices at the origin the coordinates of vertices will be 0 0 0 coordinates of this vertices will be 0, 0, 1. These vertices will be 1 x 2 components will be 0, x 3 component will be 0 for this vertices both x 1 and x 2 components will be 1 where as x 3 component will be 0. So, it is 1, 1, 0 for this vertices, all of them will be 1 for this component, I will have x 2 component 1, x 2 component 1, x 3 component will be 1 component is 0. So, this becomes 0, 1, 1 for this one we have already done and this which hidden it will be x 1 is 0, x 2 is 1 and x 3 is again 0.

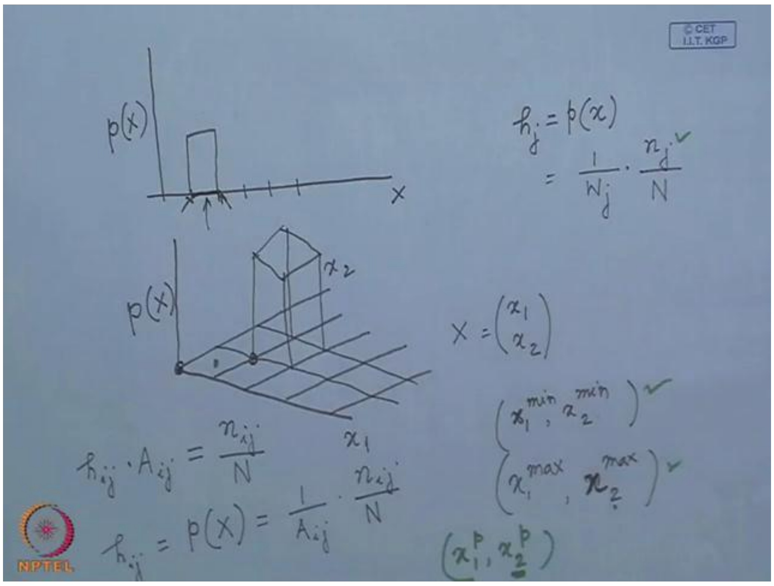
So, you find that I have got this one is left, so this one will have x 1 component is 1 and x three component is 1 and x 2 component is 0, so it is 1, 0, 1. So, I have one, two, three, four, five, six, seven, eight, so all the vertices. Now, if you look at the coordinates of all the eight vertices, you find that considering the x 1 component the x 1 component of all the eight vertices are either 0 or 1. So, the minimum is 0, x 3 component minimum is 0, so 0, 0, 0, is one of the vertices which is required to define this volume element and that is what is this one. Similarly, the maximum of all the x 1 component is equal to 1 maximum is also x 2 is also 1 and the maximum of all the x 3 component that is also 1.

So, it is 1, 1, 1 which is another vertex which is necessary for defining this particular rectangular parallelogram which is this one. So, if I simply know these vertices and these vertices and these vertices I can form this parallelogram and because this is the minimum of all the coordinates. So, I call this as min point and this beam maximum all the coordinates, so I call this as max point. So, one thing is cleared that if I know the min point and max point of every volume element then my volume element is unique identified.

Now, what is the use of this min point and max point we said before whether it is in one dimension or two dimensional cases that to estimates the probability density function in a cell or beam. So, in three dimensional case each such volume elements of the cells of the beams. So, to estimate the probability density function in each cell or in each beam I need to find out how many vectors how many points or how many training vectors are actually falling in the beam.

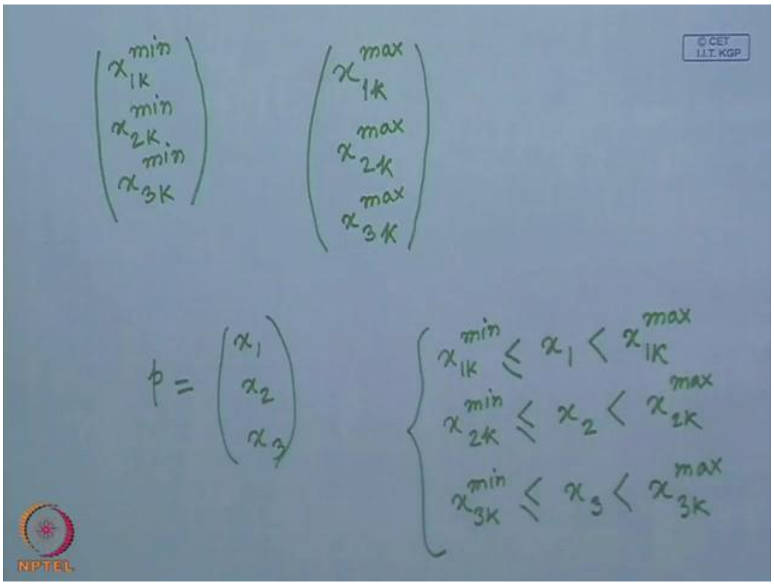
So, in case of one dimension you find that I know bounded the minimum bounded at the line segment and maximum bounded of this line segment. So, value are fixed which is greater than or equal than to this or this falls under this beam. That is my check how do I found out that, how many ten samples are falling in the j th beam you come to two dimensional case every point here also I have say for example this beam.

(Refer Slide Time: 15:04)



So, this will be the minimum point or maximum point, so this min point has coordinate in it x 1 min and x 2 minimum. So, this is the coordinate this point and coordinate of this point is x 1 max and x 2 max this is the coordinate of this point for any point p this lying within this cell. I must have say point p who is coordinate is x 1 p and x 2 p condition must be that x 1 p has to be greater than or equal to x 1 min or it has to be less than x 1 max, simultaneously x 2 p has to be greater than or equal to x to min, it has to be less than x 2 max, so if I have simply have the min point and max point I can easily determine that what are the samples which falls under which beam and I can count the number at such samples. So, the number of sample falling under the j th beam, in case of one dimension will be n j and the number of samples falling under i jth beam. In case two dimensional will be n i j, so in three dimension as we have defined this min point and max point here also find that for the min point.

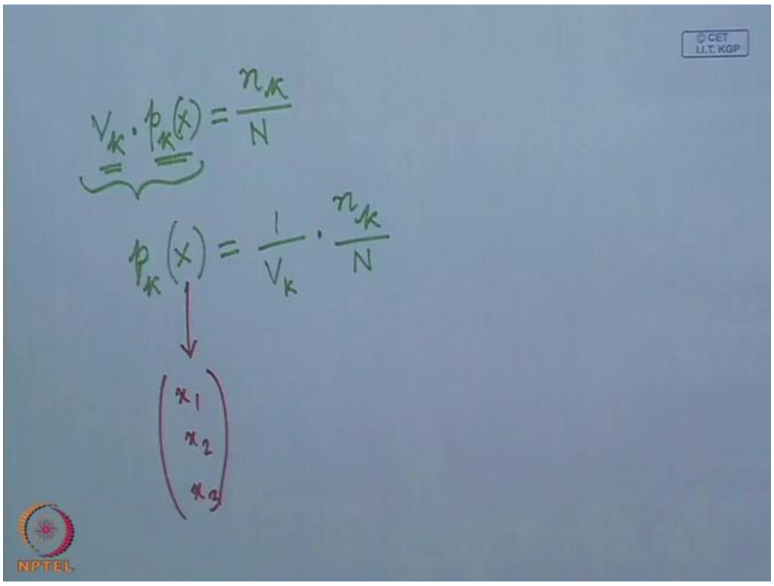
(Refer Slide Time: 16:52)



I have the coordinates, let us assume that x 1 min x 2 min and x 3 min, these are the min points of this, this is the min point of say k th beam. So, I put a x 1 k, x 2 k and x 3 k, so this is the min point k th beam. Similarly, max point I defined as x 1 max as it is for the k th beam, x 2 max k and x 3 max k, so this is the max point of the k th beam. Now, any point p at x 1 x 2 x 3 if it it has to lie within this k th beam in three dimension, I must have the condition that x 1 k min must be less than or equal to x 1 which must be less than x 1 k max.

Similarly, x 2 k min must be less than or equal to x 2 which is less than x 2 k max and x 3 k min must be less than or equal to x 3 which must be less than x 3 k max. If all these three conditions are satisfied simultaneously then only I can say that this point p falls within in the k th beam and p for whichever beam this condition. All these conditions are simultaneously satisfied the point p fall within that particular beam. So, by using this given a set of three dimensional vectors, I can find out that how many of these vectors which are falling under beam.

(Refer Slide Time: 19:43)



So, if I find that in k numbers of vectors are falling under k th beam and N is the total number of training vector then the fraction of vectors which are falling under this k th beam is given by n k by N. Now, the volume of the k th beam if that is b k and the probability density in the k th beam, if that is p k capital X i must have v k time p k X is equal to n k by n. Now, here I find that I can find analogy from one dimensional case and in two dimensional cases. In case of one dimension we said that the area under the probability density curve has to be equal to be 1, in case of two dimensional case we have said that the volume under the probability density surface is equal has to be equal to 1.

Now, following the analogy over here you find out that this is my probability density value within the k th beam, this is the volume of the k th beam. So, this beam the density and this beam the volume I can say that v k times p k represents mass, so the total mass under this probability density has to be equal to 1. So, this is a simple analogy which can be drawn from one dimension or two dimension to three dimension and later we see after some time that this can be extended to multi dimension.

So, once I have this my simple calculation is the probability density value will be in the k th beam of X, where x 1 is my three dimensional vector. So, this X is nothing but a vector x 1 x 2 x 3 which falls under the k th beam. So, the probability that a vector x will fall under the k th beam the simply given by p k x is equal to 1 upon v k times in k by capital N or v k is the volume of the k th beam, n k is the number of sample which is falling under the k th beam and N is the total number of ten samples. So, I can easily find out the probability density function in case of three dimensions.

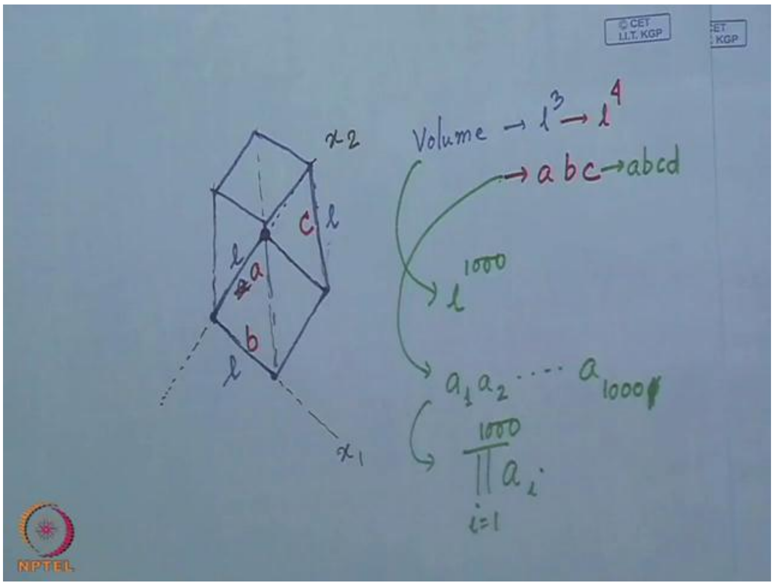
Now, let us see that how we can extend these two multi dimensional because as you have seen in earlier in our case it is not necessary that I will have only the scalar variable or vectors variables could have been two dimensional variables or three dimensional variables and so on. The dimensionality of our vectors can be much more than that because when you are dealing with the pattern recognition each vector is formed by different feature of the patterns.

I can compute one feature, I can compute two features, I can compute three features, I can compute five features. I can consider ten features, I can consider hundred features, I can consider thousand features. These depends upon what is the complex city of the that pattern we are trying to deal with, if the pattern is very simple only one feature will be sufficient in which case uni variant probability function for us. If the pattern is slightly more complicated, I may be satisfied with the three numbers of features which gives me three dimensional feature vectors.

So, probability density function in three dimension like this what you have computed that will be sufficient if the pattern is very, very complicated or I may have to hundreds of feature to represent the pattern made a thousand. So, those thousand features forming a feature vector the dimensionality of my feature vector becomes thousand. So, the space that I have to consider in which I have to estimate the probability density is neither one nor two nor three it has to be thousand dimensional spaces.

Visualization of a thousand dimensional space is not that simple may be somehow we can be visualized up to three dimension, but we cannot usually visualize a thousand dimensional space, but have do is I have to walk using that concept. So, before I go in to that let us assume, let us just look at that how I can evolve a multi dimensional space, let us consider a very simple case initially.

(Refer Slide Time: 25:11)



Suppose, I have a point, so this is the point and we all know that a point is of zero dimension A neither has the length breadth and so on. So, it has zero dimension, now if I pull this point in a certain direction along the certain lines what I get is a line segment and a line is in one dimension. So, I find that from zero dimension which was a point if I pull the line in a particular direction what I stress is a line and this line is defined in a single dimension because I know what the direction of the line is.

So, what I have is line is in one dimension because I know what is the direction of the line so what I get is and one dimensional space and a straight line is defined in one dimensional space. Now, given this if I pull the line or drag the line in a direction perpendicular to the direction of the line or in other words, I draw the line in this direction, now I get the line over here.

So, you found that I have done is a rectangle, so from one dimensional space which contains the line if I draw the line in a direction which is perpendicular to the direction of the line what I define this rectangle and this rectangle is in two dimension. So, if this was one direction dimension this is the other dimension one of the dimension I call as x 1 and the other dimension I call as x 2. So, I have a 2 dimensional space and in the two dimensional space I have rectangle or even a square which is defined in a 2 dimensional space. Now, what I can do is I can drag this square of this rectangle in a direction which is perpendicular to the plane x 1 y and x 2, so if I drag it in a direction which is perpendicular to x 2 like this.

So, this was the plane I have dragged this plane in this direction perpendicular to x 2 what I have got is this now I find that what I have actually got is a cube because if the displacement. So, the amount of drag in every direction is same that I get a cube if they are not same I get parallelogram. This parallelogram is defined in three dimensions, now come to a case that will this point was dragged by an amount say l the length of the line becomes l. If I drag this line perpendicular to the direction of the line by the same length l I get a squared, whose area is l square, if I drag this square in the direction perpendicular to the plane x 1 x 2 I get cube.

So, this is also l and I get a cube whose volume is simply l cube, now if the amount of drags are different suppose this drag is something like other this drag is a this drag is a the amount of this drag is b amount of this drag is c. What I have get this parallelogram whose volume is a b c, now let us use some other concept that I have a three dimensional space. In the three dimensional space I have a cube or a parallelogram, now I can also assume that this cube or this parallelogram is something like a plane segment.

A plane segment in a hyper plane as we say that to three dimensions, we say planes beyond three dimensions, which we cannot visualize easily at the top hyper. So, I can consider that this cube of this parallelogram is a plane planet segment in a hyper, so if I drag this cube in the direction which is orthogonal to the high per plane what I have is I have another parallelogram in four dimensions, so this is something conceptual. So, I have another cube what I another parallelogram in four dimension and if that drag the same as that drags which you have done given before. That is how drag it the same displacement l the volume of that hyper cube in four dimension will be into the power four.

So, what we have done over here that in three dimensions the volume of the cube will incur in four dimensions this will l to power of four if it is high per cube four dimension or else if I drag it by displacement d along the fourth dimension. I get hyper parallelogram in four dimensions and the volume of the hyper parallelogram will be simply a b c which was the volume of the parallelogram in three dimensions in all dimensions it will be a b c d. So, you find out the concept is very simple and I always have analogy to all dimensions, I may not stop there I can consider this is these high per cube dimension is actually a plane segment in hyper plane in four dimension.

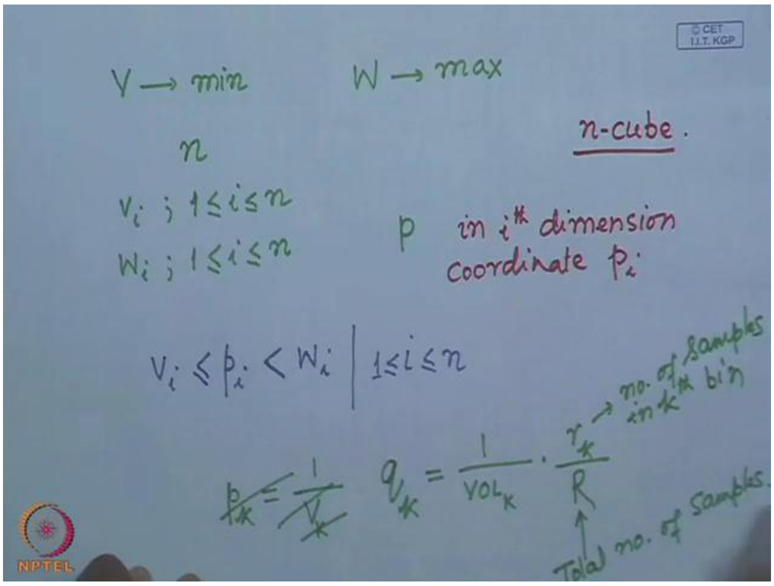
So, if I drag this four dimensional planner segments in a direction which is orthogonal to that hyper time I get another parallelogram in fifth dimension. If that amount of drag is same as l the volume of that hyper cube in five dimensions will be instead of l to the power of 4 it will be l to the power of 5. If the amount drag is e in the fifth dimension the volume of the hyper parallelogram that I get in find the dimension will be a b c d times e where e is the amount of drag in the fifth dimension.

This concept I can go on it is extending in multiple dimension, so if I have a feature space of dimension say 1,000, so volume of a hyper cube in 1,000 dimensional spaces will be given by l to the power 1,000 as 1,000 is the dimensional of the space. This is a cube in 1,000 dimensional spaces or every side of that hyper cube is equal to l considering from here. If a 1 is the length of the hyper parallelogram in first dimension a 2 is the length of that hyper parallelogram in the second dimension like this a 1,000 is the length of that hyper parallelogram in 1,000 th dimension.

Then, the volume of the high per parallelogram will be a 1 into a 2 into a three up to a 1,000, so I can simply represent this as prod at a i i is equal to 1, 2 and 100, so this will be the volume of the high per parallelogram in thousand dimension. Now, whatever we do as we have seen in case of two dimensional spaces or in case of three dimensional space that every beam can be uniquely identified by the location of two vertices. One of the vertices is the min point the other vertices is max point and for any point to lie within this beam its corresponding coordinate must be greater than or equal to the corresponding coordinate of the min point.

Then, the corresponding coordinate of the of the max point, so similarly in this multi dimensional space what we are doing is this is basically either hyper cube or a hyper parallelogram in a multi dimensional space. So, to unique identify a hyper cube or a hyper parallelogram multi dimensional space if I have the coordinates of the min point and the coordinate of the max point that is sufficient. I do not need to store any other vertices the information of any other vertices of that hyper cube or hyper parallelogram, so if I have the min point and the max point that is sufficient for me, so given this now I can write the conditions.

(Refer Slide Time: 36:30)



Let us say V is the max point, sorry V is the min point and let us say w represents the max point of a volume element. This volume element is in one dimension or in two dimension or in three dimensions, four dimensions, thousand dimension ten thousand dimension do not matter. I have defined what is the min point and what is max point and suppose this point is defined will say n dimensional space it may be one it may be two it may be three it may be anything this is n dimensional space. Incidentally, let me mention that in n dimensional space a cube high per cube in n dimensional space is actually called an n cube. Similarly, a hyper parallelogram in n dimensional space will be called in hyper parallelogram, these are the terms that we used so what I have is n dimensional space.

So, accordingly this min point V will have n number of components, it will have V i for i 1 less than or equal to i or less than equal to n. So, different coordinates of the min point similarly, for the max point ill have different coordinates which is given by W i where one less than or equal to i less than or equal to n. So, once I have defined my min point and max point, now suppose I have a point p in i th dimension the coordinate of this point p will be p i so in i th dimension or in i th direction coordinate is p i, so here again i 1 to n.

As we have in dimensional space, so p is n dimensional vector as v or w each of them are dimensional vectors, so if this p has to lie within the volume element or within the beam defined by V and W i must have the condition that V i must be less than or equal to p i. The corresponding coordinate which must be less than W i and this has to be satisfied for all i within 1 to n, so V i less than or equal to p i less d i W i V being the min point and w being the max point. So, if this is satisfied for all i between 1 to n that means in every dimension the corresponding coordinate of p must be within the minimum limit and the maximum limit of the corresponding dimension of hyper cube.

So, if this is satisfied for all the i that is every coordinate of the point p, then you say that point p is contained within the hyper cube or the corresponding hyper parallelogram. So, given this, now I find that estimating probability density function or probability density value even in the histogram technique in n dimension in as simple as we have done in case of one dimension or in two dimension or in three dimension. So, what I have to do is this n dimensional space has to divide that has to divided into a number of beams or cells or every beam or cell we have dimensional. So, these will be divided into number of beams or number of cells in n dimension then when I have a set of training samples so the training vector.

So, every training vector is of dimension n, I have to compute that how many of this training vectors falls under say k th beam and where this beam is and uni dimensional and I identify with a vector is falls in the k th beam or not. Simply by using this condition that the min point of the k th beam and at the max point at the k th beam, so every the coordinate of the feature vector must be greater than or equal to min point of the k th beam the corresponding of the min point of the kt h beam. It must be less than the corresponding coordinator or the max point of the k th beam, so by this I can identify that how many of this samples falls under the k th beam.

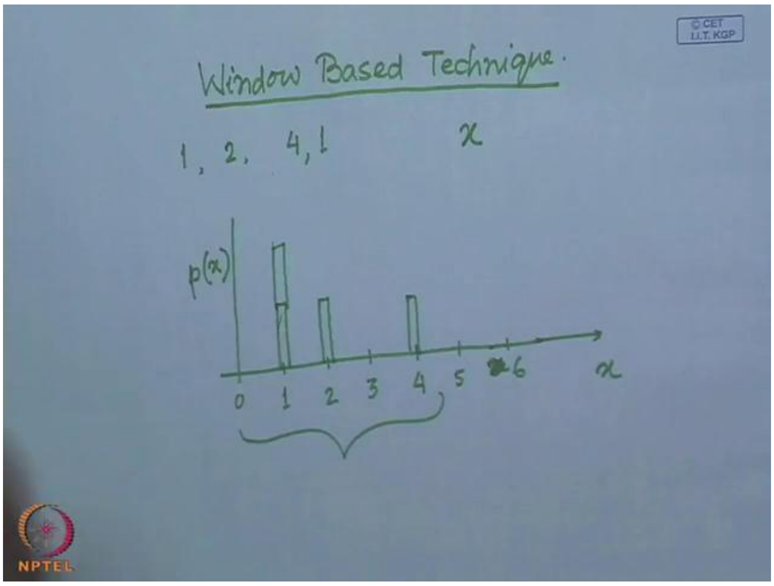
Then, the fraction of which falls under the k th beam simply k in a upon L n in case I where n k can assume that the number of samples which fall under the k th beam and capital n is the total number of samples. So, this fraction of samples falling under the k th beam divided by the volume of the k th beam are volume of the k th beam is simply either l to the power n. We are using in the dimensional case or if it is a hyper cube and l being side of the hyper cube or a, I take the product I went from one end if it is a parallelogram.

So, after all it is the fraction of the sample which falls under the k th beam divided by the volume of the k th beam that gives me the probability density estimate within the k th beam. So, a sample will fall within that k th beam with a probability density as we are estimated by this method, so what we have is a p k is the probability that a sample will falls under the k th beam it is simply 1 upon v k is the volume of the k th beam or w i n and j. Let me put it other way, so p instead of p let me call it q because what you are using as they represent point here.

So, q k which is the probability that a sample falls under the k th beam in simply given by 1 upon volume k which is the volume of the k th beam into say let me put it as say r k upon R or R k is the number of samples in k th beam and R. This total samples total number of samples, so that we try to find out that trying to estimate the density function in 1 dimension or in two dimension three dimension n dimension case does not matter, the concept is this.

I can use it in five dimensions, I can use it in ten dimension I can also use in hundred dimension it does not matter. So, this is what we have discussed about the histogram based technique for probability density estimation. We said initially that there are many other techniques, but the other techniques will have that we are going to discuss in this course is what is called window based technique.

(Refer Slide Time: 45:43)



So, that next technique that I will discuss now is what is window based; now what is called window based technique. In case of window based technique it is assumed that every sample is are representative of a probability density function. So, this representative function can be delta function or it may be any window function however our constant is that that total area under this density probability curved. We are going to estimate that must be equal to 1, so if I use a delta function are very simple case if I use that probability density function is a delta function. Every sample represents the probability density function which is a area function, so what is a delta function it is a spike with a various mole width and as before the area.

Under this delta function will be the width multiplied by the height of that spike, so if I situation something like this that I have say let us actually will go with three samples 1 2 and say 4. So, these are the three sample that we have obviously in this case we have assuming that we have a scalar variable x and these are three samples scalar of this variables.

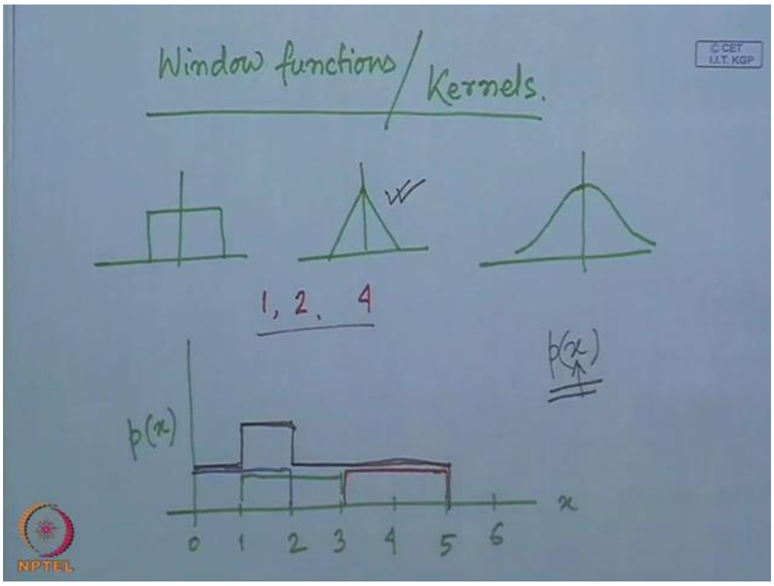
So, if I use, so this is my x and this direction I plot probability of x, so I have a sample this is 0. This is 1, this is 2, somewhere here it is 3 and this is 4, this is 5, this is 6 and so on and I said that every sample is represented by the probability density function. So, initially that assume the representative is probability density function is a delta function something like this, so I have one delta function at location x is equal to 1 I have another delta function at location x is equal to 2. I have another delta function at location x equal to 3 and I said that a constant is total area under this probability density function has to be equal to 1. So, that clearly says that area of delta function is to be 1 by three area of this delta function has to be 1 by 3 and area of this delta has to be 1 by 3.

Now, if I have any situation something like this that my samples are 1 2 4 and I have an 1 more sample at location 1, then obviously there are two delta functions which will be super imposed that location is equal to 1. So, this height of the delta function will be just a double and not only that area of individual delta functions has to be one-fourth in the earlier case I had only these samples. So, the area of individual delta function was equal to one-third, so that when I add all of them all the delta functions together the total area becomes 1, but now I am considering a case where I have four delta functions because I have two samples line at location at 1.

So, I have four delta functions and the total area has to be equal to 1 so the area and under every delta function has to be 1 by 4, so I have a situation something like this area of the delta function at x equal to 1 has to be doubled of the of the area. The delta function at location x equal 2 or area of the delta function at location x is equal to 4 total area has to be equal to 1. Now, if I have this sort of situations, then you find that this is not really approximating continues probability density function because then what I will is I have a set of spikes at set of described spikes.

So, I am not really estimating a continues probability density function, but what I have to do is I have to use the continues I have to estimate continues probability function from the rate of sample. So, the delta function to be a representative probability density function is not really suitable instead of delta function. I can use some other function that will be in the function of the kernels, which can help us estimate the probability density function in a smooth way.

(Refer Slide Time: 51:27)



So, what we will use is window functions or which are also called kernels of course they have to be properly normalized so that the total area remains to be 1. So, this window function are kernels can be different type kinds will be I can use a various rectangular window I can use angular window. So, the window if I use a rectangular window the window will be something like this which has made in box function if I use a triangular window the window will be something like this. Even I can also use a normalized normal distribution, so the window will be something like this, so different type window function is. Different types of kernels can be used to estimate the probability this function, so find that over here when I have a situation that I have three samples at location 1, 2 and 4.

The situation was something like this, so if I use the simpler case this rectangular window then at 1 I will have rectangular window. Suppose, the width of the window is let us assume two at place at two places over here at location again I have to put a rectangle. Here, whose width again equal, these will be something like this and location four, again I will put a rectangular window whose width is again equal to 2. So, I will have a window something like this, so the over all probability estimate is basically window function and I have to normalize these windows in such a way that because I have three different windows.

So, the total area under this has to be equal to 1 and by using this you find that over all probability density function is of all the density estimate will be something like this. So, this curve gives me what are the probability density estimates of p of x, so now what we have done is something visually. In the next class, we will talk about if I use instead of rectangular windows say triangular window or some other window, how can we estimate the probability density function. We will also find analytically that how we can estimate given these defined samples 1 to 4 or something like this.

These defined samples and given the kind of window function that you are going to use that how we can estimate that probability density at an unknown location x. So, I want to estimate this p x where x is not in any of this samples which are given for training. So, these samples are given your given an window function which has to be use to estimate the probability density now at an unknown x.

So, I want to estimate this P X where X is not any of this samples which are given for training, so these samples are given you are given a window function which has to be used to estimate the probability density. Now, at an unknown x i want to estimate what will be P of X, so how we can do it by using the window function, so that part we will consider in our next class.

Thank you.