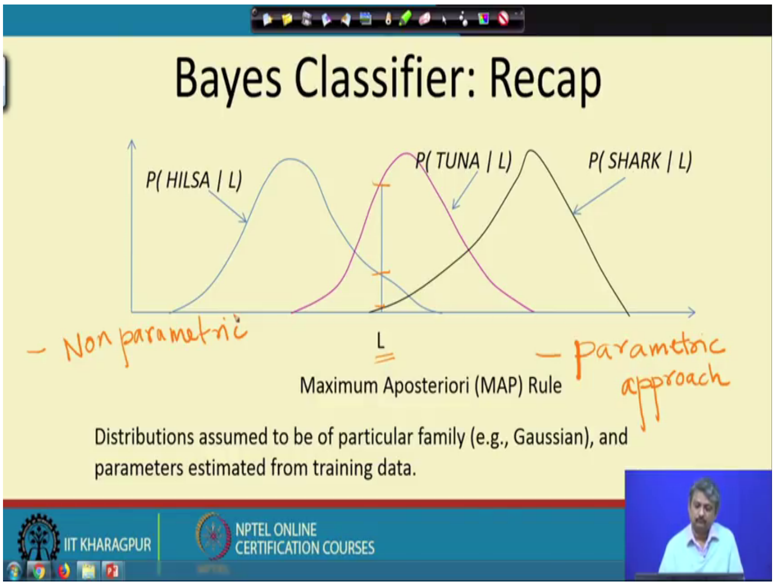
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture - 17
K - Nearest Neighbor – I

Welcome to our discussion on the classification algorithms, we have studied 2 classification algorithms so far the decision tree and the Bayesian classifier. We will next study the third classification algorithm called the K nearest neighbour classifier.

(Refer Slide Time: 00:45)



Bayes Classifier: Recap

Maximum Aposteriori (MAP) Rule

Distributions assumed to be of particular family (e.g., Gaussian), and parameters estimated from training data.

I will explain this using my previous discussion on the Bayes classifier. So, let us quickly recapitulate what we did in the Bayes classifier is that we plotted the; we sort of considered the posterior distribution, given an example of certain attribute value L belongs to certain class say tuna and for each of the classes we draw the distributions we draw the probability distributions.

And for a new point L, new whose length is L, what we did was we just considered the values of each of these distributions in this posterior probabilities which were distribution give the highest value we put it into that class and we showed actually that theoretically if we perfectly know these distributions this would give the minimum error classifier.

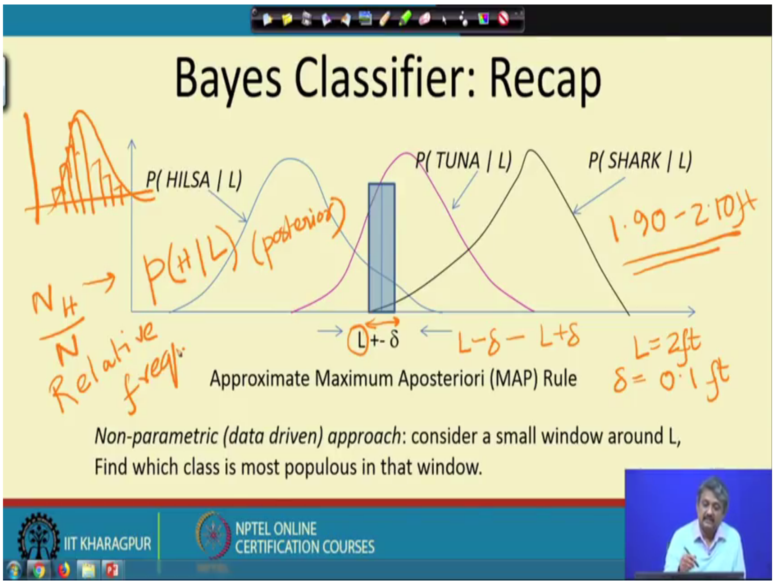
But also we observe that what happens in practice is that because of the finite sample size, we do not have infinite number of training sets, we cannot exactly know or estimate the distribution. So, we have some error in estimation we actually do not know that distributions and there are several techniques we assume them to be Gaussian, then we find out that the boundaries are linear or parabolic.

Then we assume that there is a independence between the identifying attributes and he has used the neighbours. So, this techniques of assuming Gaussian and then find from its parameter finding the decision boundary; these techniques are called parametric estimation techniques, from the data they find out the value of some parameters.

So, one critical thing about this parametrics is that you need to assume say some property of this first study of distribution. So, it is Gaussian or it has a diagonal covariance matrix, things like this you need to make assumptions whereas, the actual data might not or in the neighb Bayes independence assumption actual data might not follow these assumptions. So, there is an alternate approach which often people make known as the nonparametric approach, nonparametric. So, where we will soon see no not this example you will soon see that we need not makes any sense of assumptions.

So, we need not estimate some parameters of under this assumption and we use only the data nothing, but the data to come up with the classification. So, let me explain how to do that.

(Refer Slide Time: 04:46)



Bayes Classifier: Recap

Approximate Maximum Aposteriori (MAP) Rule

Non-parametric (data driven) approach: consider a small window around L, Find which class is most populous in that window.

So, what I do is the following, I suppose a face of length L is what I want to is, what I want to find out the class of I want to know a new face of length L which class it belongs to. What I do is that I make a small window of say a L minus delta 2, L plus delta plus minus delta small 2 delta with window I consider around L so; that means, for example, if my L is say 2 feet, 2 feet long face.

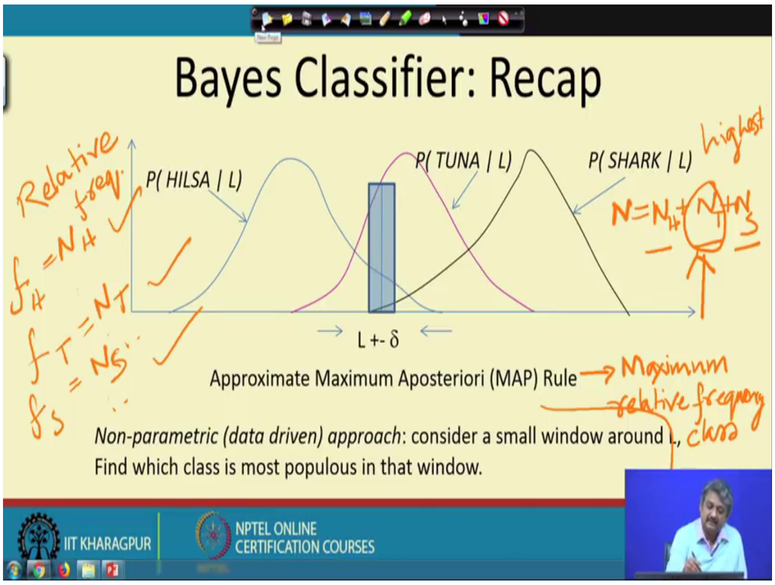
And delta is say point one feet, this consider the interval of 1.90 to 2.10 feet this interval I consider, I what I actually want to find out is that what is the class of a face of length L, 2 feet. So, now, what I do is that how many of the hill surface have a length in this inter lying in this interval ok.

So, in my training set what is the count of face having length between 1.9 to 2.10, feet let me call that count as say N H, note one thing that this is nothing, but originally when you when we draw these curves you considered histograms.

Which are, which are basically a series of bins, where we find the count of number of points in each bin and so the distribution was actually a collection of small histograms bar charts and when these bins are small enough this would look like a instead of a bar chart a small smooth function. So, the idea is that if I make my delta small enough this value N H number of which is a having length between L plus delta and minus, that sort of is a good approximation of the probability a piece having length L belongs to class the posterior probability. So, think on this a little bit. So, what I do is that I count how many phase is of course, ice made a small mistake is that it is not this count, but count by total number of piece that I am considering.

So, the something called a relative frequency, relative frequency. So, what I want to say is that if my intervals are small enough the relative frequency will give me the posterior probabilities and so what I do is that I repeat the same thing.

(Refer Slide Time: 09:28)



Bayes Classifier: Recap

Approximate Maximum Aposteriori (MAP) Rule

Non-parametric (data driven) approach: consider a small window around L, Find which class is most populous in that window.

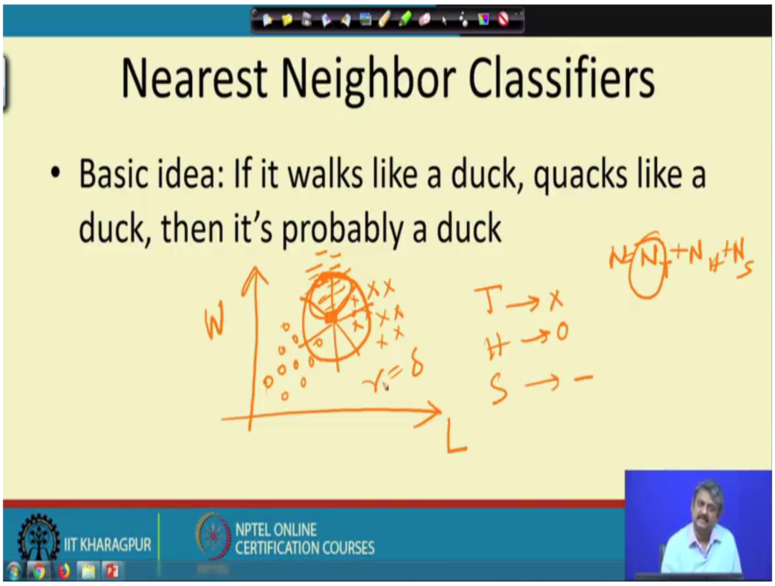
So, as I let me denote this relative frequency as f 8. So, it is NH by NN is the total training set size similarly I compute for the tuna, I compute f tuna and similarly I compute f tuna is number of tuna piece having length in that interval similarly f shark.

So, 3 relative frequencies, what the Bayes classifier, what the map rule says is that you put it in the class having the highest posterior. Since my relative frequencies are a good approximator of the posterior probabilities, I can to restate in an approximate form the map rule as, put it in the class having the highest relative frequency here that would turn out to be tuna. So, not further that I can actually do one more thing, since this denominator n is same for all these classes I can just drop it, I can just drop it still the the top class would have the highest count.

So, what I am basically doing is that for a piece of length L, I am taking an interval around it in I call it a neighbourhood around it and count how many specimen in the training set heavy length in that neighbourhood, in that interval and among them suppose that length is in I said that number is in. So, basically in N of the training N number of training examples Pauline L plus delta n minus delta and out of this N NH are hilsa NT are tuna and NS are shark and I can sort of rewrite my Mac rule as check which of these counts are highest put L into the corresponding class. Suppose NT is highest tuna is highest put it into class tuna if this is highest among these 3. So, I can do that.

So, you note that this is for 1 dimension, I could have extended it to 2 or higher dimension also how do I do that. So, let me take an example let me see if I have some if I have take something already n I do not have. So, let me delete in this suppose I have or.

(Refer Slide Time: 14:09)



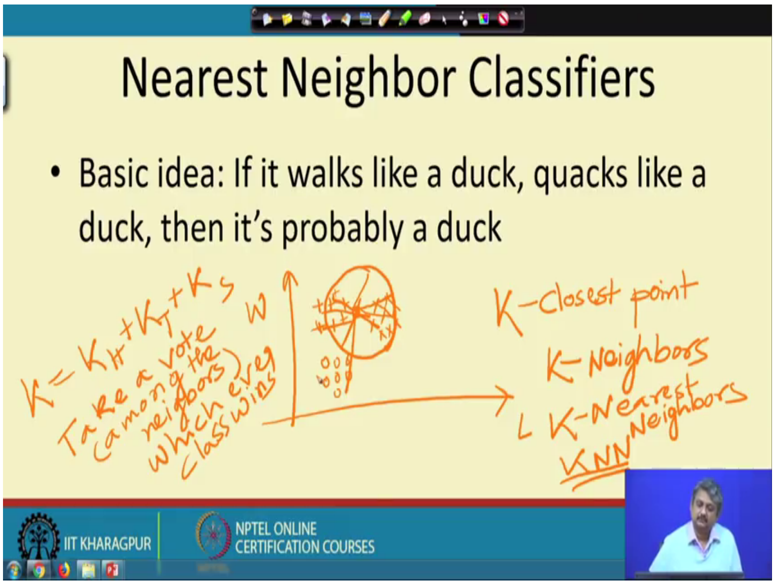
Nearest Neighbor Classifiers

- Basic idea: If it walks like a duck, quacks like a duck, then it's probably a duck

A 2 attributes a length of the piece and weight of the piece. So, every piece is a point in this 2 dimensional space. So, what I do I if I look at the training examples let me I each of the tuna maybe they will I am writing them as T or maybe write them as cross the T I represent ith cross point hilsa of ia circle point and shark maybe valued S.

So, this is all my tunas if I plot the length and weight this is all my hilsas and this is all my shark, there are some small sharks, some big sharks. So, this is like this. So, when a forget about what I have this duck example I am coming to that later. So, if I if I have a new face let me write it as a box, what I do I took an interval in 1 dimensional case instead of an interval I will take a small circle around it, it is like radius or is equivalent to the interval I take radius or in all direction delta, delta, delta say r equal to delta. R r r this in all direction I draw and make a neighbourhood with the point a lwpr some new piece that is the centre and I see among these training sets, how many of them fall in this circle suppose n number of them fall in this circle account. So, these false, this false, this false, this false, this false may be these false and see out of this n that falls how many are, tuna how many are how many are tuna. I just count that NT how many are hilsa how many are shark whatever is the highest I put it in that class, same principle same principle. So, instead of drawing these delta and all these sort of thing deciding on some delta and all these sort of thing because see it will depend on how much dimension you do and how do you choose a value of delta I can sort of have an equivalent rule as follows.

(Refer Slide Time: 17:40)



Nearest Neighbor Classifiers

- Basic idea: If it walks like a duck, quacks like a duck, then it's probably a duck

$K = K_H + K_T + K_S$
W
Take a vote (among the neighbors) whichever class wins
K-Closest point
K-Neighbors
K-Nearest Neighbors
KNN
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

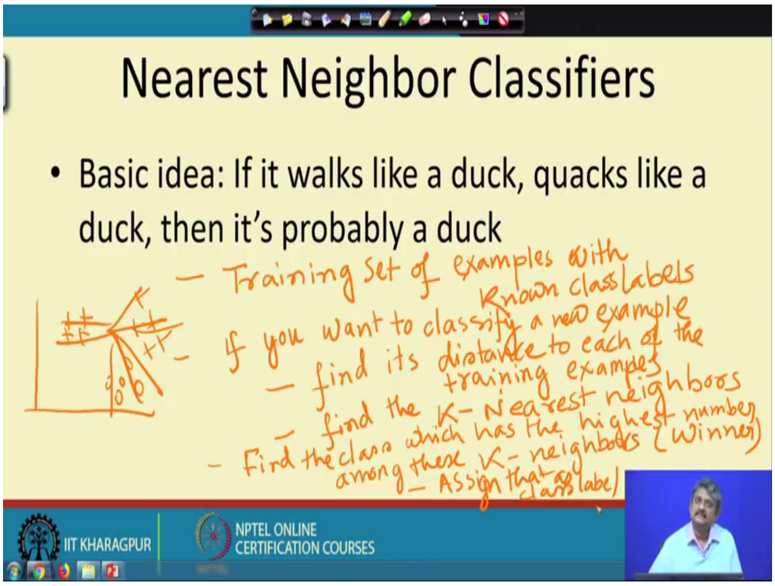
So, this is my attribute space or training space sorry feature space and I have my training points whose class levels I know and when a new point is to be classified instead of drawing a delta. I sort of do a ranking I find that among these training points if I sort them in order of how far they are from this new point in the order of how far they are from this new point I sort them the some will be closed some will far I just sort them. So, it is with respect to this new point how far these training points are and I take among these training points the top K closest points to the new point.

Which are have the shortest distance to this new point and I call this set of points which are. So, this K set up now there are K points because I am sorting them and finding the K th closest point and draw a circle which includes this K point. So, I find the so basically what I do I find the K th closest point draw a circle with this new point at centre and enclosing its K closest point smallest circle and this K points I will call as that K neighbours of this new point or sometimes the K nearest neighbours, K nearest neighbour K NN and I repeat what I did earlier.

That means, among this K I find out among its neighbours, how many are hilsa, how many are tuna and how many are shark and whichever is the highest stage. So, I can alternately take it like the think of it like this as if I am conducting and vote and election among its neighbours and whoever whichever class wins the election. That means, has number of members are there from that class among these K neighbours, I classify the new point to this winner. So, take a vote among the neighbours see how many neighbour belong to this class, how many neighbour belong to that class and whichever class wins put the new point whose class level I yet not know into this winner class.

So, for example, here maybe the cross class is the winner. So, classify this as a cross. So, this principle is called that K nearest neighbour classifier, it is a classification principle it is not like building some tree or something what it does is it does takes a training set and when a new point is to be classified it finds the distance of the new point to the training sets training set and sorts the distance take the smallest K distances and among the smallest K closest neighbour it may take support whichever class wins the both put it in that class simple. So, let me actually write down.

(Refer Slide Time: 22:39)



Nearest Neighbor Classifiers

- Basic idea: If it walks like a duck, quacks like a duck, then it's probably a duck

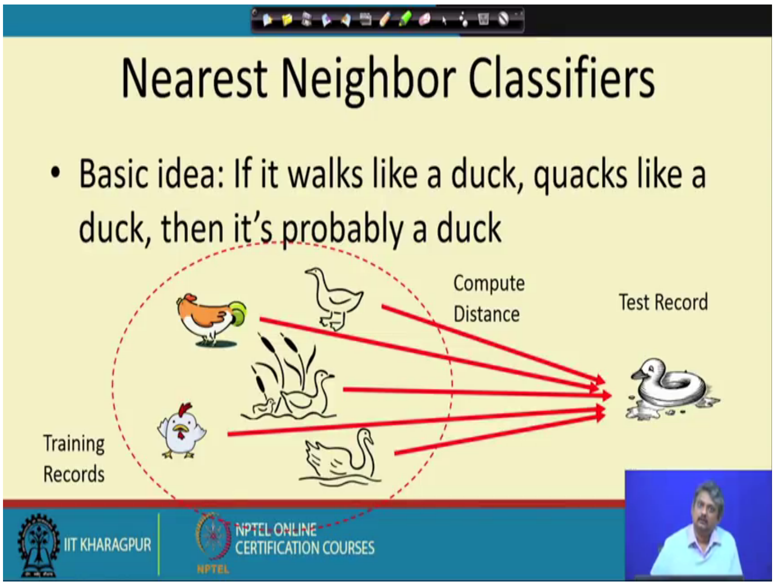
- Training set of examples with Known class labels
- If you want to classify a new example
  - find its distance to each of the training examples
  - find the K-Nearest neighbors
- Find the class which has the highest number among these K-neighbors (Winner)
  - Assign that as class label

Finds its distance to each of the training set example so a new point comes they that is a little bit worry some point; that means, when a new point comes you have to in order to find the K nearest neighbour, you have to find its distance to each and every training set point each and every it is otherwise you do not know which are the closest. Then once you compute this distance find the find the K closest neighbours actually K nearest neighbours, then find the class which has the highest number among these K examples, K neighbours see which class has the highest and assign that class that as the class of the. So, I this is the winner class after the vote.

So, the vote is just something to visualize it, actually means that basically you are finding which class among these K neighbours have the highest number of points. So, among this K most of the points come from which class and assign that as the class level of the new point, the class level. So, very simple very simple rule a little bit computationally expensive because you have to find the distance to all these training points and you have to sort them to find out the top K, but otherwise it is a very simple rule the good thing is that we have already shown that this is a, this is a approximation of the Bayesian classifier a nonparametric approximation of the Bayes classifier.

The, but this principle actually was philosophically sort of known to people before it is like suppose there is suppose there is a new. So, what we have is that let me draw this picture.

(Refer Slide Time: 26:42)



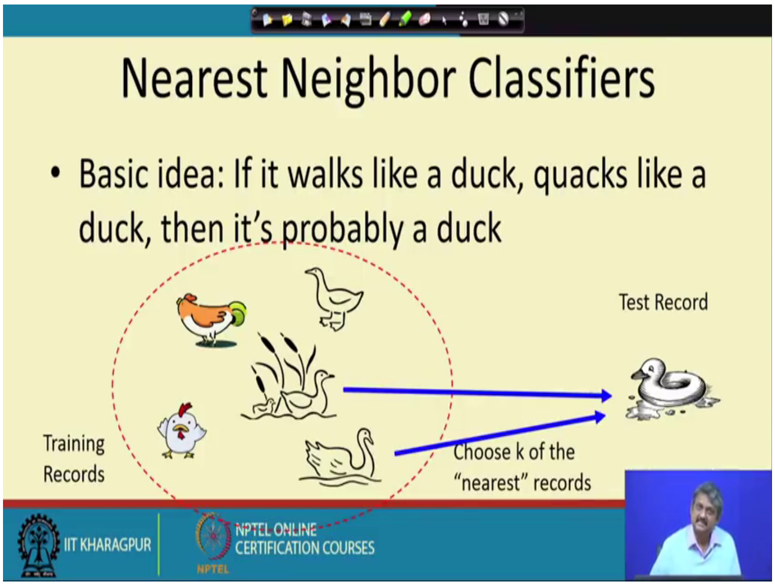
Nearest Neighbor Classifiers

- Basic idea: If it walks like a duck, quacks like a duck, then it's probably a duck

So, in my training set I have different type of birds. So, this is a bird, this is a bird, this is a bird these are my training records and they are described by some attributes say the nature of their, the their quacks their colour and so on. So, you describe them each of the bird by some attributes and I have a new bird whose species I do not know, all that I do is to in terms of the attribute for the new record I find this distance to the known training birds, note that training said my content more than 1 duck species there are different.

Types of ducks it may contain some hen species, it may contain some angry bird species.

(Refer Slide Time: 27:58)



Nearest Neighbor Classifiers

- Basic idea: If it walks like a duck, quacks like a duck, then it's probably a duck

So, I find out in terms of the attribute these distances see which is the closest, K closest say K equal to 2 here and among these 2 in this case both are ducks, they may be in the if if K was a 3 it may be 2 duck 1 hen. So, the majority of the members of this nearest neighbour that class is the class of this new record is the class of this new record then is the simple rule. Now, I have so this is a very popular and simple rule there are some things we need to consider for example.

What is the value of K? How do you compute distance? How do you solve this computational time problem and how do I construct a training set? So, these are certain issues that one needs to think of before using it as an as a determining algorithm all right. So, in our next lecture we will discuss some of these topics. So, I hope you have got a overall idea of the basic principle of K nearest neighbour classifier and how it approximates the Bayes classifier. We will go into details of applying this for our particular problem in hand in the next lecture and keep in mind that this is one of the most successful and popular classifier used in industry all right. So, thank you for today we will continue in our next lecture or discussion on this.

Thank you.