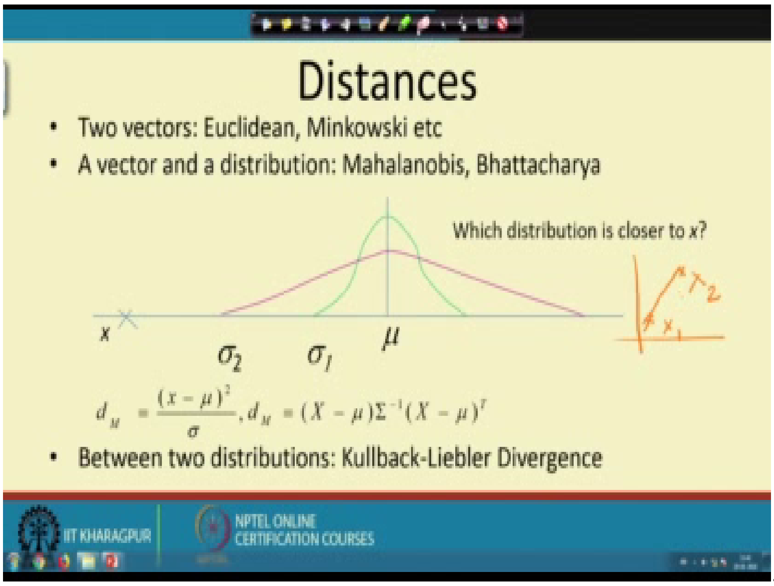
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture – 15
Bayes Classifier IV

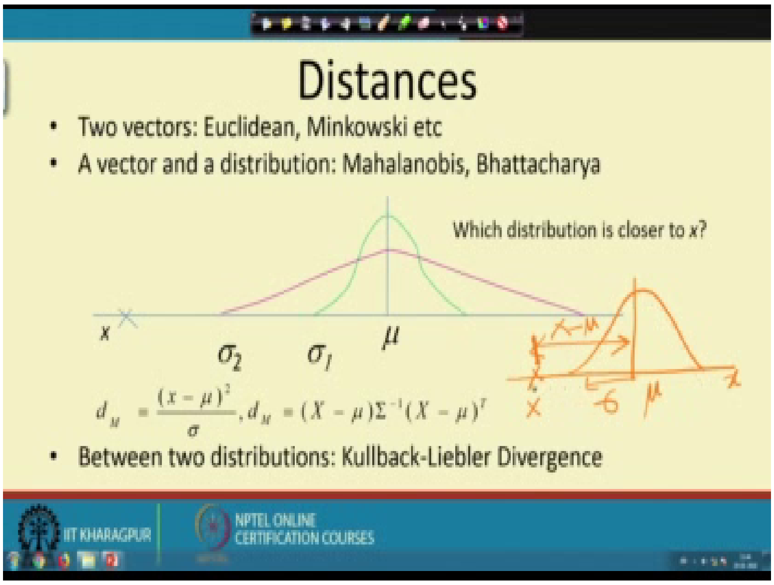
We are discussing the finding out a decision boundary, when you have two bivariate Gaussians as the posterior probability distributions, and the Gaussians are still circular diagonal covariance matrix with equal value in both the dimensions, but they are of different sizes for the case, where both are equal sizes we find that it is the it is the perpendicular bisector all points which are equidistance from both the means ok. So, let me digress a bit and discuss another concept.

(Refer Slide Time: 01:09)



So, as you know if we have two points x 1, and x 2, you can define various types of distances, between the Euclidean distance the straight line distance, Murkowski distance, we studied earlier, but how do we measure the distance between not two points, but a point n n distribution. So, what I am talking of is like this.

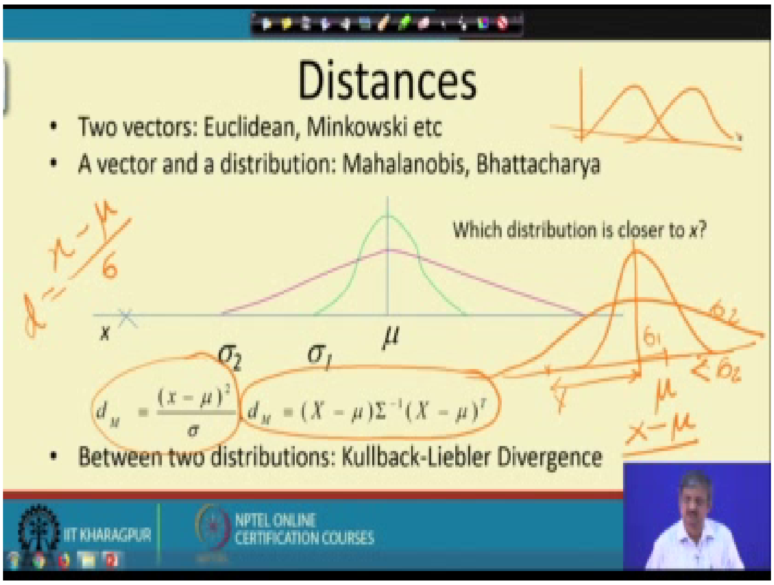
(Refer Slide Time: 01:41)



So, I have a distribution let say one dimensional for the moment we can extend it to multiple dimensions. So, we have a distribution with say mean mu and some variance sigma and I have a point x I want to say I want to find out how far is this point from this distribution, say one typical case may be something like this that I have the distribution of length of all tuna fishes and I catch a random fish, how much is it tuna miss; that means, how far is it from the distribution how separate it is from the distribution.

So, one obvious distance that probably you can consider is nothing, but the distance between this point under consideration x and the mean of the distribution. So, x minus mu so that is my distance how far it is from the mean value. So, in the case of tuna suppose the mean tuna length is 2 feet or 3 feet and suppose a fish is another fish is 4 feet. So, it is 4 minus 3, 1 feet is the distance, but there is one problem suppose the length of tuna distributed like this whereas, maybe the length of another fish another species of fish is distributed like this while let me draw a bigger picture.

(Refer Slide Time: 03:24)



So, this is one distribution mean at mu divergence sigma 1 and this is my x and this is x minus mu x stays in the same place, mu stays in the same place, but instead I have a distribution which is more spread apart in other words it has a sigma 2, which is greater than, sigma one I have another distribution with more divergence than this one. So, this has more sigma two then sigma one.

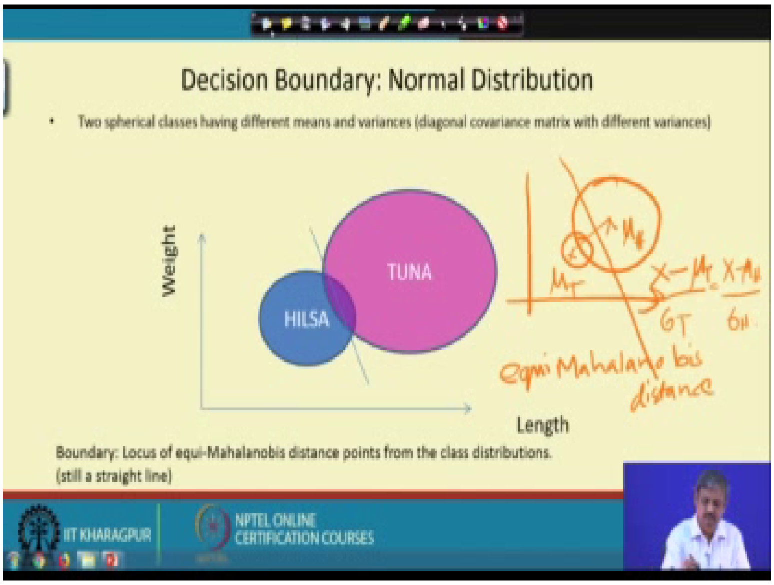
Since the mean is unchanged, if we consider the earlier distance normal distance the distance would will be the same, but intuitively this x is closer to the sigma 2 distribution than the sigma 1 distribution. So, maybe I mean the argument is like this maybe all tuna fish they have a average length of 3, and the maximum variation in the length is say 2.5 to 3.5. So, a 4 feet fish is less likely to be a tuna whereas, maybe another fish whose mean is still 3 feet, but length varies from say 1 feet to 4 feet more divergence in that case, x is 4 feet fish is closer to that distribution.

So, how do we take into account this instead of x minus mu, what I do is I normalize it instead of x minus mu as my distance I normalize it by the spread of the dispersion of the distribution. So, now you see this though x minus mu is same for both this cases, the one with a higher dispersion will have the closer distance smaller distance. So, this concept this new distance d is defined as the mahalanobis distance for a multivariate case, when instead of a sigma you have a covariance matrix it generalizes to this definition.

So, if we do not have this sigma it is just the euclidean distance to you put the sigma in between you have the mahalanobis distance, sigma inverse in between ok.So, this is one type of distance between a point and a distribution, there are other Bhattacharya and other coefficients which can be used. In fact, one can generalize it further which will use later in our course is that you can define a distance between not just a point and a distribution, but between two distributions.

So, that would be called a Kullback Liebler divergence, other measures are also there let us not discussing it for the moment all right; so mahalanobis distance what we will do with this, we show that for the case where, we have two circles one large and one small.

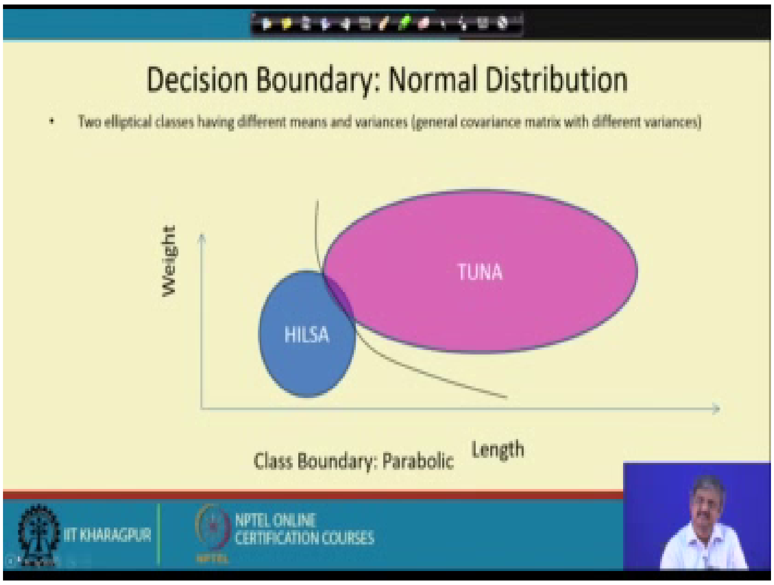
(Refer Slide Time: 07:04)



The optimal based is sun boundary is not the perpendicular distance, which is equi euclidean distance from both the centres, but it is all the points which are equi mahalanobis distance. So, if these are mu h and mu; mu h and mu t x would be all point where x minus mu t by sigma t would be same as x minus mu h by sigma h, again I am not doing the algebra if we work out the definition of the base decision boundary and if you find out equi probable points you would get this expression.

So, this is the general equation, for more general cases where we have ellipses and tilted ellipses maybe.

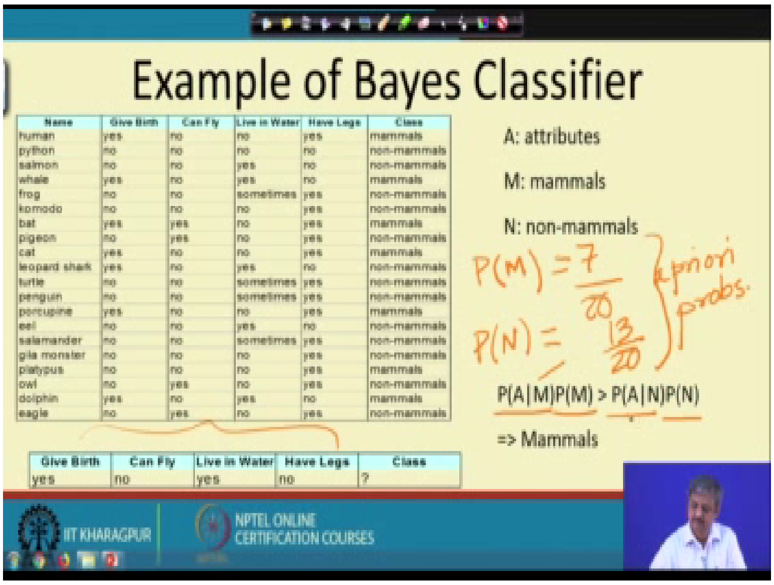
(Refer Slide Time: 08:22)



Then the optimal boundary is no longer a straight line, but takes a quadratic surface, again I leave you at a leave it to you as an exercise that if you just work out the algebra you will easily get the expression for this quadratic surface the idea is find out all x. So, I have the probability values for each of these distributions are same where assuming each of the distribution of normal with some mu and sigma all right.

So, this is one class assume that distributions are normal. Let us, look into finding out this Bayesian classifier for another specific case, I will probably come back to our old example, where the attributes are not the old example is slightly different form of that the attributes are discrete.

(Refer Slide Time: 09:19)



In the sense that they take on a finite set of values earlier it has continuous length can be any number of continuous values. So, we have a discrete set of values and it will be clear to you that in such a case, if you have a training set, I will explain it with this particular table as the training set easy to find out what the Bayes classifier is.

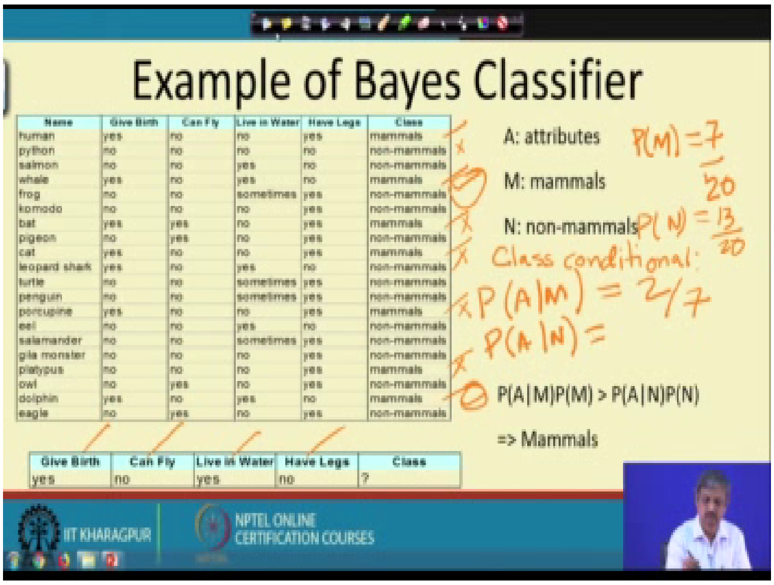
So, let us look at this example, so this is my training set so I have some species whose which is in the first column. So, each of these are my training instances, so I have a human a python is salmon a well a frog and. So, on up to an eagle so I have all this and they belong to one of the two classes either they are mammals or non-mammals and they are each of these species are described by four attributes whether they give birth to an young generation whether they can fly, whether they live in water or whether they have legs.

So, each of these are discrete values. So, this has yes or no, this is yes or no see this yes no or sometimes, which is the amphibian case and legs is also yes or no and I would use the Bayes classifier to classify a new species which I know gives birth, fly cannot fly lives in water does not have legs what is the class mammal or non-mammal.

So, how do I go about this first you answer me let us do the easiest thing first what is the prior probabilities of mammal and mammal is m non mammal is n. So, easy to find p m mammal is you just count how many of these training instances are mammal the bias in the training set so 1 and 2 and 3 and 4 and 5 and 6 and 7. So, it is 7 out of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 out of I may be wrong, I may make count in mistakes, but the way I am doing is I am counting how many mammals are there in the training set divided by total number of examples in the training set. So, similarly non mammal would be 13 by 20.

So, this gives me the prior probabilities apriori probabilities ok. So, now the next thing we want to find is the class conditional; that means, known that an instance belong to the class mammal what is the probability the attribute values are this because, I want to classify this example similarly known that an instance is non-mammal, what is the probability that the attribute values are this.

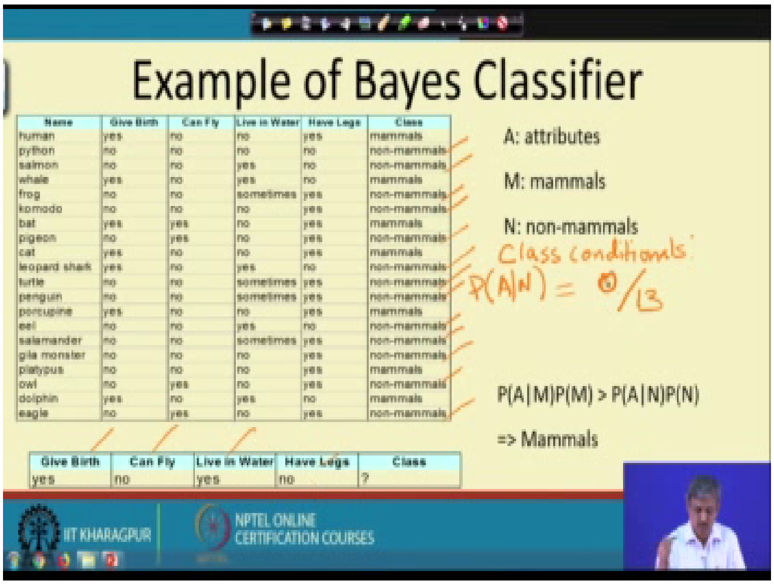
(Refer Slide Time: 13:46)



So, in other words, I write down the priors here, known that it is mammal what is the probability I have this set of values let us see. So, pick up the mammals there are 7 of them 1, 2, 3, 4, 5, 6 and 7 find out how many of these 7 cases, have this set up attribute values.

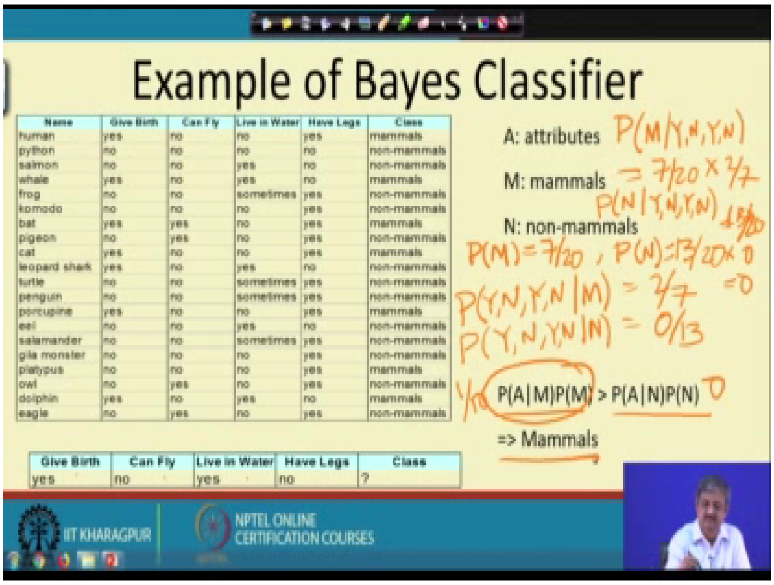
So, yes, no, yes, no, yes, no, no, yes, not this second one yes no yes no yes this one, one yes, yes, no, yes, not this one yes, no, no, yes, not this one yes, no, no, yes, not this one this one no, no, no, yes, not this one yes, no, yes, no, yes, this one. So, two I have found two mammals I have found whose attribute values are exactly this. So, my prior probability is two out of this 7; I sorry, not prior; the class conditional probability is two out of this seven repeat this for non-mammals.

(Refer Slide Time: 15:44)



Let me take the non-mammals, 1 and 2 and 3 and 4 and 5 and 6 and 7 and 8, 11, 12, 13. So, something by 13 out of this 13 how many match; this no does not match, does not match, does not match, does not match does any of this match does any of I can't find any maybe you can find something. So, I want a non-mammal whose attribute values match this set does it match just check out if it matches yes no yes no no I do not change. So, it is 0, so if I summarize what I had got.

(Refer Slide Time: 17:12)

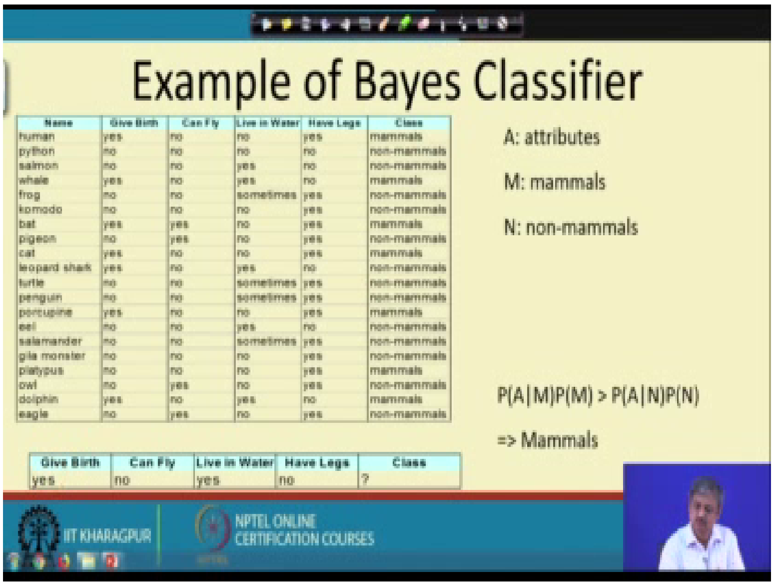


I had got p m to be 7 by 13 and p n sorry it is not 7 by 13 (Refer Time: 17:21) 7 by 20 and p n is 13 by 20 and p a let me write a as yes no yes no yes no yes no yes no given it is mammal is 2 upon 7 and the probability let the attribute values is yes no yes no given it is non-mammal is 0 upon 13 ok.

So, you just multiply these two quantities to get the posterior. So, one side I get probability mammal given yes no yes no applying the Bayes rule is prior times class conditional 7 by 20 into 2 by 7 probability non mammal, is 13 by 20, 13 by 20 into 0 which is 0. So, one side I have 7 by 20 into 2 by 7 which is 1 10th. So, this side I have 1 by 10 and this side I have 0. So, the map maximum aposteriori probability tells me that it is more likely to be a mammal than a non-mammal higher posterior.

So, I classify it as mammal by the map rule. So, this is what I do I have the steps in detail, but I have the steps in detail in the lecture notes you can consult that, but I guess you have got it.

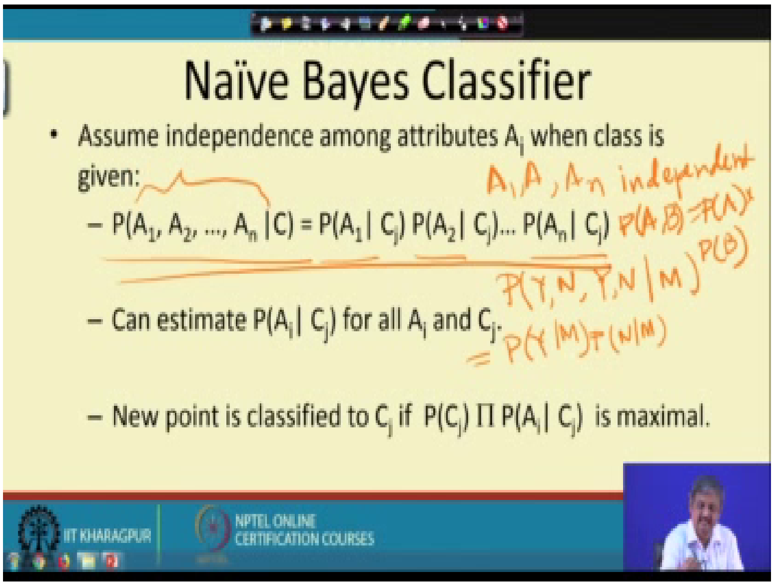
(Refer Slide Time: 19:24)



So, theoretically this is approach, but the problem is that we have seen one of the probability went to 0 because, this sample is not big enough to get an example of a non-mammal, which matches this attribute pattern so; that means, what is happening is that because of the small sample that we have only these many my probability estimates are not good enough not good enough I should not say not good maybe they could have been better.

How they could have been better look see, the main problem was I didn't find a non-mammal sorry, I didn't find a non-mammal having this attribute values together all of them at a time, but I do have a number of non-mammal and for mammal where individually these attributes there. So, lot of non-mammals having yes not of non-mammal having no and so on and can I do that use that information.

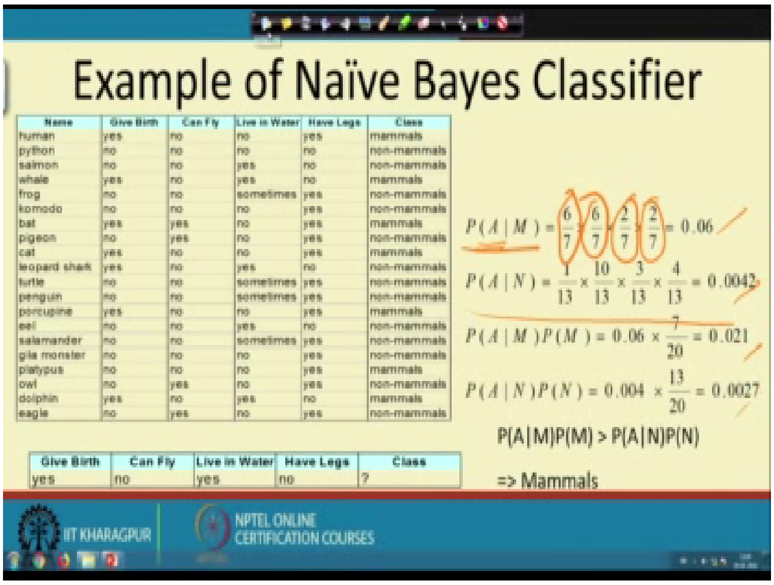
(Refer Slide Time: 21:01)



So, that is used in one particular setting of this Bayes classifier called a naive Bayes classifier. It uses one assumption; it says that the probability of all the attributes taking on a set of values together given a class is product of the individual attributes taking on those values. So, probability yes no yes no yes even mammal for that attribute and.

So, on this is true when we have something called that a1, a 2 these attributes are independent; that means, we say that two variables a and b if they are independent and p a b joint probability is product of their individual probabilities ok. So, that is what I use recent is an assumption called the independence assumption, if I make this assumption then life becomes easy.

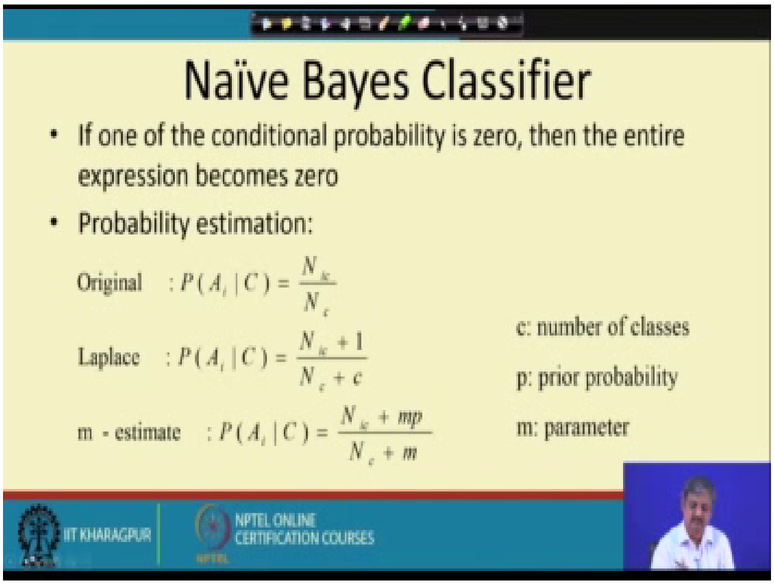
(Refer Slide Time: 22:53)



Now, now let us see. So, what happens when we make the independence assumption probability of yes no yes no is probability of yes given mammal for give birth, no given mammal for can fly no live in water yes live in water for mammal and. So, on if you count this value; that means, how many mammals out of 7 have give birth equal to yes count that that will be 6. So, I get this similarly how many can cannot fly, how many lives in water and how many have legs.

Because they are independent if I just multiply them of I will be having the probability they are jointly having this set of values and I repeat this for non-mammal and I multiply them and still I get this the point to be noted is that I do not have a 0 estimate, anymore it is easier I have with the same amount of training data I can get a better estimate provided I this assumption is correct ok.

(Refer Slide Time: 24:29)



So, this is the net classifier sometimes one additional thing is done in a classifier is that if some of the probabilities still go to be 0, I do something called a smoothing. So, I add some random value and make it better make it better. So, that will help me, so I think this is clear how to calculate the net Bayes, if you I will give you some exercises which if you do it will become even more clear. So, with this I close my discussion on the Bayesian classifier will go into the next topic in my next lecture.

Thank you.