

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

Lecture – 12
Bayes Classifier I

Let, me start my lecture on the next classification algorithm Bayes classifier. So, in the previous lectures we have discussed the classification algorithm known as decision trees based on the training example, we have the tree which gave us the class information. One extension to that would like to do is that instead of using the exact training set we will use some kind of statistics or general, I mean overall statistical properties of the training set to design this classifier. The advantage is moment it being in this probability we have some kind of confidence score and we can do away with noises and other inaccuracies more than that we can do in a decision tree. So, I will explain next the Bayes classifier.

(Refer Slide Time: 01:25)

A Simple Species Classification Problem

- Measure the *length* of a fish, and decide its class
– Hilsa or Tuna

The slide contains two photographs. The left photograph shows a man in a light blue shirt holding two fish. The right photograph shows two men in white shirts and caps holding a large fish. At the bottom of the slide, there are logos for IIT Kharagpur and NPTEL Online Certification, along with a navigation bar.

Let, me start with a example, a small example which will help you understand. So, I want to build kind of a automated robotic system, which will classify between two species of fish. So, it is like a fish stroller which catches fish and a robot picks up one type of fish and puts in one stack and picks up another type of fish and puts in a separate stack and what I am going to do is that.

So, I have for the moment, I will consider two classes a class of fish known as Tuna and another class of fish known as Hilsa. And what this system will do is the following, at the moment it catches a fish it will measure the length of the fish and depending on the value of the length, it will put it into one of the two classes. More specifically, I will have a rule as follows; I will have its kind of a boundary length. So, any fish whose length is less than this boundary length, I will put it in one class and any fish which measures longer than this boundary length, I will put it in the other class.

So, initially let me consider only two classes of fish, this is one class Hilsa and this is another class Tuna. So, as you can figure out from this figure that this measurement or attribute length is indeed an indicative attribute because one fish the Tuna fish the right hand side is the Tuna is usually longer than the other class the left hand side the Hilsa, but, so using length if I follow this rule that, if I follow this rule sorry, if I follow this rule that any fish measuring less than some l star is Hilsa and any fish measuring larger than some l star is Tuna. As a data miner your job is to find out a good value of this boundary length l star, so that I can as accurately as possible classify these fishes.

And so how do I do it, you try to think how do I do it. So, the thing is first thing I do if I have to find out this l star is I will go to the market.

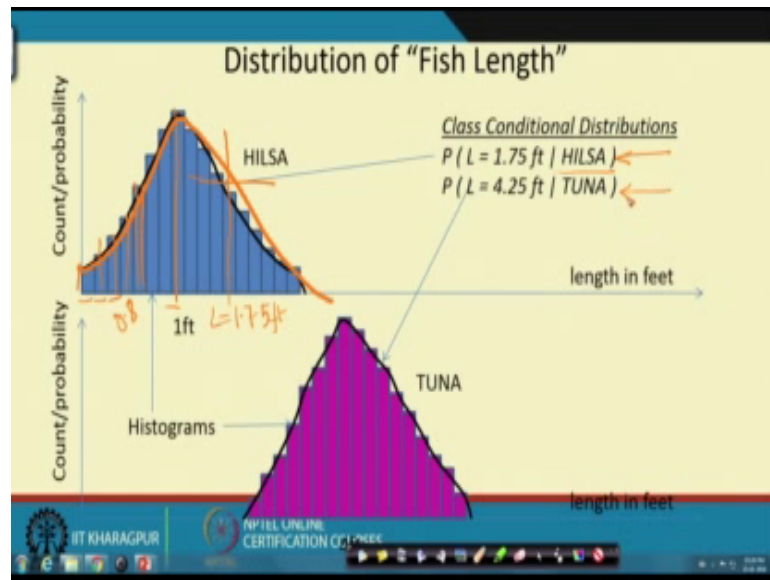
(Refer Slide Time: 04:39)

The slide is titled "Collect Statistics ...". It features two photographs. The left photograph shows a large pile of small fish, labeled "Population for Class Hilsa". The right photograph shows four men standing on a boat deck, each holding a large fish, labeled "Population for Class Tuna". At the bottom of the slide, there are logos for "IIT KHARAGPUR" and "NPTEL ONLINE CERTIFICATION COURSE". A small video inset of a man is visible in the bottom right corner.

I will buy a lot of or rather measure a lot of Hilsa fish and I will go to the market and measure a lot of Tuna fish measure the length of lot of Tuna fish. So, that I get an idea

usually how long are Tuna and usually how long are Hilsa. So, in more formal terms I will collect statistics about a population of Hilsa the length and a population of Tuna. And let us, say I have, I will measure the length of thousand Hilsa and I will measure the length of thousand Tuna and then what I do with these measurements?

(Refer Slide Time: 05:29)



We do the following; I will draw a graph like this, what the graph will do? The x axis of the graph will be length of the fish, in let us say feet and I will break down this x axis into small intervals, small bins of length say one inch and this thousand Hilsa are that I have collected statistics, about I will measure how many have length between zero to one inch, how many have length between one sorry, one inch to two inch, how many have length between.

So, what I will do I will measure how many have length between zero to one inch, one to two inch, two to three inch. I will count out of this thousand how many have and that number I will plot in the corresponding y axis. So, if I actually draw this plot, I will sort of get a bar chart like this. So, these bars I will get so basically, I can't do a little more sophistication instead of exactly plotting in the y axis how many fish have length between zero to one inch one to two inch, I will plot what fraction of this thousand fishes have length between zero to one what fraction of this thousand fishes have length between one to two and so on.

So, this number actually will be between zero and one y axis and I will plot it. So, this bar chart will be actually called a histogram of the distribution of length values of a population of Hilsa. So, you can see that the distribution has a kind of a peak at let us, say 1 foot, so that is most common length that an Hilsa has and then there are very few Hilsa of this length, small length and very few Hilsa of very large length also, I do this. So now, in my population size is large and if my bins are very, very small these intervals are very, very small then, this bar chart.

If I draw it will gradually approach a smooth curve like this, it will gradually approach a smooth curve like this, what is the meaning of the curve? The meaning of the y axis of the curve actually gives some probability values, what is the probability value? This tells that if the fish is known to be Hilsa suppose 1.75 feet, x equal to L equal to 1.75 feet, if the fish is known to be Hilsa this y value that I read off tells me what is the probability, that fish will measure 1.75 feet. In other words this y axis will give me what is the probability a Hilsa fish measures 1.75 feet, and each value gives like that. In fact, I will call this smooth curve something called a probability distribution; it is a continuous version of the histogram.






What type of probability distribution? This probability distribution I will call a class conditional distribution meaning that if it is known that the class is Hilsa what is the probability if given the class is Hilsa what is the probability L takes on a value 1.75. Similarly, given the class is Hilsa. So, what is the probability L takes on a value let say 0.8. So, this curve I will call a class conditional distribution of the length. Now I will repeat this experiment for this Tuna population also, I will make small, small bins I will draw this histogram, I will draw the probability distribution and I will make the class conditional distribution for Tuna also alright.

Next, what I do is the following after drawing these two constructing these two distributions from the population, my job is to find this decision rule find a boundary B .

(Refer Slide Time: 11:59)

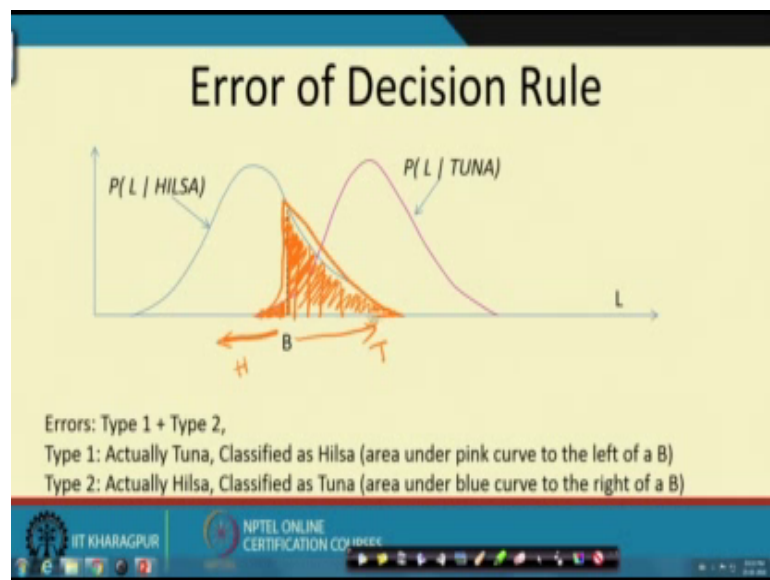
Decision Rule

- If length $L \leq B$
 - HILSA
- ELSE
 - TUNA
- What should be the value of B (“boundary” length)?
 - Based on population statistics



So, that any length less than b is Hilsa otherwise Tuna, based on these two class conditional distributions give me a good value of b let us, see how to do that.

(Refer Slide Time: 12:30)



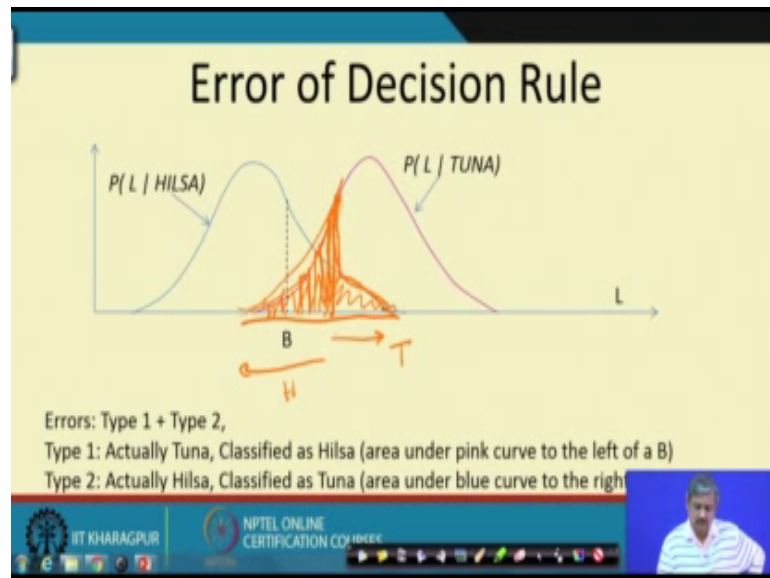
Let me place these two distributions side by side. So, x axis is L and the y axis this blue line is probability L , given Hilsa and this is probability L given Tuna. So, let me draw these two distributions side by side. And let, me assume that some value B here is my boundary of the decision rule. So, let me see how much error I will commit, if I use this decision rule with this value of B . So, there will be two type of error, one type is it is

actually Tuna, but classified as Hilsa. So, anything to the left of this is classified as Hilsa, anything to the right of this is classified as Tuna.

So, this y axis is the count of the number of under this pink curve is the number of fishes, which are specimens which are actually Tuna, but this rule tells Hilsa. So, actually the area under these curves gives me the probability of a Tuna being classified as Hilsa, the area under this curve.

Type two it is actually Hilsa, but classified as Tuna. So, similarly these so many fish plus, so many fish plus, so many fish plus, so many fish plus, so many fish plus, so many fish plus they are all rule says you are Tuna, but they come from a Hilsa population. So, area under this curve is type two errors. So, the total error type one plus type two is this much area this much area, now suppose I place B somewhere else, I place B let us say here.

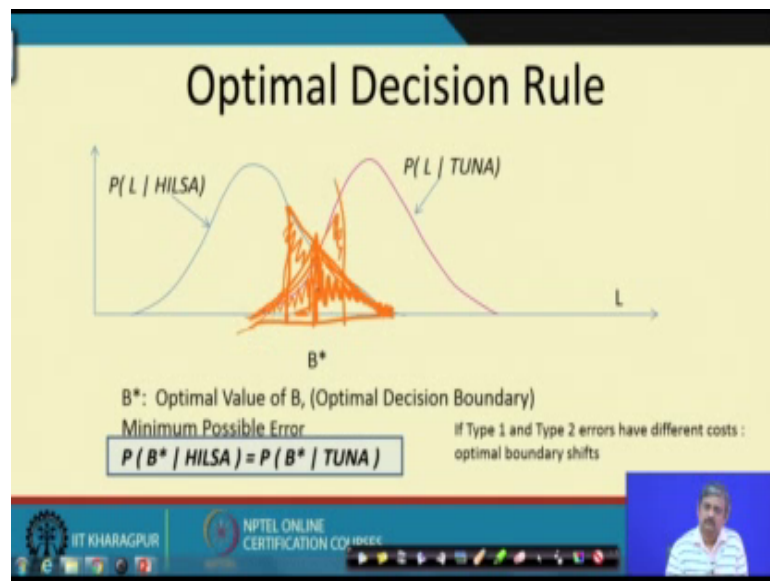
(Refer Slide Time: 15:54)



So this is my Tuna now, this is my Hilsa now, so what is type one error now, what is type two error. So, total error is this much this area. So, I hope you understand why this is this area because see remember y axis has how many fish, how many Tuna fish in this range of values, how many Tuna fish in this range of value, how many Tuna fish in this range of value. So, I add them all up these heights. So, I integrate this region area under this curve and I get the area.

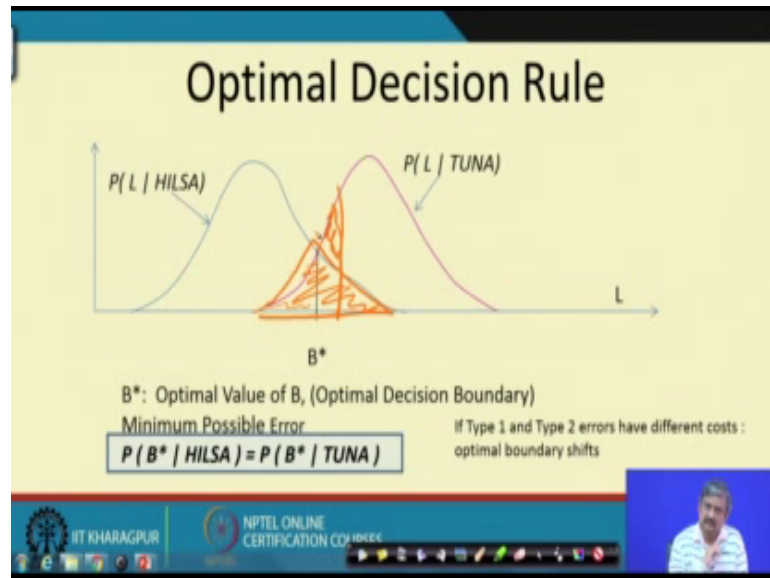
So, this height plus, this height plus, this height plus, this height plus, this height in the limit will be this area. So, my total error is now this alright so depending on, where I place my B, I will get defined amount of error now, I ask you the question where should I place B. So, that I get the smallest amount of error so where should I place B so that I get the smallest amount of error. So, as you might have guessed already that probably.

(Refer Slide Time: 17:51)



Let us, see the best place to place B, I call it B star is at the intersection edge of these two curves, it has the intersection of these two curves. Why? Because if I place it there type one error is this area, type two error is this area. So, total error is this much this much area anywhere else I place B you can check by placing B, somewhere else the error will be some other area, that area will always be this area plus something. Place it here, it will be this area plus this much, it will be more than the error by placing at the intersection. So, what I claim is that if we place in the intersection.

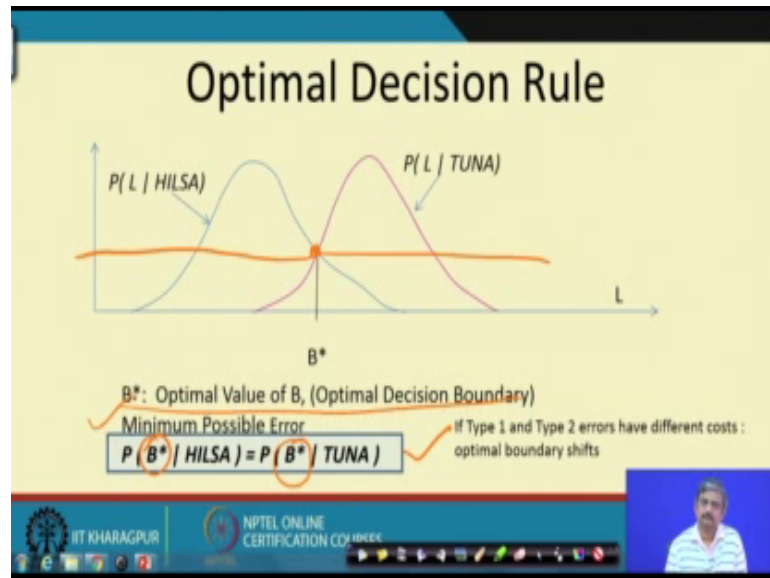
(Refer Slide Time: 19:18)



The amount of error you get is smallest among all placements of B , my argument is if you place me somewhere, if you place B , somewhere else you will always commit an error which is this area plus some extra anywhere you place B . So, it is the intersection of these two probability density functions probability distributions of Hilsa and Tuna, where I would like to place my B star. So, that I get the minimum possible error.

So, let me examine what this intersection basically means. So, intersection basically means it is that value of L , where the Hilsa distribution gives a probability values same as there is given by the Tuna distribution.

(Refer Slide Time: 20:34)



The intercept so the y value by of this pink curve at this B is same as the y value of this blue curve at this B, it is such an B. So, basically I can say that my B star is nothing, but a value where P B star given Hilsa, equals P B star given Tuna the class conditional probabilities are same for both classes, this thing now, this is what I found out and I have shown that it is this particular value of B, which is the optimal giving minimum possible error.

(Refer Slide Time: 21:55)

Species Identification Problem

- Measure lengths of a (sizeable) population of Hilsa and Tuna fishes
- Estimate Class Conditional Distributions for Hilsa and Tuna classes respectively
- Find Optimal Decision Boundary B^* from the distributions
- Apply Decision Rule to classify a newly caught (and measured) fish as either Hilsa or Tuna
– (with minimum error probability)

So, in summary, what I do is the following I measure the length of a population of good population of Hilsa and Tuna from these measurements, I estimate the class conditional distribution curves from the histogram, find out where the curves intersect that is my B star for a new fish, I measure its length check less than or greater than B star, put it into Hilsa Tuna class and I have shown that this will give me the smallest probability that I make a misclassification.

(Refer Slide Time: 22:51)

Location/Time of Experiment

- Calcutta in Monsoon
 - More Hilsa few Tuna
- California in Winter
 - More Tuna less Hilsa
- Even a 2ft fish is likely to be Hilsa in Calcutta
- a 1.5ft fish may be Tuna in California

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSE

But one more factor, now I should look into also see I have built these curves of the distributions by taking a population. What if this population is biased suppose I have collected this population in Calcutta during the monsoon season? So, my out of this two 1000 fish, 1000, 2000 fish that I have collected from the Bay of Bengal, there will definitely be almost no Tuna and large number of Hilsa. Similarly, if I do this get this population in California in winter season and from the Pacific Ocean, I draw this population definitely there will be lot more Tuna than Hilsa.

So, if I using this population the first population if I draw the distribution curves, I will get a boundary which will be biased, because even a two feet fish in Calcutta, will probably be a Hilsa not a in Tuna, similarly the curves that I get from the second population even a 1.5 feet fish will be a Tuna, so the curves not a real reflection of the actual length of Hilsa and Tuna.

So, how do I account for this so basically, what I am trying to say is the following, if I do experiment in Calcutta, this is my length may be the Hilsa curve will look like this, and the Tuna curve will look like this. So, I will get B start somewhere here even. So, maybe this is 2 feet, but because there are few Tuna, but that is not the reality because most of the Hilsa there is rarely a 2 feet Hilsa.

So, this B star is not a good B star similarly, if I do it here, I get lot of Tuna and few Hilsa so I get a B start here, so it is again not a good B star. So, what do I do? How do I solve this problem? If I have knowledge of the relative distribution of Hilsa and Tuna in my population, so one thing I can do is the following, I can modulate, I can multiply these two curves I can boost up one curve and bug down another curve.

So, that B star comes back, so what I am saying is that, what I am saying is that, I multiply, if I am doing in Calcutta, I pull down this distribution and pull up the Tuna distribution. So, that B start shifts to the left if I am doing in California, I boost up the Hilsa distribution by multiplying it by some constant factor and pull down the Tuna distribution. So, that B star comes back, so I kind of do a scaling of the class conditional distributions using the bias factor or the dominance factor of the classes, let us see how to do that.

(Refer Slide Time: 27:09)

Apriori Probability

- Without measuring length what can we guess about the class of a fish
 - Depends on location/time of experiment
 - Calcutta : Hilsa, California: Tuna
- Apriori probability: $P(HILSA)$, $P(TUNA)$
 - Property of the frequency of classes during experiment
 - Not a property of length of the fish
 - Calcutta: $P(Hilsa) = 0.90$, $P(Tuna) = 0.10$
 - California: $P(Tuna) = 0.95$, $P(Hilsa) = 0.05$
 - London: $P(Tuna) = 0.50$, $P(Hilsa) = 0.50$
- Also a determining factor in class decision along with class conditional probability

I do this scaling using another probability term called the apriori probability. So, that what apriori probability means, if I randomly blindly pick up a fish in Calcutta without

measuring its length, if I randomly pick up a fish what is the probability that it is Hilsa and what is the probability it is Tuna. So, basically I tell that in my population, from my population, if I randomly pick up a fish blindly pick up a fish I don't measure its length what is the probability that I get a Hilsa and what is the probability I get a Tuna .

So, these probabilities do not depend on the length of the fish unlike the class conditional, this actually depends on the population distribution the bias in the population, how many fish how many Hilsa is there in the population, how many Tuna is there. So, some sample values you can check that maybe if I in Calcutta Hilsa has a high probability Tuna has a lower California, other way around London maybe 50-50, now just for example, I don't know whatever fishes are available in London. So, these values I use to sort of scale up and down my earlier distribution ok.

(Refer Slide Time: 28:55)

Classification Decision

- We consider the product of *Apriori* and *Class conditional* probability factors
- *Posteriori probability (Bayes rule)*
 - $P(\text{HILSA} | L = 2\text{ft}) = P(\text{HILSA}) \times P(L=2\text{ft} | \text{HILSA}) / P(L=2\text{ft})$
 - *Posteriori* \approx *Apriori* \times *Class conditional*
 - *denominator is constant for all classes*
- *Apriori*: Without any measurement - based on just location/time – what can we guess about class membership (estimated from size of class populations)
- *Class conditional*: Given the fish belongs to a particular class what is the probability that its length is $L=2\text{ft}$ (estimated from population)
- *Posteriori*: Given the measurement that the length of the fish is $L=2\text{ft}$ what is the probability that the fish belongs to a particular class (obtained using Bayes rule from above two probabilities).
 - Useful in decision making using evidences/measurements.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSE | NPTEL

So, I will do so think on this think on how to do this and I will explain it in the next lecture thank you I will do it in the next lecture.