Indian Institute of Technology Madras
Presents

NPTEL
NATIONAL PROGRAMME ON TECHNOLOGY ENHANCED LEARNING

Pattern Recognition

Module 02

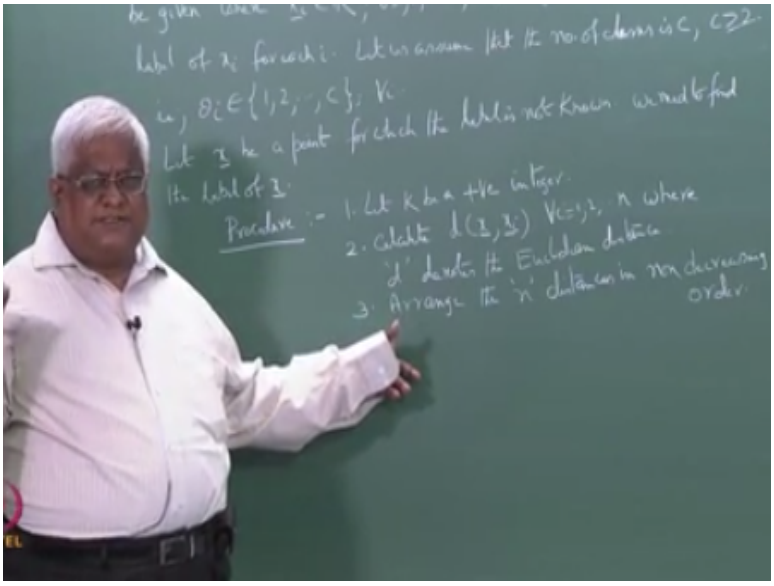
Lecture 12

K-NN Classifier

Prof. C.A. Murthy
Machine Intelligence Unit. Indian Statistical Institute. Kolkata

I shall talk about K nearest labour rule now.

(Refer Slide Time: 00:22)



This is a procedure for supervised classification. Here Xi θi, i=1 to n, they are given to us Xis are the points, they belong to M dimensional Euclidian space, θi denotes the label of Xi f or each i. Label what I mean is the class from which the observation Xi has come, the class from which the observation Xi has come. So let us assume that the number of classes is C, C number of classes, where C is naturally greater than or equal to 2.

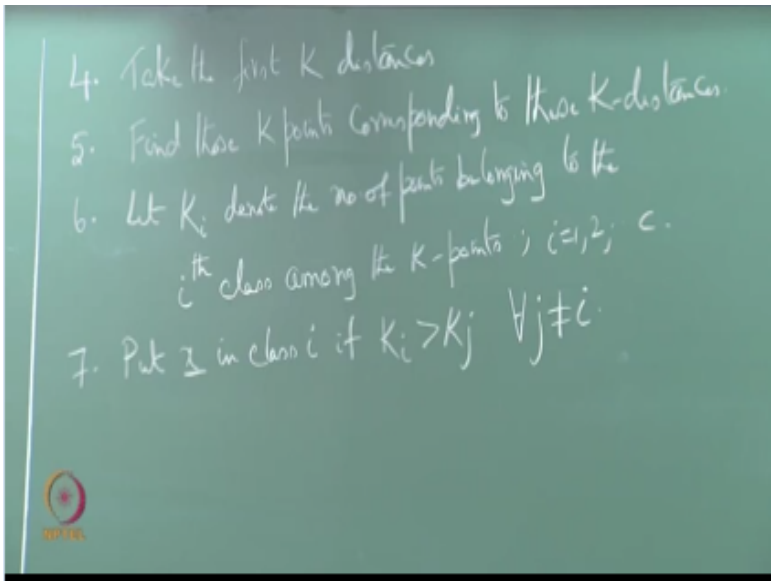
Where C is naturally an integer, and it is greater than or equal to 2. That means that is each θi, it can take values from 1 to up to C for all i. For all i, each θi, for all i, θi can take values from this set, 1, 2 up to C. θi=I means Xi observation has come from class 1 to Xi observation has come from class 2 etc… And the problem is that, let X be a point for which the label is not known, that is the class 2 which X should, X belongs to that label is not known.

So how to get the label from X1, X2, Xn… that is the basic problem. The procedure is the following, we need to find the label of X. This is the basic issue right, now what is the procedure, the procedure is the following. The first step is let K be a positive integer. Note that we are talking about K nearest neighbor decision rule. So we are going to take a value of K, let K be a positive integer.

How to choose the value of K, we will come to it later okay. What we will do is that, from this point X we calculate distances, X to X1 the distance X to X2 the distance, X to X3 the distance and up to X to Xn, we calculate N distances. Calculate d(x, xi) for all i=1 to n, where 'd' denotes the Euclidian distance. So we calculate distance from X and X1, X and X2, X and Xn, Xn, so we have in total N number of distances.

And the third step is arrange these distances in increasing order or to be precise non decreasing order okay. Arrange these N distances in non decreasing, there are N distances arranged in non decreasing order.

(Refer Slide Time: 07:59)



Take the 1st k distance that is if k = 1 take the least distance if k = 2 if you take the 1st two distances so in general for any k if yiu take the 1st k distances and find those points for which these 2st k distances occur okay find those k points corresponding to these k distances find those k points corresponding to this k distances 6, let k I denote the so we have k points from this k points fine those points belonging to the class 1 the number of such points such points belong to the class 1 we call class k1 the number of points belong to the class 2 call it as k2 similarly the number of points belong to class c call it as kc so I ranges from 1 to c it may so happen that some class may not have any point for example let us just say c = 10. 10 classes are there.

And we have taken the value of k to 1 we have taken let us say we have take the value of k to be = 1 then we have n distance where in them increasing order non decreasing order and you find the 1st distance so you will have exactly 1 point let us just say this point belongs to class 2 that means for k2 the value will be 1 for k1 the value will be 0 k3 the value will be 0 dot up to k10 the value will be 0.

Expect for K2 for all the other class the value should be 0 so it will not necessarily true that ki is a positive value it can 0 also so k I denote the number of points belong to the i-th class among the k points among these k points now the root is put x in class I if ki is > kj for all j ≠ i this is the rule that means the class for which number of members is maximum among this k put the point x into that class so this is rule.

(Refer Slide Time: 12:43)

So let me give you a graphical example in the graphical example let us just say these 3 points they belong to 1 class these 4 points they belong to 1 class these 2 points they belong to 1 class and let us say this is the point that should be classified so in class 1 if I call this class cross as cross class1 so they are from class 1 there are 3 points from class 2 they are 4 points from class 3 there 2 points.

So totally 9 points 2+ 3, 5 + 4, 9, 0 points are there so smaller value is 9 small n value is 9 so if I apply this rule let me just say the value k is = 3 let us just say k = 3 then what will happen you have to find the 9 distances and array in the increasing order probably the last one is probably this and the 2nd distance is probably this and probably the 3rd distance is this is so from k1 from this class we have 2 points from k2 this class there this 1 point k3 from this class there is 0 points.

So k3 is 0 k2 is 1 k1 is 2 so highest is happening for the class1 so put this point in class 1 this is basically the rule so here there are probably you are going to have very many droughts let me first tell that droughts that you may be having let me first tell them 1 by 1 the first dough is how to choose the value of k that is the 1st doughty

And there is a 2nd dough what will happen if there is equality here what will happen if there is an equality here equality in the sense that supposing I have taken here k = 4 and then my 4th one is let us just say this one the from class 1 there are 2 representatives from class 2 also there are also

2 representatives and from class 3 there is no representatives then to which class I should put this point where I have to put it in class1 how do I put in class2 that is another doubt.

Now let me tell you few more doubts is it necessarily true that for different values of k you will get the same result is it necessarily true that for different values of k you will get the same result the answer is no, the answer to this question is no it is not necessarily true that for different values of k you will let the same are different values of k you may get different results then the next question is different values of k if you get different results.

Then how to choose k this is I mean how do I say that some particular case better than the other one how do I say that some particular value of k is better than the other value of k and you have a much more fundamental question what is the theoretical justification of this root just because some people have given this rule though I need to follow it what is the theoretical justification of this rule first let me tell you the history the history is that around 1950s fix and hard just they had come out this rule.

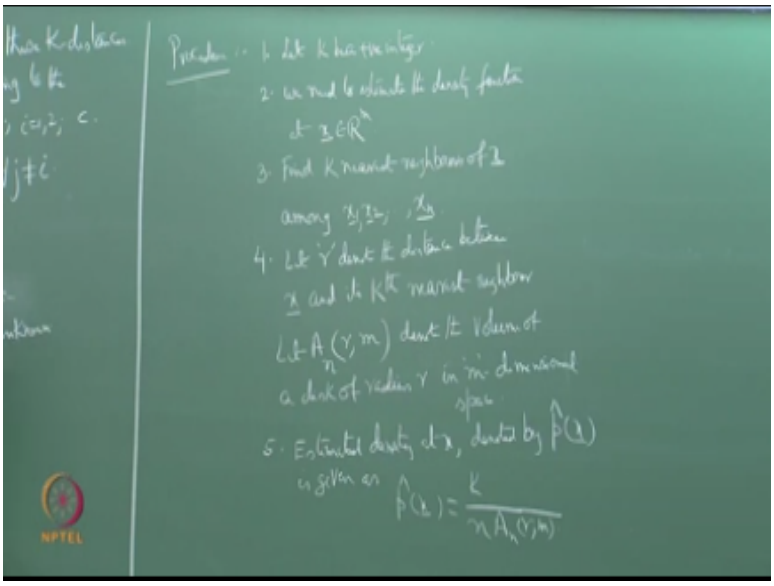
K nearest able decision rule we should go through the papers you will find this was one of the earliest differences fix and hard just in 1950s early 50s they wrote the first paper un case decision rule when they gave this rule there was no theory and biggest apply it on few data sets and they are found that the rule is working well and then that problem was written the problem still remain how to choose the value of k the problem still remain how to choose the value of k the problem still remain.

How to choose the value of k and it can be easily found that for different values of k you will get different results okay that can be very easily found that for different values of k you will get different results, now in 1965 the Loft's garden in the year 1965 that paper is published in AMS of mathematical statistics AMS of mathematical statistics 1965 Loft's garden he wrote a paper on k nearest neighbor density estimation procedure I have to tell you this thing, so let me actually tell you what this procedure is.

The procedure is that you have there is a probability the density here is the probability density estimation okay, so we have x1 x2 xn independent and identical distributed random vectors independent in radical distributed random vectors they follow a probability density function P

where P is unknown P is the probability density function that P is unknown where P is unknown, so now the question is on the basis of this n points how do you estimate the probability density function well the procedure is the following the procedure is the following so let me write down the procedure.

(Refer Slide Time: 20:59)



Procedure to estimate P okay I will write on the procedure the procedure is again the first step is as it is let KB a positive integer let KB a positive integer to let us say at that point x you would like to estimate the density function let us say we need to estimate the density function at x I am assuming that all these points they belonging to they are belonging to RM so I am not going to write one second so all these exercise they belong to RM okay so this x also belongs to RM we need to estimate the density function at the point x belonging to RM.

Okay so what do we do so what we do is that find k nearest neighbors find k in nearest neighbors of x among x1 to xn find k nearest neighbors of x among x1 to xn well a we have found them I hope you know the meaning of k nearest neighbors here that is you calculate the distance of x to x1 distance between x and x1distance between x and x2 and up to distance between x1 and xn and the distance function is always Euclidian the distance function is always Euclidian for example in the k nearest.

Neighbor decision rule here I wrote that it is Euclidian so there must be question if it is non Euclidian then what are you going to do many people apply k nearest label decision rule where

the decision distance function they take as non Euclidian some other metric the how do I sort of justify that all these things w are going to look at it at least in some amount of detail may not be much so when I say here k nearest neighbors of x the distance function is Euclidian only it is not a non Euclidian function the distance function, the distance function is Euclidian so you get hold of n distances again they may non decreasing order and find the first k distances.

So you are going to get K nearest neighbors okay, now what you would, what is done is that let R denote the distance between x and its kth nearest neighbor, then what we do is that you get hold of m dimensional volume. Let A(r,m) m is the dimension r is the radius and this is the function of n, n is the number of points.

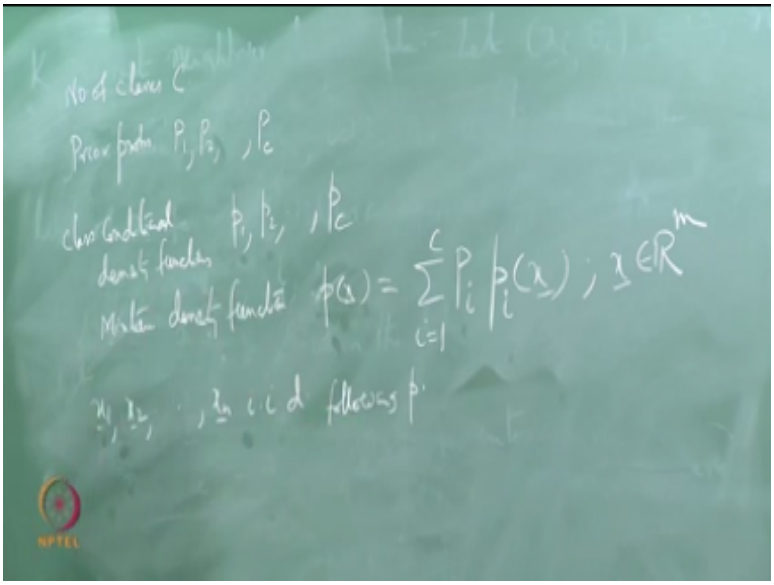
Let this denote the volume of a disk of radius r in m dimensions. Since this r is a function of n I wrote it okay, I wrote n you need not write if you want if you do not want to okay, since this r is a function of n depends on what x1 to xn we have got okay. So this denote the volume of a disk of radius r in m dimensional space and 5 is when the estimated density at x we shall denoted by $\hat{p}x$ is given as k/nA(r,m) the volume.

k/nA(r,m) now what Lauft garden had shown was that under some conditions on k this is going to be asymptotical n based and consistent what he had shown was that under some conditions on k this is going to be asymptotical n based and consistent. Now let me basically tell you the institution behind this thing, when did we probably come across the word density I think we came across the word density probably in class 8 or 9, right.

Probably in class 8 or 9 we came across the word density where we define density has mass by volume okay, note that here there is volume this is volume A and (r,m) this is volume of a disk of radius r this is volume within this volume how many points are there, there are k point out of n points so this basically looks like mass had there been more number of points probably I mean one can write that, within this volume how many points are there, k points are there.

And how did this k come k is out of n, so this looks like at least it is similar to the word density that we had used in classes I think 8 or 9 okay, this is and what Lauft Garden had shown was that it indeed goes to the probability density of the point x as n goes to ∞ , as n goes to ∞ this indeed goes to the probability density at the point x under some conditions on k, what are the conditions on k, the conditions on k are the following the first.

(Refer Slide Time: 30:24)



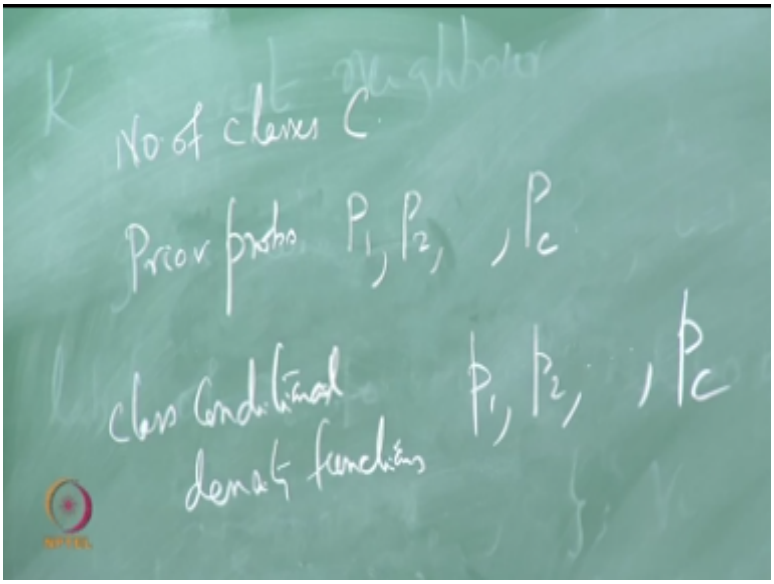
Let me just write conditions on k, the first thing is that k is a function of n, n is the number of points, k is a function n so let me denoted by we shall denoted by kn, then the conditions on k are first kn should go to ∞, as n goes to ∞ and the second one is Kn/n should go to 0 as n goes to ∞. Kn should go to ∞, as n goes to ∞ and Kn/n should go to 0 as n goes to ∞, if these two conditions are satisfied and if x is a continuity point of the density function P.

Then as n goes to ∞ this one it is an asymptotical n based and consistent estimate of Px, see if b is satisfied if the conditions in b are satisfied and x is a continuity point of p then if p^ x is an asymptotical unbiased and constant estimate of p what is the meaning of asymptotical unbiased? Asymptotical unbiased means the expected value of this one this is after all a random variable based on n number of observations, so the expected value of this that is the average as n goes to infinity this one the expected value it will go to the original one that is why it is unbiased but asymptotical as n goes to infinity this one the expected value it will go to the actual value of p.

Consistent means the difference between p^ and P as n goes to infinity the difference gets reduced that is basically consistent, so this is the density estimation process this was given in 1965 now using this density estimation procedure you can apply based decision rule fro based decision rule you have the prior probabilities and the probability density functions. So estimate prior probability is by the proportions estimate prior probability by the proportions, estimate the density functions by this.

And use based decision rule on the estimated prior probabilities and estimated density functions then you will get k nearest neighbor decision rule, I will do it now so there are c classes. Number of classes c prior of probabilities P1 P2 Pc and class conditional density functions they are p1 p2 to pc okay.

(Refer Slide Time: 35:35)



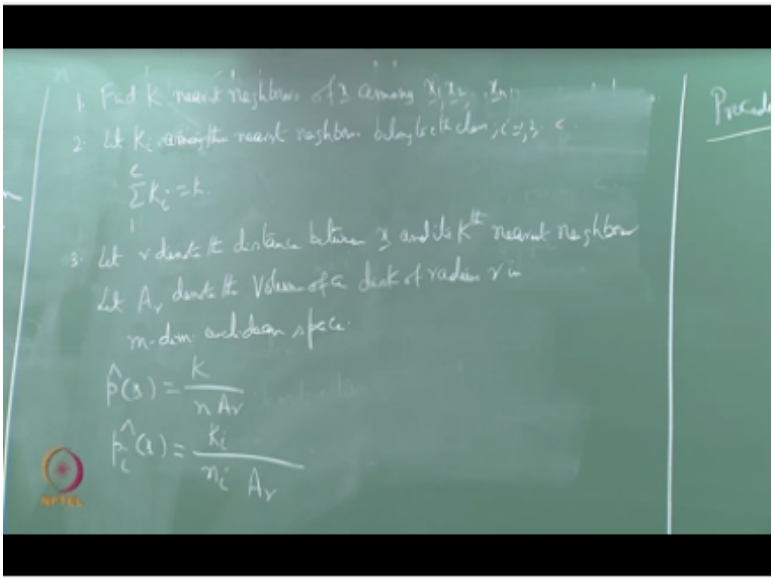
So what is the mixture density function? The mixture density function is if I write it as px that is equal to Σ I = 1 to c Pa x pax then now what are our x1 x2 xn, x1 x2 xn they are iid independent and identically distributed and what is the density function for them the density function is small p each of them they are coming from small p following independent and iid is independent and identically distributed and they are follow small p this means let me just tell you how it is done it is just an example to I mean just to think to make you understand how these things are done.

It is like this generate a random number from 0 to 1 if the values lies if the value is less than or equal to P1 then you generate point from the density function p1 if it is lies between p1 and p1 + p2 generate points from p2 if it is lies between p1 + p2 to p1+ p2 +p3 then you generate a point from p 3 then p4 p5 up to pc. So initially you need to generate a point a random number from 0 to 1 and on the base of that you decide from which distribution you need to generate a point and then you generate point from that distribution randomly.

So when you are generating a point from the distribution randomly you have got here point and you not only have the point you have the label of the point also, the point and as well as the label that is your first observation. Similarly x2 the second observation for each observation you need to generate a point you need to generate a random number from 0 to 1 and wherever it is lying the corresponding density function you need generate a point randomly.

So you are getting the label of the point and as well as the value,, values are exercise so with them you have a label also automatically given to you. X1 x2 xn are iid they are following small p let ni times out of these n points belong to class i, i = 1 to upto c that means what, that means now are estimate of the prior probability is ni/n. now let us x is point to be classified, so what do I do?

(Refer Slide Time: 40:17)



X is point to be classified, so what do I do? I find k nearest neighbor of x among x1 to xn, find k nearest neighbor of x among and let k of these ki among nearest neighbor belong to ith class i= 1,2..c, so this naturally means that ∑ 1 to c = k. Now so we know what our p^ 6 so let to find the estimate, let r denote the distance between x and it is kth nearest neighbor, let r denote the distance between x and it is kth nearest neighbor.

So now let again Ar denotes the volume of the disc of radius r in m dimensional space, so what do we know, P ^ x = k/n x Art hi is true but what is pi^ x this is going to be out of this volume Ar you have a Ki point/ ni, out of this volume Ar you have a Ki point/ ni, i= 1 upto c. now you apply base this so Pi^x let me than ≥ because = if you remember the equality part does not have the elements it is not going to give you any extra error.

So let me just maintain the equality Pi^x for all j ≠ I, what are Pi^x is ni/n and this is ki/ni x Ar ≥ nj/n and k/nj x Ar and then you cancel out everything ki > k. now a doubt about equality what will happen if I take a particular k and I get from 2 classes same number of points and that is maximum, the question is which class I have put. The answer from this thing is obvious that is the first rule that we are using, as n goes to ∞ these things happen.

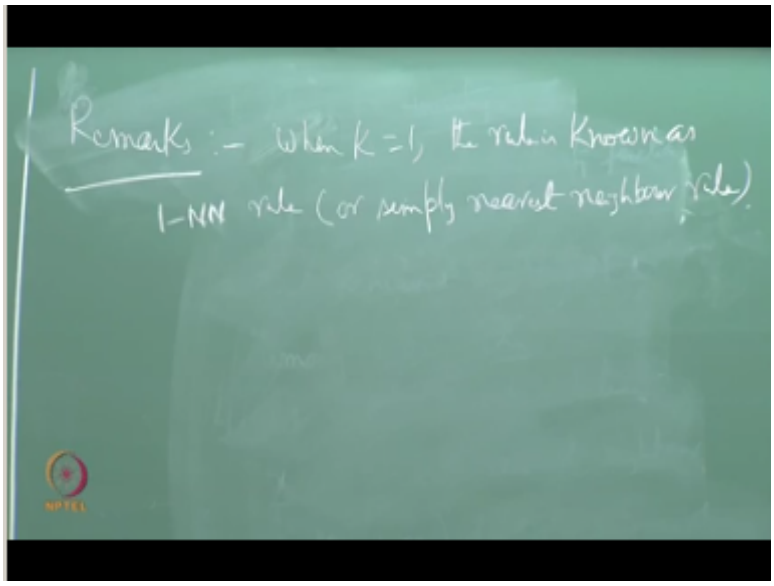
As n ∞ these things are happening and unfortunately we are always finite spaces our n never goes to ∞ our n is finite, so if there is a confusion and extreme that you are starting from smaller value then increase k/1 then the next question is fine then there is a equality then you have to modify the value of a, then the next question is how do you choose the value of A.

This question is still remaining, this question is still remaining let me tell you the meaning of this still remaining, the meaning actually I have to put in several ways one of the popular ways in which people choose the value of k now is by cross validation one of the popular ways that people choose the value of k now is by cross validation okay but then it still does not answer the question completely it is still does not answer the question completely if you choose k by cross validation then how do you know that I mean it is among the what we are doing is that if you are choosing a k by cross validation.

What exactly we are trying to say is that for this set of data this is the best that we are able to do but still it does not say then answer the next question if following the same distribution if one more point comes in then will your k help the answer to that thing is exactly not clear what cross validation and these sort of principles what they state basically is that for this set of data this is what I am trying to do I choose.

Some value for k may be tenfold 15 fold or something some value I will choose and then I do the best that I can so this is a partial answer not a complete answer if you look at the history of k and rule you will find that which have made differences over in many years one of the papers was by one of the students was saying covariance and hard is the they wrote in paper 1967 the paper actually what they say is how to calculate the error probability of error for the nearest neighbor rule what is nearest neighbor rule is.

(Refer Slide Time: 49:30)



Let me write down remarks when k=1 this rule is known as 1n rule are simply nearest rule k=1 when the rule is known as or simply nearest label rule the rule is known inn rule or simply nearest label in 1967 covariant and heart they wrote a paper where for k=1 they found the probability of error for k=1 case that they have found it to be function of the error for the based they have found it to be function of error of waste.

And in India they did a quite a lot of works on k nearest neighbor rule and several of these modifications many modifications are done on kn nearest neighbor rule so some of them you will find in wrote a book on k nearest neighbor rule I do not know whether quit is available now or not in India the author is darasdhi here, he wrote a book on k nearest neighbor rule and some people he gave some theoretical procedure for choosing.

The value of k at least in some situations that was one of the results have did some work on how to choose the value of k then there are many people who tried to take the value of k adopt that means adaptively means that for different points you will not take the same value of k for class you have many point in the test may be for one point you will choose you will have one value of k but for another point we may have another value of k.

So how to do it that they are some people who have done some work on that so there are how to choose the value of k for any data set is still not completely answered to the satisfaction of all the persons there are partial solutions available but only after that I will grow I do not want to say anything more about the partial can be compute solution or not and I stop with it.

Online Video Editing /Post Production

K.R.Mahendra Babu
Soju Francis
S.Pradeepa
S.Subash

Camera

Selvam
Robert Joseph
Karthikeyan
Ram Kumar
Ramganesh
Sathiaraj

Studio Assistants

Krishankumar
Linuselvan
Saranraj

Animations

Anushree Santhosh
Pradeep Valan .S.L

NPTEL Web &Faculty Assistance Team

Allen Jacob Dinesh
Bharathi Balaji
Deepa Venkatraman
Dianis Bertin
Gayathri
Gurumoorthi
Jason Prasad
Jayanthi
Kamala Ramakrishnan
Lakshmi Priya
Malarvizhi
Manikandasivam
Mohana Sundari
Muthu Kumaran
Naveen Kumar
Palani
Salomi
Senthil
Sridharan

Suriyakumari

Administrative Assistant

Janakiraman.K.S

Video Producers

K.R. Ravindranath
Kannan Krishnamurty

IIT Madras Production

Funded By
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in