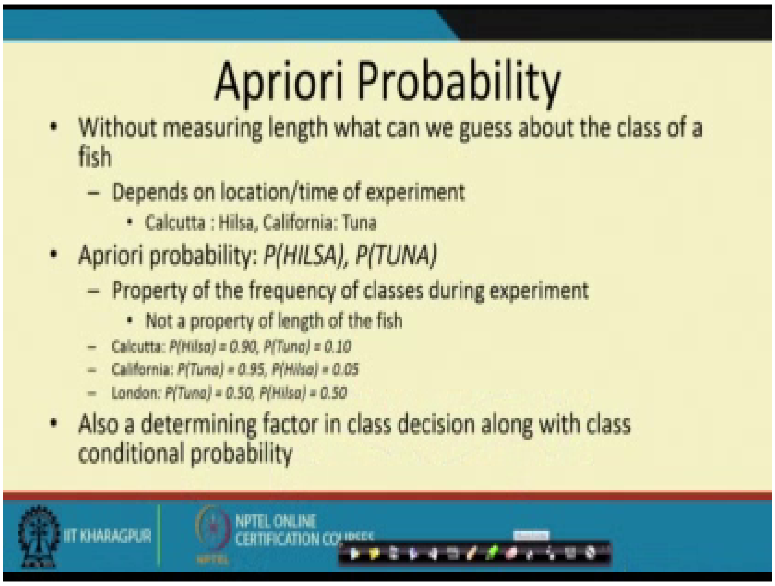
Data Mining
Prof. Pabitra Mitra
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur

Lecture – 13
Bayes Classifier – II

We continue our discussion on the Bayes classification rule.

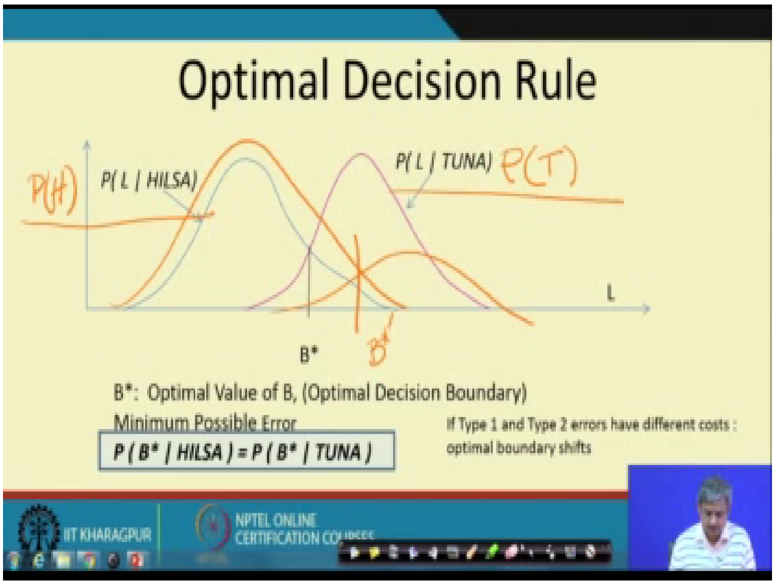
(Refer Slide Time: 00:25)



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

(Refer Slide Time: 00:33)



Optimal Decision Rule

B*: Optimal Value of B, (Optimal Decision Boundary)

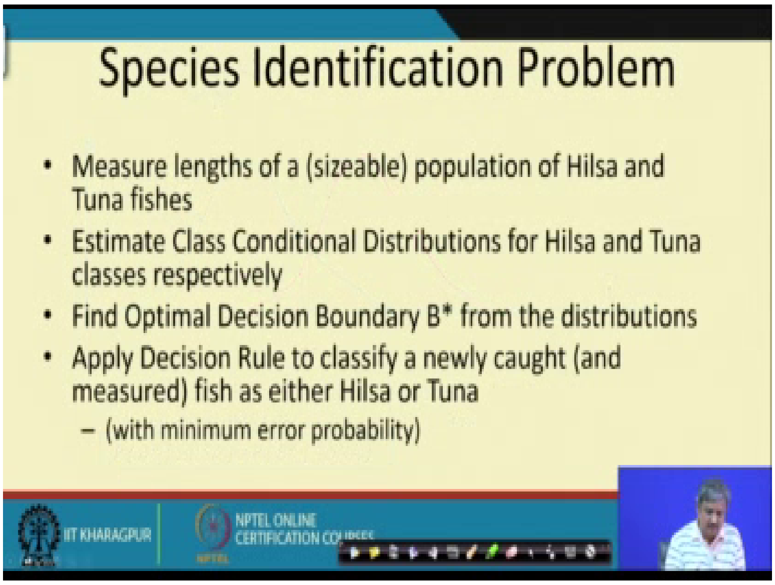
Minimum Possible Error

$\boxed{P(B^* \mid HILSA) = P(B^* \mid TUNA)}$

If Type 1 and Type 2 errors have different costs : optimal boundary shifts

So, to quickly review what we have studied so far, we draw the class conditional distribution that is probability that; if I know the classy HILSA, what is the probability that that fish has a length L. Similarly, for TUNA and I found that the optimal decision boundary was at the intersection of these two form points.

(Refer Slide Time: 00:56)



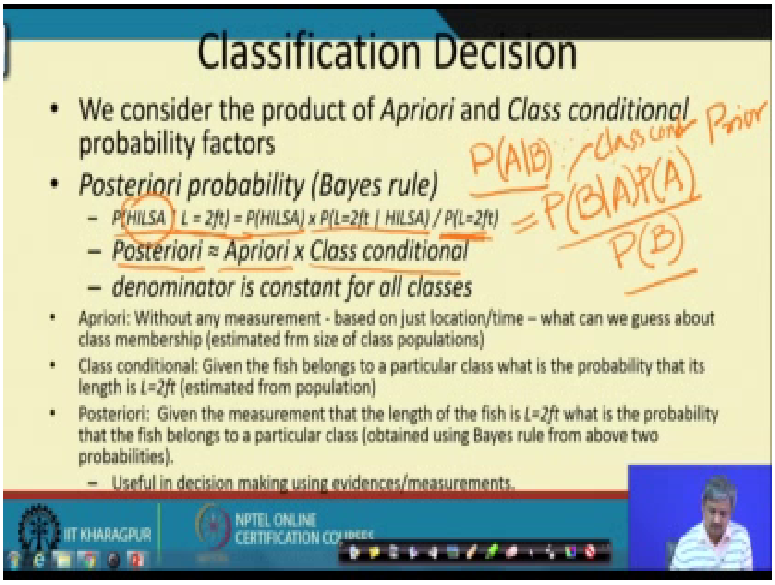
Species Identification Problem

- 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, one problem with this approach is that it will this this curves may be biased depending on where we take our population from to compensate for that we introduce a notion called apriori probability. So, apriori probability is probability that a randomly chosen phase belongs to class HILSA and a randomly chosen fields belongs to class TUNA know L involved.

So, what I do is the following in my original curves let me show you the curve I multiply this class conditional curve by this apriori probability, by this apriori probability and I multiply this class similarly by this apriori probability. So, depending on what my P H equal to P T the curve remains as it is if P H is greater this curve goes up, the other curve goes down and vice versa. So, I scale the class conditional by the prior multiply it by the prior, and my nu B star e the place where this multiplied curves intersect, that is my B star say prime. So, let me write this down algebraically.

(Refer Slide Time: 03:00)



Classification Decision

- We consider the product of *Apriori* and *Class conditional* probability factors
- *Posteriori probability (Bayes rule)*
  - *P(HILSA | L = 2ft) = P(HILSA) x P(L=2ft | HILSA) / P(L=2ft)*
  - *Posteriori ≈ Apriori x 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 frm size of class populations)
- Class conditional: Given the fish belongs to a particular class what is the probability that its length is *L=2ft* (estimated from population)
- Posteriori: Given the measurement that the length of the fish is *L=2ft* 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.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

class cond

Prior

So, I multiplied the class conditional by the prior, and I divide it by the probability of length being two feet for example, in this case. So, I multiply the class conditional by the prior what I get is something called a posterior a posterior distribution meaning why posterior you know the Bayes rule which says that P of A given B into P B equals probability of B given a into P A.

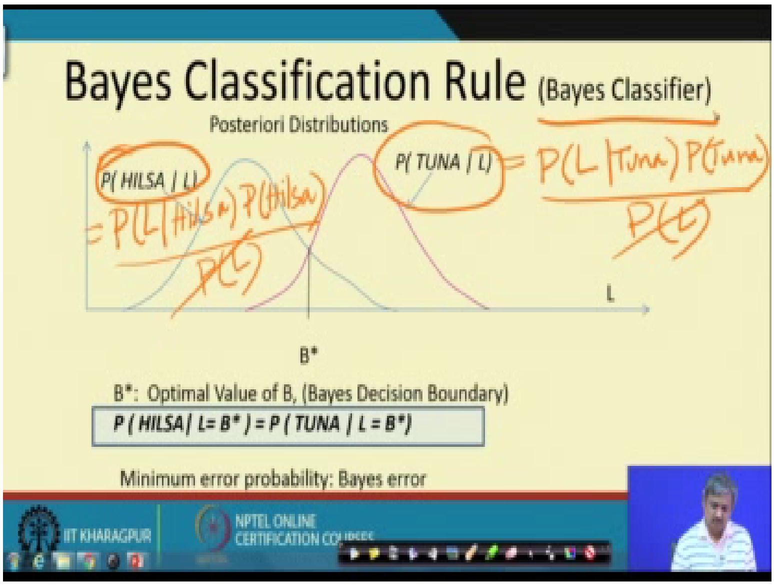
So, what I actually do is kind of, and of bring this to the denominator of the right hand side doing this to the denominator of the right hand side and make it p b. So, a P A is the prior, prior probability aprior probability, P B given a is probability length equal to two feet given it belongs to class a, this is probability of a random piece belonging to class a, this is class conditional. So, class conditional into prior divided by P B note that this denominator whatever class I take is same it is independent of class. So, if I multiply if I multiplied both the curves by one by P B both will equally scale. So, the intersection will not change. So, the denominator is kind of a is a scale factor same for both classes.

Bayes rule tells me that this quantity equal’s probability of a given B means probability of f is being HILSA given that it lengths is two feet. So, if I measure a length to be of f is to be up two feet long what is the probability it belongs to class HILSA, how do I get that I use the class conditional that is distribution of length of all hill surfaces multiplied by the prior probability of a random piece being HILSA and I get the posterior, all right.

So, this I have written out the definitions, class conditional is this prior is this posterior is the product, divided by a constant factor for each of the classes.

So, what I actually do is that I find the posterior for both the classes, I find the posterior for HILSA, I find the posterior for TUNA and instead of earlier I was drawing the intersection of two class conditionals, I just change that rule and I say I draw the distribution of the posterior probabilities so.

(Refer Slide Time: 06:57)



Bayes Classification Rule (Bayes Classifier)

Posteriori Distributions

$P(HILSA \mid L) = \frac{P(L \mid Hilsa)\, P(Hilsa)}{P(L)}$

$P(TUNA \mid L) = \frac{P(L \mid Tuna)\, P(Tuna)}{P(L)}$

B*: Optimal Value of B, (Bayes Decision Boundary)

$P(HILSA \mid L = B^*) = P(TUNA \mid L = B^*)$

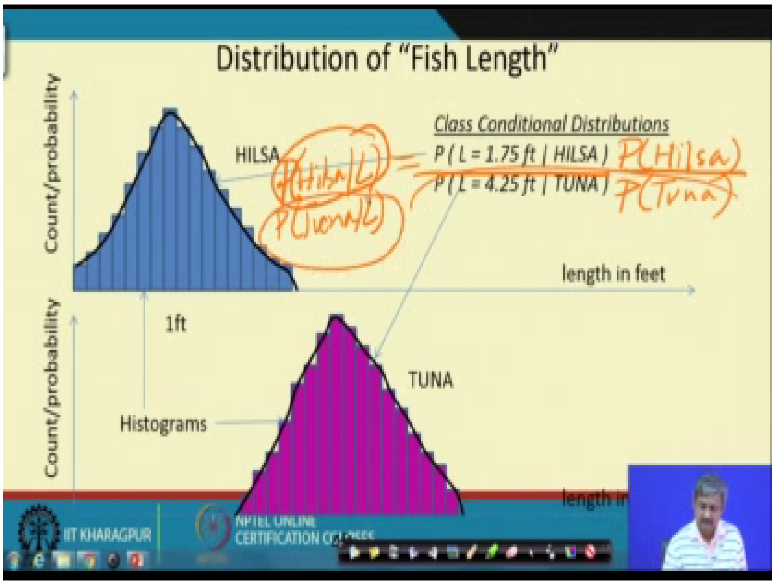
Minimum error probability: Bayes error

.

This is divided by some constant factor probability of a random piece having length L, irrespective of class. So, actually if I drop this in both sides, I will still get the same intersection and my B star boundary is now the intersection of these posteriors not the class conditional as I had mentioned earlier, because it takes best better care of the bias in the population it still has the minimum error.

So, this is my definition of B star it still has the minimum error you can actually seen by looking at the area and this is what I lose. So, this classifier since, it uses this Bayes rule is called a Bayes classifier, and it is actually sometimes called a Bayes optimal classifier and B star is called a Bayes optimal boundary, it has the minimum error possible. So, I repeat my steps what I had done, repeat my steps with these pictures. So, repeat my steps so what I actually do let me look back a little I collect a population.

(Refer Slide Time: 09:18)



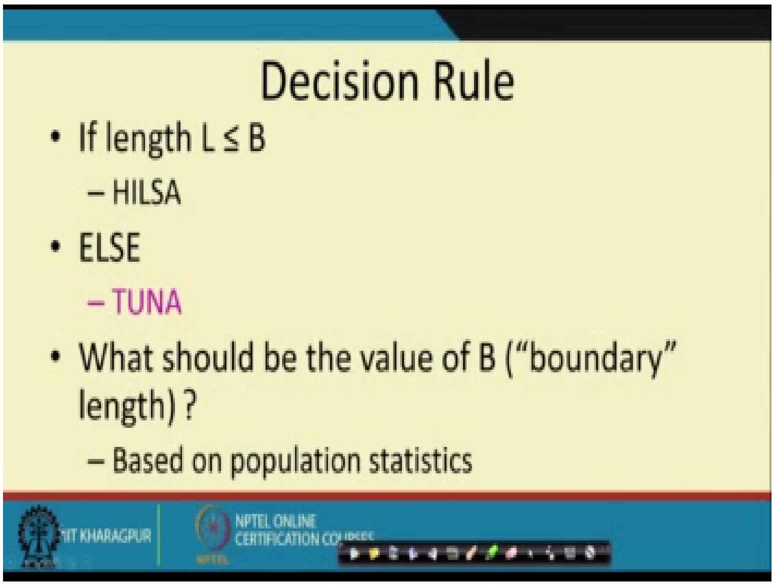
Distribution of "Fish Length"

Class Conditional Distributions

$$\frac{P(Hilsa|L)}{P(Tuna|L)} = \frac{P(L = 1.75\ ft \mid HILSA)}{P(L = 4.25\ ft \mid TUNA)} \cdot \frac{P(Hilsa)}{P(Tuna)}$$

I draw this histogram for HILSA, distribution of length I draw the class sorry, I draw the class conditional.

(Refer Slide Time: 09:26)



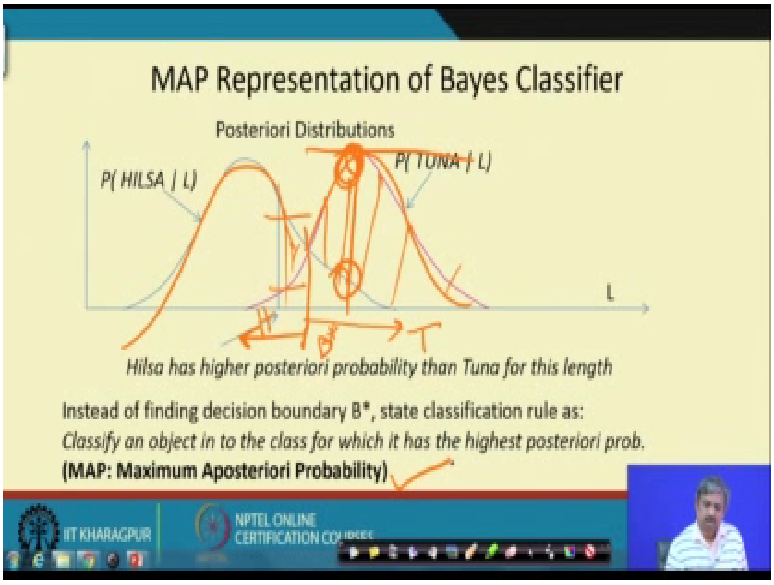
Decision Rule

- If length L ≤ B
  - HILSA
- ELSE
  - TUNA
- What should be the value of B ("boundary" length) ?
  - Based on population statistics

Distributions for class conditional distributions for each of the class, then I estimate the prior probabilities, how I estimate the prior probabilities just count in the population how many are HILSA, how many are TUNA, that fraction is my probability that fraction is my probability and what I do is that this last curve. After I find a prior I multiply this curve, this I multiply by the HILSA, this I multiply by P TUNA, and what I get as a result is the reverse P HILSA given L the posterior P TUNA even L.

So, first compute this from the population, then compute this from the population, multiply them ignore the constant factor in the denominator get the posterior, repeat this for the TUNA class get the posterior, draw these two I have drawn this two draw these two distributions and same as before, set your B star to be the intersection of the posterior distributions and it can it will still have the minimum error, all right. So, now, this is fine this is this I call as the Bayes classifier, let me make life a little bit easier for me.

(Refer Slide Time: 12:10)


MAP Representation of Bayes Classifier

Hilsa has higher posteriori probability than Tuna for this length

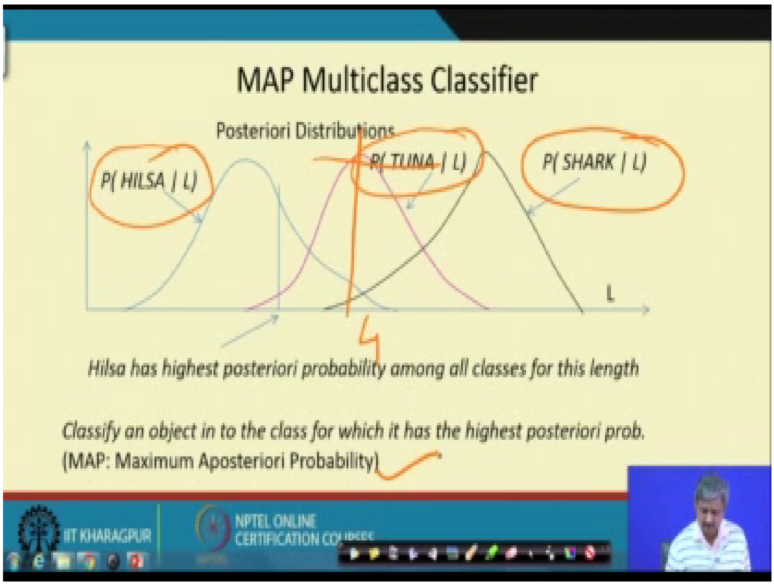
Instead of finding decision boundary B*, state classification rule as:
Classify an object in to the class for which it has the highest posteriori prob.
(MAP: Maximum Aposteriori Probability)

Instead of finding the intersection solving these two equations and finding the intersection, I do the following you make a observation that to the left of this B star the HILSA region, the blue curve the HILSA curve is above the pink curve the TUNA curve, and to the right of this distribution, which is the TUNA region the TUNA curve the TUNA curve stays above the HILSA curve, the pink is above the blue pink is about the blue. So, here pink is above the blue here blue above the pink.

So, instead of finding out what the value of B star is I can do the following given a new fish, you find out its probability of belonging posterior probability of belonging to TUNA, find its posterior probability of belonging to HILSA, compare these two values whichever value is higher, put it in that class, whichever value is higher find this posterior of one class find the posterior of the other class compare these values whichever is higher classify it into that class.

So, I can sort of restate my Bayes classifier as the maximum posterior classifier these curves are the posterior probability curves and my rule is put it into that class, which has the higher highest posterior probability. The map rule you follow the map principle maximum aposteriori probability principle, it is same as the Bayes classifier.

(Refer Slide Time: 14:33)



MAP Multiclass Classifier

Posteriori Distributions

Hilsa has highest posteriori probability among all classes for this length

Classify an object in to the class for which it has the highest posteriori prob.
(MAP: Maximum Aposteriori Probability)

.

These, particular formulation actually helps me generalize this Bayes classifier to more than two classes. So, suppose there are three classes of fish in general let us say there are k classes of fish, let us say HILSA and TUNA and SHARK, I using that Bayes rule, I find this posterior sorry, I find out this posterior, which is a class conditional inter prior for HILSA, I find out the posterior curve for TUNA, pink I find the posterior curve for SHARK black.

Now, for a new fish a measure it L, suppose L is this value L1 I draw a vertical line L1. At that particular point out of these three whichever curve is the top most curves, put that fish of length L1 into that class. So, basically follow the map principle and this will give me the Bayes classifier.

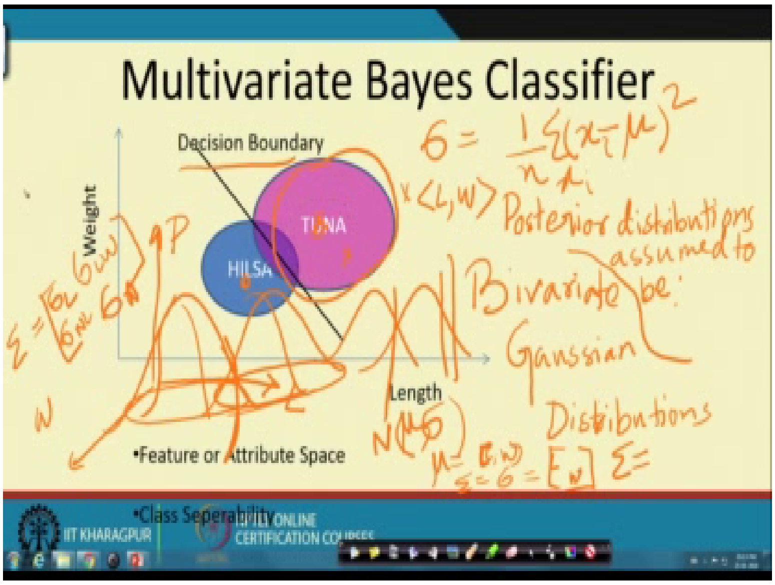
So, again to summarize if I have k class, take a population find out the class conditional distributions for each of the class, find out the prior probabilities for each of the class multiply them to get the posteriors though the posterior distribution, for each of the class for a new fish L1, classify it into a class which has the highest posterior for that value of L map, good I think you have understood up to this.

So, now you may ask the question that well you say that this is the Bayes classifier, this has the minimum error. So, why not use this always, why do I use this son tree or say support vector machine or something why not always use Bayes classifier the thing is that if I know these posterior distribution, if you know the class conditional distribution, then I can only find out the Bayes classifier of course, if somebody tells me that this is that class conditional. I will definitely use the Bayes classifier, but nobody will tell me what the class conditional distribution is what I have done in this first example, by going to the market and drawing this curve is only an estimate of this distributions of course, this estimate will be correct if I draw a large sample, but it is only an estimate it will have an error, it will it is only, it is not the it is not exactly the unknown distribution it is an estimate of the unknown distribution there are estimation errors.

So, that is the problem, so even though Bayes classifier is the best you as a precondition to applying the Bayes classifier you have to know the distributions. If you do not know the distribution you estimate it and that why there will be a some error, and because of that error it may not might not give you the optimal classifier.

So, what I will do next for one dimension I have seen that the decision boundary was as a point L star, B star suppose my classes are multivariate to bivariate distribution.

(Refer Slide Time: 19:53)



Multivariate Bayes Classifier

- Feature or Attribute Space
- Class Seperability

Decision Boundary
Weight
Length
TUNA
HILSA
$\sigma = \frac{1}{n}\sum_{i}(x_i-\mu)^2$
<L,W>
Posterior distributions assumed to be:
Bivariate Gaussian Distributions
P
W
L
$\Sigma = \begin{bmatrix}\sigma_L & \sigma_{LW}\\ \sigma_{WL} & \sigma_W\end{bmatrix}$
$N(\mu,\sigma)$
$\mu = [L, W]$
$\Sigma = \sigma = [W]$
$\Sigma =$
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

So, instead of characterizing a fish only by its length, I measure the length as well as weight. So, any fish is a point in, any fish is actually a point in two dimensional space is the point in two dimensional space, some length fell some weight fell. So, now if I take a number of TUNA, they will have some mean length weight and there will be some distribution variance around that there will be for HILSA also, let us assume that this individual distribution are in normal distributions bivariate Gaussian distributions.

So, how does bivariate Gaussian distribution actually look like? So, for example, if this is my L and this is my W and if I in the z axis if I draw my probability, they look like bells, bells aped curves, if I project them they will look like circles or ellipses. So, they will be like a bell over L W (Refer Time: 22:27) L W plane. So, I have one such bell for that TUNA note that, so what I am drawing here is actually the projection of the bell to L W space, I am looking from the top this is a picture looking from the top similarly, I have a bell for HILSA and this projections will either look like an circle, if this bell is symmetric or look like an ellipse, if the bell is elongated more in one direction than the other, now earlier I said if when the univariate case the boundary was the intersection between two distributions, but now if we put two such bell if we put to such bells their intersection is not a point, but a curve between them in L W space, L W plane if we just look at values where these two bells intersect I will get a curve.

So, this curve is actually my decision boundary what will work out is if I know the parameters of this class I have the posterior distributions and assume them to be Gaussian or normal distribution.

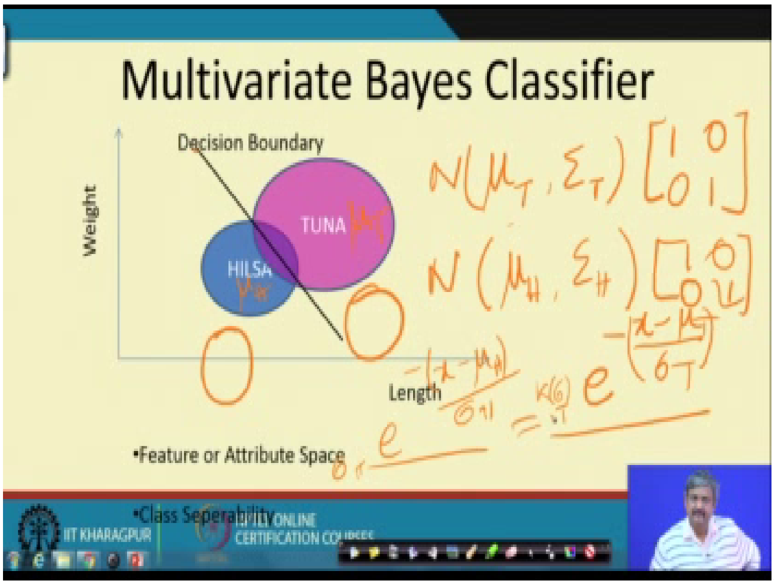
How does this decision boundaries look like, what type of curve they are you know a Gaussian distribution, a normal distribution is parameterized by two parameters the mean value and a variance value mu sigma, in the case of 2 dimensional bivariate Gaussians the mu will be nothing, but a 2 dimensional vector representing this centre it will dimensional vector and sigma will be a 2 by 2 matrix. So, this will mu will be a 2 dimension L W vector and sigma will be a, let me write it as capital sigma will be a 2 dimensional matrix.

Where the entries are the following the first entry is variance of L with itself the second entry is variance of W with itself. So, if let me define what variance is so sigma roughly is if I have n samples, it is the variation of x from the mean square of that summed over all the x. So, if you take x 1 x 2 x 3 x n. So, x i if each of this x i, I find how far it is from mu take the difference square it up edit up over all n, take the average value of this squared error, it measures the kind of spread of the distribution whereas, mu measures the central location of the distribution.

If I have a bivariate the, it is not just a sigma is no longer a scalar it is a matrix. So, you know, what does the matrix look like? Matrix will look like I am getting 30. So, the matrix will look like, it is the sigma of L mean value of length and variation of length of each fish from this mean value for TUNA class or HILSA class mean value of W, the average weight of a TUNA and how the weight of individual TUNAs differ from this average weight squared data, some data and then the covariance’s.

So, covariance is a cross term; that means, variation of length with respect to average W variation of W with respect to average L same thing, but the central cross. So, suppose somehow by estimation or something I know the mean value and sigma value of TUNA, I know mu T sigma T.

(Refer Slide Time: 28:00)



Multivariate Bayes Classifier

Decision Boundary
Weight
Length
TUNA
HILSA
μ_T
μ_H
$N(\mu_T, \Sigma_T) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
$N(\mu_H, \Sigma_H) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
$e^{-\frac{(x-\mu_H)}{\sigma_H}} = K(\sigma_T)\, e^{-\left(\frac{x-\mu_T}{\sigma_T}\right)}$
IIT KHARAGPUR
NPTEL ONLINE CERTIFICATION COURSES

- Feature or Attribute Space
- Class Seperability

And I know mu H sigma H, then I want to derive the decision boundary as a function of mu T mu I sigma T sigma H, if you remember the Gaussian distribution next it takes this particular form.

So, e to the power minus x minus mu by sigma square and then some constant term of sigma, so this is how the probability of x belonging to this Gaussian distribution, looks like. So, I will use that expression and I will derive under various conditions on mu and sigma what does this boundary look like actually I will give you as an exercise you consider two special cases, if the sigma is a diagonal matrix and unit diagonal matrix. So, basically these are unit spheres each of them both of them have same sigma, and unit diagonal matrices hence, some value of mu H and mu T, some value of mu H and mu T.

Then putting this thing where does posterior assume class conditionals are same 5.0, 5.0, 5 for both I sorry assumed 5 are same 5.0, 5.0 for both when does the posterior distribution give at what values of what is the locus of the values of x, which keeps same value of probability for both the distributions mu T sigma T mu H sigma H, I am not writing the exact form you should write it.

So, where does this to equate over which locus of x that is finding surrounded by definition of the (Refer Time: 30:26) classify. So, we will derive it for different assumptions on mu and sigma and we will see they have a nice geometrical form. So, with this assignment you should work out I end today's lecture I will continue discussion on this in my next lecture.

Thank you.