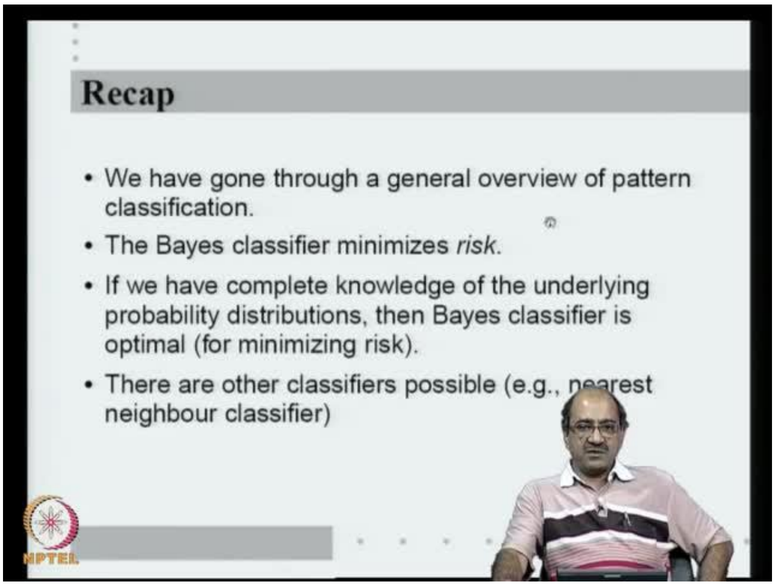
Pattern Recognition

Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 3
The Bayes Classifier for minimizing Risk

Hello, so to continue with our pattern recognition. In the first two lectures we have gone through a general overview of pattern classification. We looked at what is the problem of pattern classification, we defined what pattern classifiers are and we have looked at the two block diagram model, that is given a pattern you first measure some features. So, pattern gets convert to feature vector and then the classifier essentially maps feature vectors to class labels.

(Refer Slide Time: 01:17)



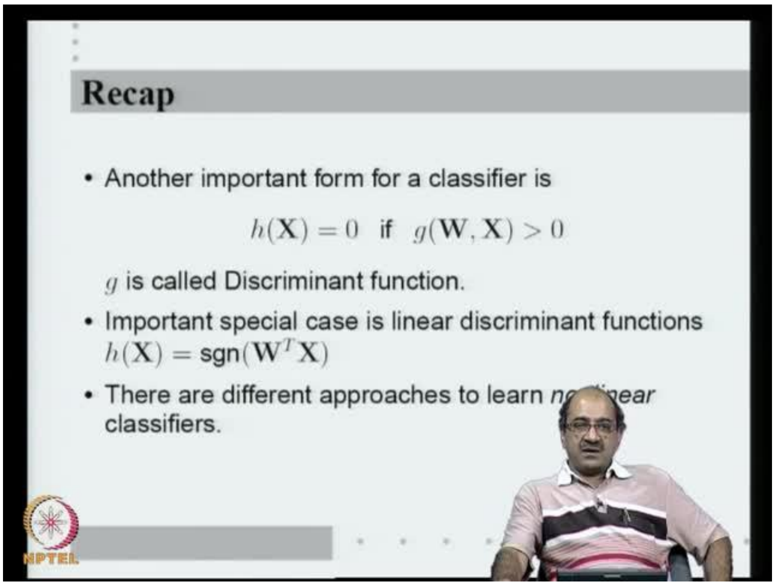
Recap

- We have gone through a general overview of pattern classification.
- The Bayes classifier minimizes *risk*.
- If we have complete knowledge of the underlying probability distributions, then Bayes classifier is optimal (for minimizing risk).
- There are other classifiers possible (e.g., nearest neighbour classifier)

So, they said the course is about classifier design, we have looked at a number of classifier design options as an overview. Now, from this lecture we will go into details. So, just to recap what we have done so far; we have gone through a general overview of pattern classification, a few points from the general overview that I would like to emphasis. So, here is a classifier that we already looked at is the Bayes classifier, the Bayes the the for the Bayes classifier we are taking a statistical view of pattern recognition. Essentially what it means is that a feature vector is is is a essentially random.

So, the variations in feature values when you measure pattern from the same class are captured through probability densities. And given all the underlying class conditional densities we we see in that base classifier minimises risk, we saw the proof for only minimising problems classification we will see the general proof this class. So, Bayes classifier essentially puts a pattern in the class for which the posterior probability is maximum and it minimises risk. If you have the complete knowledge of the underlying probability distributions then Bayes classifier is optimal for minimising risk.

(Refer Slide Time: 02:15)



Recap

- Another important form for a classifier is

$$h(\mathbf{X}) = 0 \quad \text{if} \quad g(\mathbf{W}, \mathbf{X}) > 0$$

g is called Discriminant function.

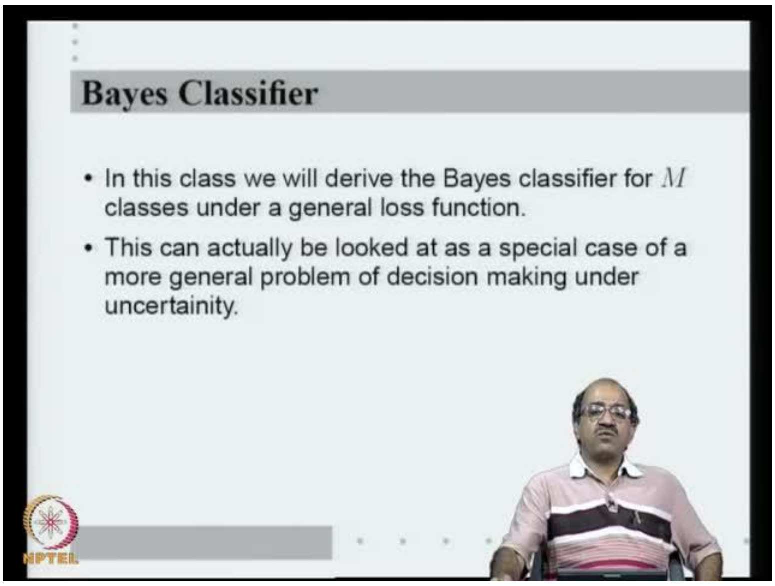
- Important special case is linear discriminant functions $h(\mathbf{X}) = \text{sgn}(\mathbf{W}^T\mathbf{X})$
- There are different approaches to learn *nonlinear* classifiers.

There are other classifiers for example, we seen nearest neighbour classifier, among the other classifier of course, nearest neighbour classifier we will come back to again. But one other classifier that we have seen which I would like to emphasis again or a discriminant function based classifiers; that is the classifier. We we use h for the classifier function so h of X is 0 in a two class case that is X is put in class 0. If some other function g W comma X is greater than 0 W is the parameter vector for the g function, g is called the discriminant function. So, we can design a function a discriminant function g or find the appropriate values for the parameter vector W among a class of discriminant functions. And a classifier of this kind h X is equal to 0 if g W comma X is greater than 0, is called a discriminant function based classifier.

Sometimes it is simply called a discriminant function or we say the classifier is the discriminant function though g is actually the discriminant function. So, if a classifier is based on discriminant function, we call a discriminant function or a classifying function based classifier and so on. A special case of discriminant functions is a linear discriminant function, which also we considered where h of X is simply sign of W transpose X or as I said we normally take an augmented vector X. So, that the constant in the linear form is incorporated.

So, this is essentially stand for W summation of W X a plus a constant W naught, which can be viewed as a simple (( )) product W transpose X. We always having a extra component one in the feature vector X that is called an augmented feature vector, then the W vector contains also the constant. So, essentially when g is a linear function like this and h X is 0 of g X is greater than 0 otherwise. So, I can essentially think of h of X as sin of W transpose X and such a classifier is called a linear discriminant function. We have we have seen this also in our first two classes and this is another important structure for a classifier. We have also seen different approaches for learning non-linear classifiers. So, we will consider all of them through this course.

(Refer Slide Time: 04:22)

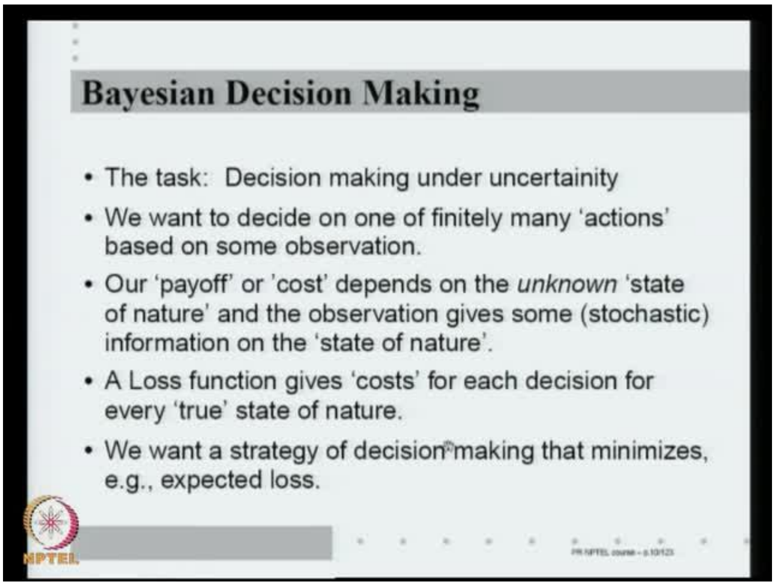


Bayes Classifier

- In this class we will derive the Bayes classifier for M classes under a general loss function.
- This can actually be looked at as a special case of a more general problem of decision making under uncertainity.

Now, in this lecture we will start by giving you more details on the Bayes classifier. So, we will derive the Bayes classifier, where as general M class case and for general loss function. Not a 0 1 loss function, earlier we looked at it for 0 1 loss function and two class case. Now, we look for a generic M class classifier under a very general loss function. Before I start with the base classifier this actually is a special case of what is called as Bayesian decision making or the problem of decision making under uncertainty. Since, this is more generally applicable is worth whiles pending a bit of time looking at what decision making problem is about.

(Refer Slide Time: 05:04)



Bayesian Decision Making

- The task: Decision making under uncertainity
- We want to decide on one of finitely many 'actions' based on some observation.
- Our 'payoff' or 'cost' depends on the *unknown* 'state of nature' and the observation gives some (stochastic) information on the 'state of nature'.
- A Loss function gives 'costs' for each decision for every 'true' state of nature.
- We want a strategy of decision making that minimizes, e.g., expected loss.

PR NPTEL course – p.10/123

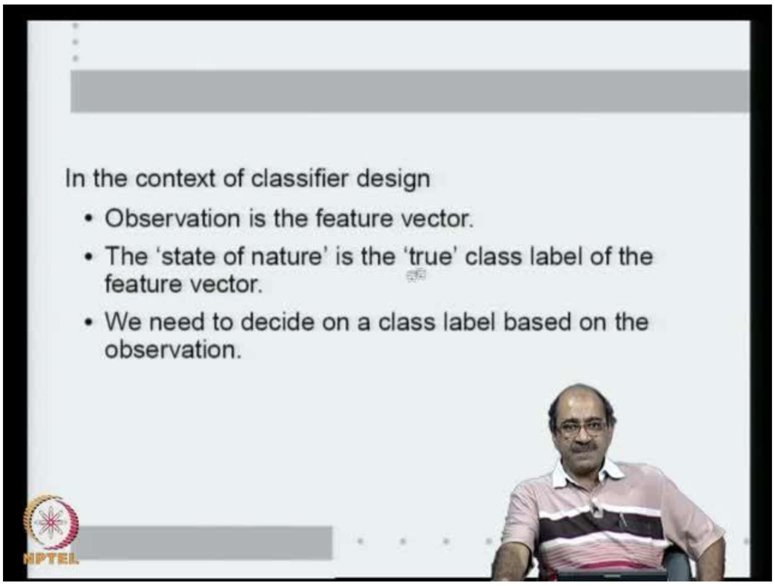
In a decision making problem the task of course, obviously is to make a decision, but the reason why it becomes a problem is that there is uncertainty about what is the right decision. What is the form in which the uncertainty comes up? Essentially we want to decide on of finitely many actions, based on some observation. For example, in a pattern classification problem, for in many other situations you you observe the state of some system and based on that you should take an action. Say for example, a control problem is also decision making problem you you look at the current output at the plant and then based on that you have to take some control action.

In the pattern classification cant, for example; I look at the radar reflection signal and based on that observation, how to make a decision of whether there is an enemy aircraft coming or not or if I am thinking of identity authentication as a classification problem. I look at a fingerprint image and an identity claim that is my observation based on that I have to take an action yes or no. The uncertainty is because the cost of my decision, depends on some unknown state of nature. So, in the in the radars example actually out there either there is a enemy aircraft or there is no enemy aircraft that is the true state of nature. But that state of nature is a unknown to me. What I have is some observation namely my radar reflector signal, which gives me some information not necessarily exact 100 percent dependable information, but some stochastic information on the true state of nature.

So, based on this observation that gives me some information on the state of nature, I have to make my action. And for a given action my payoff or cost depends on the actual state of nature which is unknown to me. So, once again in this example of deciding whether or not there is an enemy aircraft, based on my observation of the radar signal I make a decision of sounding a alarm for bombing.

Then whether that is right or wrong and hence, the cost incurred depends on truly whether or not there is an enemy aircraft. A loss function gives cost for each decision and every true for true state of nature, that is if I take this decision and the true state of nature happens to be something, then I incur some cost. Since, I do not know the true state of nature I only have the observations we need some strategy of decision making that minimises some objective function for example, expect a value of loss. This is the kind of scenario we are in.

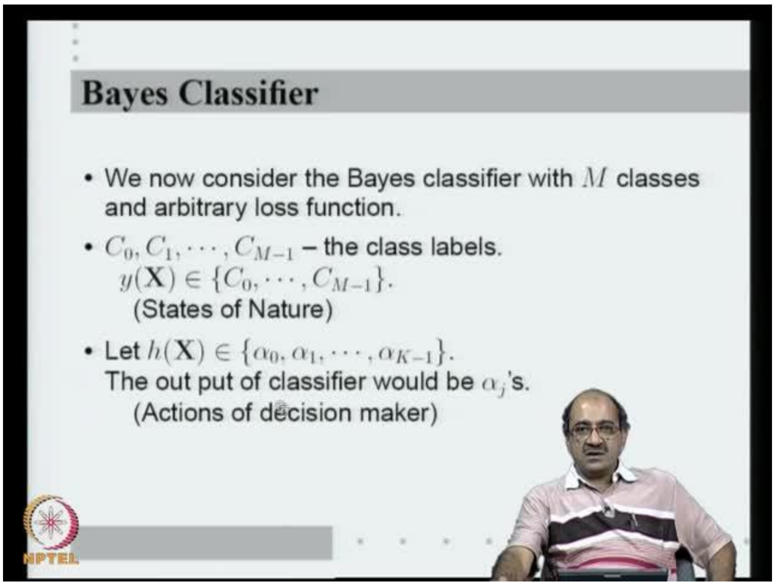
(Refer Slide Time: 07:57)



In the context of classifier design

- Observation is the feature vector.
- The 'state of nature' is the 'true' class label of the feature vector.
- We need to decide on a class label based on the observation.

(Refer Slide Time: 08:25)



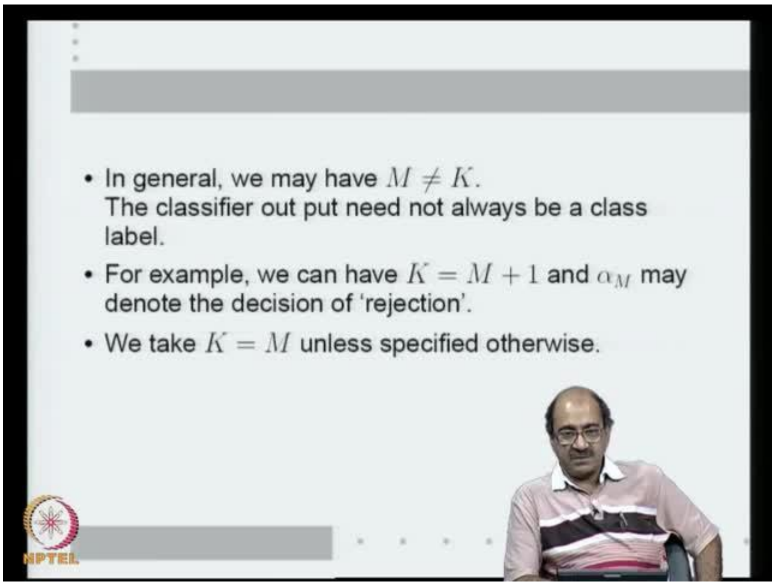
Bayes Classifier

- We now consider the Bayes classifier with M classes and arbitrary loss function.
- $C_0, C_1, \cdots, C_{M-1}$ – the class labels.
  $y(\mathbf{X}) \in \{C_0, \cdots, C_{M-1}\}$.
  (States of Nature)
- Let $h(\mathbf{X}) \in \{\alpha_0, \alpha_1, \cdots, \alpha_{K-1}\}$.
  The out put of classifier would be α_j's.
  (Actions of decision maker)

So, to sum up observation is the feature vector, the state of nature is the true class label for the feature vectors presented to be. So based on my observation of the feature vector I have to take an action, the action may be calling the class label or to say I cannot classify or many other things, but most of the time actions are simply class labels. So, we need to decide on a class label based on the observation the feature vector. So, with this general interaction we now consider a Bayes classifier M classes and arbitrary loss function. So, let us denote the M classes by C 0, C 1, C M minus 1, these are the class labels. So, what is that means y of X there is a variable been used into denote the class label of a the true class label of a feature vector X. Because, of random variability's y as a function of X could be random.

So, but this this variable y, which is a function of X can take values as one of the class labels, C 0 to C M. Let us say the action that are quite possible for the classifier or alpha 0, alpha 1, alpha K minus 1. Because, in general K will be M and alphas will be same as C, but let us just for now keep the notation different. So, the output of a classifier would be one of the alpha's that is why I said h of X belongs to alpha 0 to alpha K minus 1. So, the output of classifiers would be alpha s whereas, the actual class labels will be C s. The reason why we put different symbols is that essentially we want to think of the class labels at the states of nature, which is not known to us. And the alphas are the actions of decision maker alphas are what the classifier calls out.

(Refer Slide Time: 10:38)



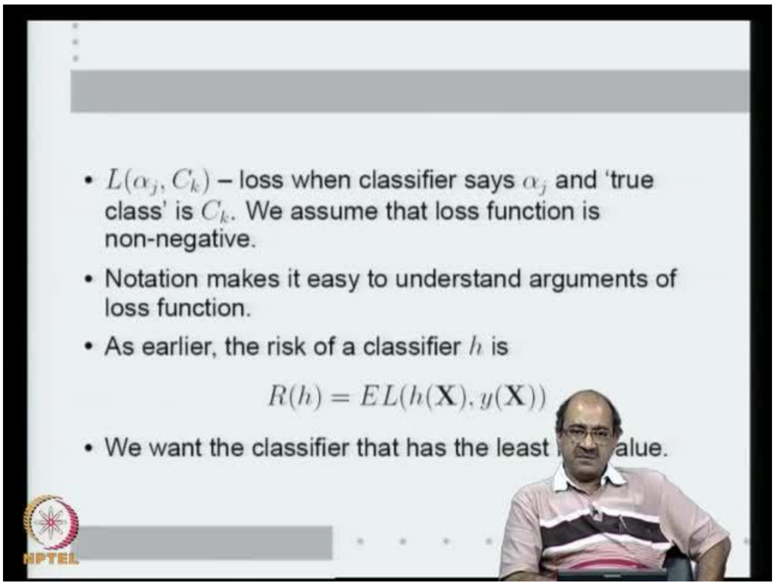
- In general, we may have $M \neq K$. The classifier out put need not always be a class label.
- For example, we can have $K = M + 1$ and α_M may denote the decision of 'rejection'.
- We take $K = M$ unless specified otherwise.

In general of course, alphas can be class labels, but is important to note that in general we do not have to have M equal to K. So, we we put K actions K possible actions that is K possible outputs for the classifier and M possible class labels. So, not only alphas need not have to be class labels even K and M need not be same. It is a simple example of why am I want to do it, for example, I may take K is equal to M plus 1 or in the other alphas. I make alpha 0 same as C 0 alpha 1 same as C 1.

So, alpha 0, alpha 1 up to alpha M minus 1 will be class label that is I am I am calling over the class. But I have one more extra action possible for the classifier right, will be alpha M alpha M minus 1, M will be alpha M and alpha M it denote the decision of rejection. So, for example, I can design a classifier which looking at a looking at a class label, it can say that yeah I will take this looking at a feature vector. I can say this is this class or say that no this feature vector does not look sufficient for me, I cannot take a decision.

For example, if I am doing a identity authentication system, somebody presents a fingerprint under identity claim. I can either say yes is a authorised user or say no he is not an authorised user or I can say this image does not look nice. So, I do not want to take a decision. So, in practice what it turns out to be that may be the image was not captured well. So, tell the user can you please wipe your finger and you know give your finger print again. So, sometimes in in certain situations having such a rejection option is useful. So, this is one example why the output of classifier may be different from the class label, they might be few extra actions for the classifier. But any way for now when we are discussing the Bayes classifier we take K is equal to M and we assume that the output of classier is also class labels.

(Refer Slide Time: 11:54)



- $L(\alpha_j, C_k)$ – loss when classifier says α_j and 'true class' is C_k. We assume that loss function is non-negative.
- Notation makes it easy to understand arguments of loss function.
- As earlier, the risk of a classifier h is

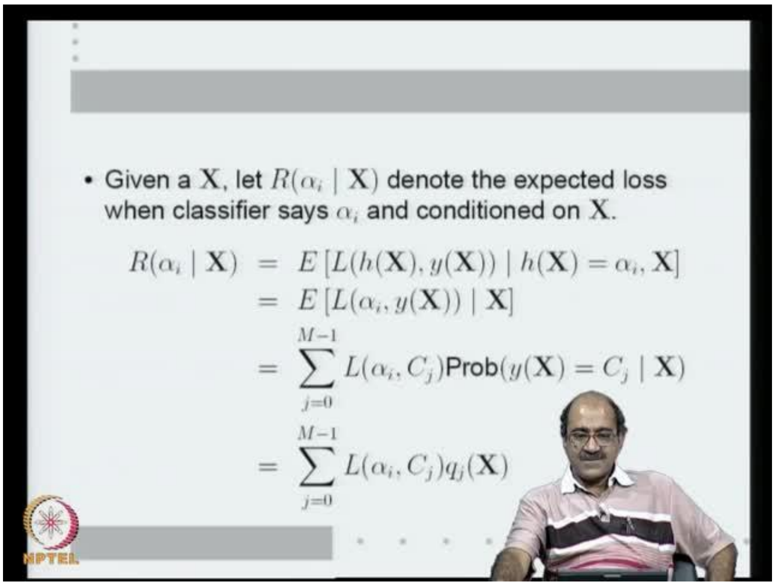
$$R(h) = EL(h(\mathbf{X}), y(\mathbf{X}))$$

- We want the classifier that has the least alue.

So, let the last function L as we have already seen is two arguments; one is what the classifier says and one is the true class. So, L alpha j, C k is the loss when the classifiers say alpha i and the true class is C k. We assume that the loss function is always non negative. Now, earlier we are saying L j comma k in this particular derivation we would saying L alpha j comma C k. So, that we know the first argument differs to the output of the classifier and second argument refers to the true class.

So, this kind of notation is a little easier to understand the arguments. As earlier risky risk of a classifier h is the expectation of the loss. So, each classifier h to each classifier h I can assign a figure of merit, which I called risk, which I write R of h, which is expectation of L of h X comma y X, where the expectation with respect to the distribution of x and whatever randomness they may be in y of X. So, expectation of loss is the risk and the object to the base classier is to find a classifier at the rule h that minimises the risk.

(Refer Slide Time: 13:07)



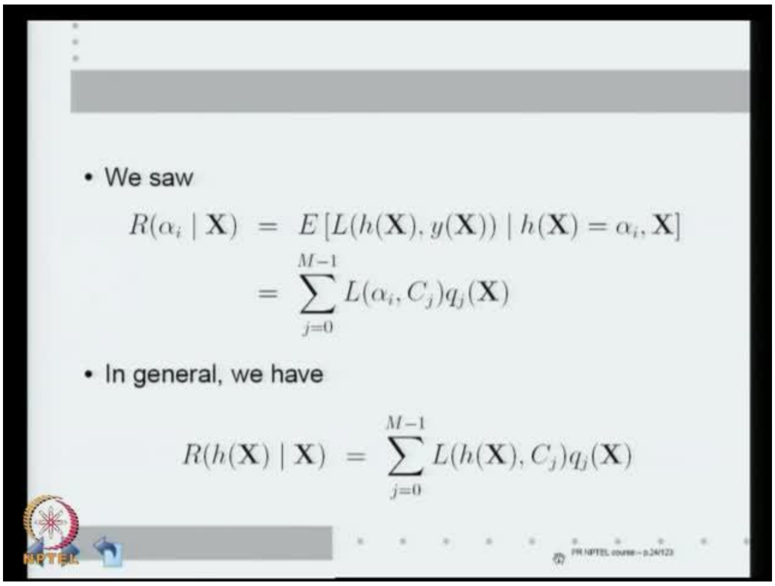
- Given a $\mathbf{X}$, let $R(\alpha_i \mid \mathbf{X})$ denote the expected loss when classifier says α_i and conditioned on $\mathbf{X}$.

$$\begin{aligned} R(\alpha_i \mid \mathbf{X}) &= E\left[L(h(\mathbf{X}), y(\mathbf{X})) \mid h(\mathbf{X}) = \alpha_i, \mathbf{X}\right] \\ &= E\left[L(\alpha_i, y(\mathbf{X})) \mid \mathbf{X}\right] \\ &= \sum_{j=0}^{M-1} L(\alpha_i, C_j)\text{Prob}(y(\mathbf{X}) = C_j \mid \mathbf{X}) \\ &= \sum_{j=0}^{M-1} L(\alpha_i, C_j) q_j(\mathbf{X}) \end{aligned}$$

Now, to derive the base classifier let us say given a particular feature vector X lets denote the R alpha i given X. The expected loss when the classifier it says i conditioned on X, that is given X undefined say alpha X what is my expected loss. So, what does R alpha given X, correspond to R alpha given X is expectation of the loss that is L of h X comma y X. But it is not in a unconditional expectation, it is an conditional expectation, conditioned on h X equals to alpha i that is classifier say alpha i unconditioned of the random variable X. So, R alpha i given X is expectation of L of h X comma y X, condition on h X is equal to alpha i and X. In a conditional expectation if any of the condition random variable say pair in the expectation they can be replaced by whatever their condition.

So, this same as conditional expectation of h X can be replaced with alpha i expected value of L alpha i comma y X condition now only on X. Now, what is this expectation, this expectation of the random variable L alpha i comma y X with respect to the distribution conditioned on X, y X can take values C 0, C1 up to C m minus 1 with different probabilities. So, I can write this conditional expectation as this L of i comma y X will take value L alpha i comma C j with probability y X equals to C j given X. When I sum it over j is equal to 0 m minus 1 that gives me the conditional expectation. Now, we know what probability y X equal to C j given X is that is what we call the posterior probability for which you are already have a notation q j of X. So, R alpha i given X now is summation j equals to 0 at m minus 1, L of alpha i comma C j, q j X.

(Refer Slide Time: 15:10)



- We saw

$$R(\alpha_i \mid \mathbf{X}) = E\left[L(h(\mathbf{X}), y(\mathbf{X})) \mid h(\mathbf{X}) = \alpha_i, \mathbf{X}\right]$$
$$= \sum_{j=0}^{M-1} L(\alpha_i, C_j) q_j(\mathbf{X})$$

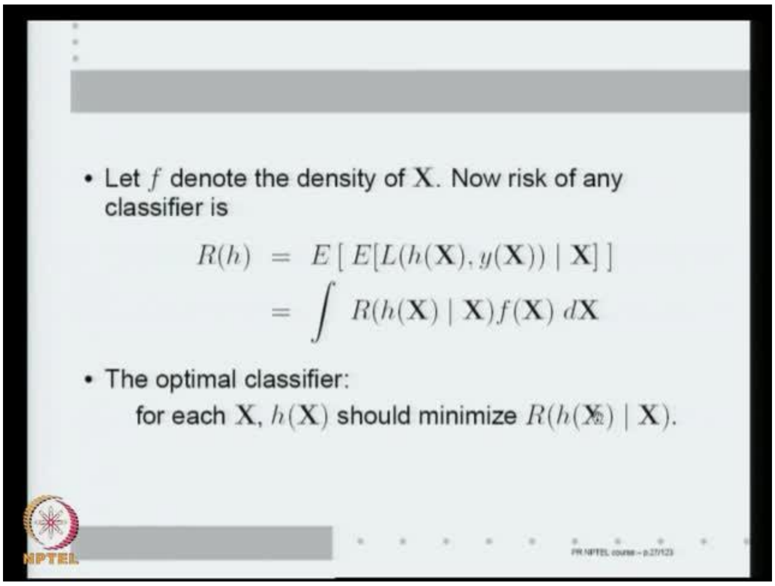
- In general, we have

$$R(h(\mathbf{X}) \mid \mathbf{X}) = \sum_{j=0}^{M-1} L(h(\mathbf{X}), C_j) q_j(\mathbf{X})$$

PR NPTEL course – p.24/123

Now, using this we can now find a find an expression for the true risk. What will be the true risk? We have seen that R alpha i given X is expected value of L of h X comma y X, conditioned on h X equal of y comma X for which you derived this expression, L of alpha i comma C j, q j X summed over j is equal to 0 to m minus 1. So, this is the case for L alpha i given X. So, in general we can also write this equation as R of h X given X is the same summation where alpha is replaced with h X. So, we can write R of h X given X as summation j is equal to 0 to m minus 1, L of h X comma C j, q j X. So, this is the expression that we are going to use in deriving our final expression for the risk. Once we derive the final expression for risk then we will see how to minimise this.

(Refer Slide Time: 15:59)



- Let f denote the density of $\mathbf{X}$. Now risk of any classifier is

$$
\begin{aligned}
R(h) &= E\,[\,E[L(h(\mathbf{X}), y(\mathbf{X})) \mid \mathbf{X}]\,] \\
&= \int R(h(\mathbf{X}) \mid \mathbf{X}) f(\mathbf{X})\, d\mathbf{X}
\end{aligned}
$$

- The optimal classifier:
  for each $\mathbf{X}$, $h(\mathbf{X})$ should minimize $R(h(\mathbf{X}) \mid \mathbf{X})$.

PR NPTEL course – p.27/123

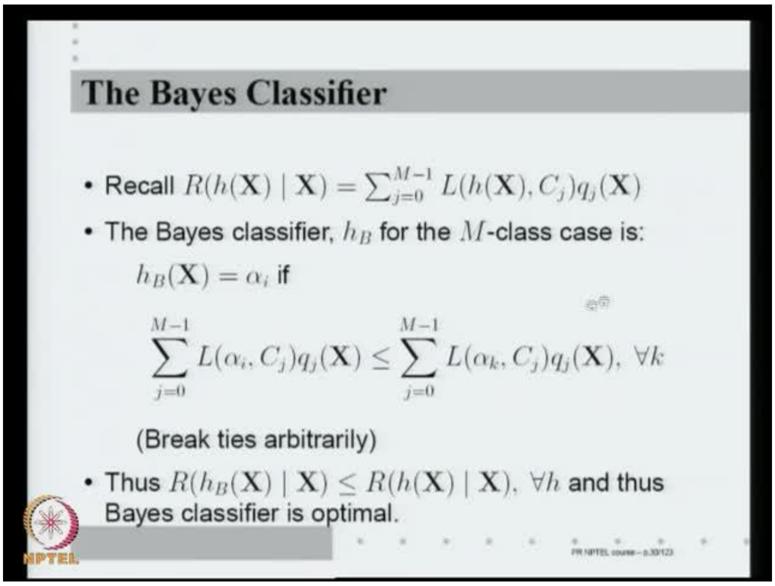
So, risk as we know is the expectation of L h X comma y X. Now, as you know any expectation can be written as expectation of a conditional expectation. So, that is how we have written R h here the the angle expression is expected expectation of L of hX comma y X conditioned on X. Now, if I take another expectation it becomes the unconditional expectation of L h X comma y X. So, I have written the unconditional expectation which is the risk R h as expectation of a conditional expectation, what is the reason for that the reason for that is I already have an expression for the general conditional expectation. General conditional expectation because the conditional expectation is only a function of X, I have an expression for this.

Now, the outer expectation is the unconditional expectation of some function of X. So, I know how to do it because I know the density for X so let us expand this expression. So, the inner conditional expectation which is conditional expectation L h X comma y X condition on X is what we call R of h X given X and the outer expectation is simply expectation of this as a function of X. So, the total expectation becomes integral R h X given X into f X, d X where f X is the density function for X. This is just the conditional expectation integral, because if X happens to be discrete it will become summation, but in general let us use integrals.

So, so that where essentially considering X all X is to be continuous random variable. So, this is the expression for true risk so from this expression there is something that we can say, if there is a h if we can find a classifier h, such that R of h X given X is less than R of h prime X given X for every X and any other h prime. So, h that minimises every term in this integral, minimises the integrand for each X would be the optimal, optimal classifier. Mind you a minimiser of R h does not necessarily have to at have to have the least value for every X, minimiser for R h only means the entire integral should be minimised by changing the h, but in case there is a h which takes least value for each X.

So, that R h X given X is is small for every X then that will certainly be a minimiser of X. So, let us look for an optimal classifier by asking for each X, h X should minimise R h X given X. So, we want to choose a classifier h such that for every X, h of X is such that R h of X given X is smaller than R of anything else given X. So, that that is certainly will be an optimal classifier and let us ask is there such an optimal classifier that be confined.

(Refer Slide Time: 19:06)



The Bayes Classifier

- Recall $R(h(\mathbf{X}) \mid \mathbf{X}) = \sum_{j=0}^{M-1} L(h(\mathbf{X}), C_j) q_j(\mathbf{X})$
- The Bayes classifier, h_B for the M-class case is:

  $h_B(\mathbf{X}) = \alpha_i$ if

  $$\sum_{j=0}^{M-1} L(\alpha_i, C_j) q_j(\mathbf{X}) \leq \sum_{j=0}^{M-1} L(\alpha_k, C_j) q_j(\mathbf{X}), \ \forall k$$

  (Break ties arbitrarily)
- Thus $R(h_B(\mathbf{X}) \mid \mathbf{X}) \leq R(h(\mathbf{X}) \mid \mathbf{X}), \ \forall h$ and thus Bayes classifier is optimal.

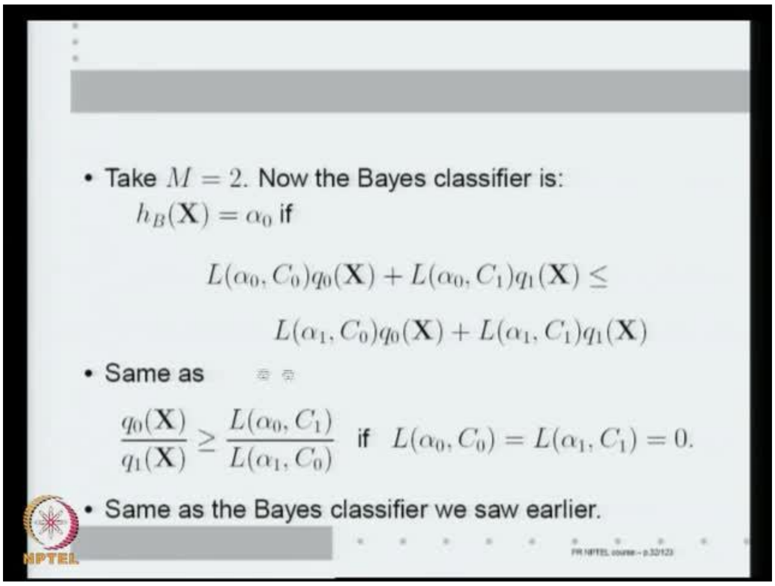
PR NPTEL course – p.30/123

So, this is the this is the expression for h of X. So, if I want to say h of X is equal to alpha i if I put h of X is equal to alpha i this, then that should be smaller than what any other classifier can assign to X. Now, every classifier all it can do is given an X it has to assign 1 of the classes to it and we have we have we have we are denoting the classes assigned by the classifier by alphas, which brings us to the following definition. The best classifier will be 1 which assigns, let us define the classifier h B as follows given an X h B of X assigns alpha i. If L of alpha i comma C j, q j X summed over j, is less than or equal to L of alpha k comma q j, C j summed over j. This will immediately mean h B is optimal why because h B assigns alpha i, what out to X, what any other classifier can do. Is assign one of the other alphas to X that is all any classifier can do.

So, compared to assigning any other alpha assigning alpha i is better, because by assigning alpha i, h of X given X it becomes smallest. So, if such a classifier would certainly minimise this integral because they for every X that classifier has the least value for all h X given X. So, this is known as Bayes classifier. So, Bayes classifier is now simply defined by, the Bayes classifier assigns alpha i to X, if L alpha is C j, q j some dou j is less than or equal to L alpha k, C j, q j some dou j for all alpha k for all k. So of course, there can be more than one alpha which may attain the same minimum value, but that makes no difference we can break ties arbitrarily. By arbitrarily I mean by can have a consistent policy if alpha 2 and alpha 3 both attain the same minimum value. Then I will choose alpha 2 saying the one with the smallest index or I can choose alpha 3, say in one with the largest index.

Any such arbitrary way of breaking ties is good enough because that will make my classifier a function. Given any X it unambiguously attains a classifier and the way we constructed it, it is clear that R h B of X given X is always less than equal to R of h X given X for all h. Because, as you said h B gives alpha i and but any other h can do is give one of the alpha case and no matter what alpha k this h gives. R h X given X for the Bayes classifier is less than or equal to what any other classifier can achieve. There is R h B of X given X is less than R h X given X for all h and all X and thus Bayes classifier is the optimal. So, this is the general optimal Bayes classifier for minimising risk.

(Refer Slide Time: 22:25)



- Take $M = 2$. Now the Bayes classifier is:

  $h_B(\mathbf{X}) = \alpha_0$ if

$$L(\alpha_0, C_0)q_0(\mathbf{X}) + L(\alpha_0, C_1)q_1(\mathbf{X}) \leq$$
$$L(\alpha_1, C_0)q_0(\mathbf{X}) + L(\alpha_1, C_1)q_1(\mathbf{X})$$

- Same as

$$\frac{q_0(\mathbf{X})}{q_1(\mathbf{X})} \geq \frac{L(\alpha_0, C_1)}{L(\alpha_1, C_0)} \text{ if } L(\alpha_0, C_0) = L(\alpha_1, C_1) = 0.$$

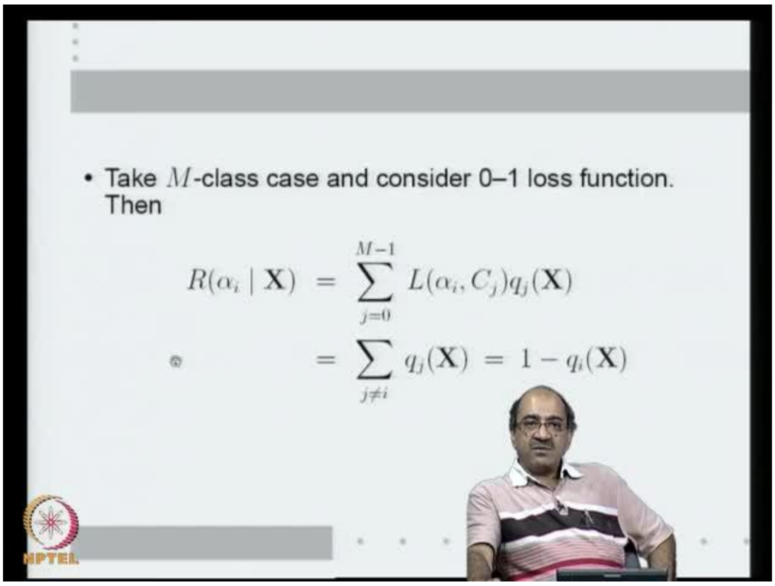
- Same as the Bayes classifier we saw earlier.

PR NPTEL course – p.32/123

So, let us take some special cases we will start from this expression and explore Bayes classifier for many special cases. Take M is equal to 2, if there is only 2 then the summation has only two terms; one for C 0, one for C 1 and h also has only two values alpha 0 and alpha 1. So, what does the Bayes classifier now say h B X will say alpha 0 if L of alpha 0, C 0, q 0 plus L of alpha 0, C 1, q 1 is less than or equal to L of alpha 1 C 0, q 0 plus l of alpha 1, C 1, q 1, where this assigning alpha 0 to X incurs less risk then assigning alpha 1 to X. Now, let us go back to our old loss function where we normally put Ll alpha 0, C 0 to be 0, L alpha 1, C 1 to be 0 because if we take the correct decision there is no loss.

So, if we ensure L alpha 0, C 0 is equal to L, alpha 1 C 1 is equal 0 then what does this mean this term will drop out this term will drop out. So, if I bring q 1 this side this expression is same as q 0 by q 1 is greater than or equal to L of alpha 0, C 1 by L of alpha 1, C 0. This is the Bayes classifier for two class general loss function that we specified last in the last lecture that time we said we will prove it later. So, from our general proof for optimal Bayes classifier, we see that for the two class case this is the optimal Bayes classifier, this is same as what we saw earlier. Now, this completes the proof of that expression also.

(Refer Slide Time: 24:10)

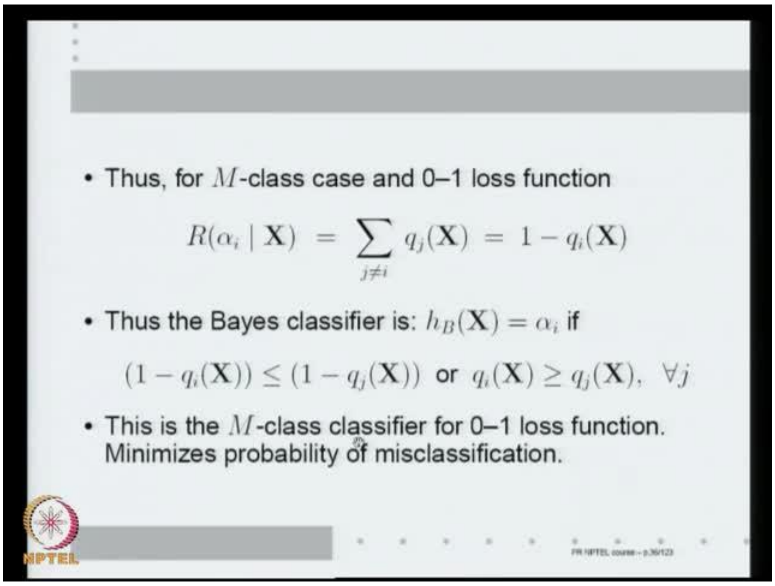


- Take M-class case and consider 0–1 loss function. Then

$$R(\alpha_i \mid \mathbf{X}) = \sum_{j=0}^{M-1} L(\alpha_i, C_j) q_j(\mathbf{X})$$
$$= \sum_{j \neq i} q_j(\mathbf{X}) = 1 - q_i(\mathbf{X})$$

To consider another loss function another special case let us take the M-class case, but now, instead of general loss function, consider a 0-1 loss function. Once again this is general expression R alpha i given X is this. Now, if i consider 0-1 loss function L alpha is C j is 1, if alpha is not equal to C j and 0 if alpha is equal to 0. Now, alpha is are same as C j's for us where one can considering the classifier action same as, whenever alpha is is not equal to C j the loss is 1. So, only those terms contain and alpha is equal to C j the loss is 0. So, this expression becomes summation over all j not equal to y 2 j of X because for a given X has to be one of the classes all the all the posterior probabilities together should sum to 1. So, sum over j not equals to q j X is nothing but 1minus q i X.

(Refer Slide Time: 25:19)



- Thus, for M-class case and 0–1 loss function

$$R(\alpha_i \mid \mathbf{X}) = \sum_{j \neq i} q_j(\mathbf{X}) = 1 - q_i(\mathbf{X})$$

- Thus the Bayes classifier is: $h_B(\mathbf{X}) = \alpha_i$ if

$$(1 - q_i(\mathbf{X})) \leq (1 - q_j(\mathbf{X})) \text{ or } q_i(\mathbf{X}) \geq q_j(\mathbf{X}), \ \forall j$$

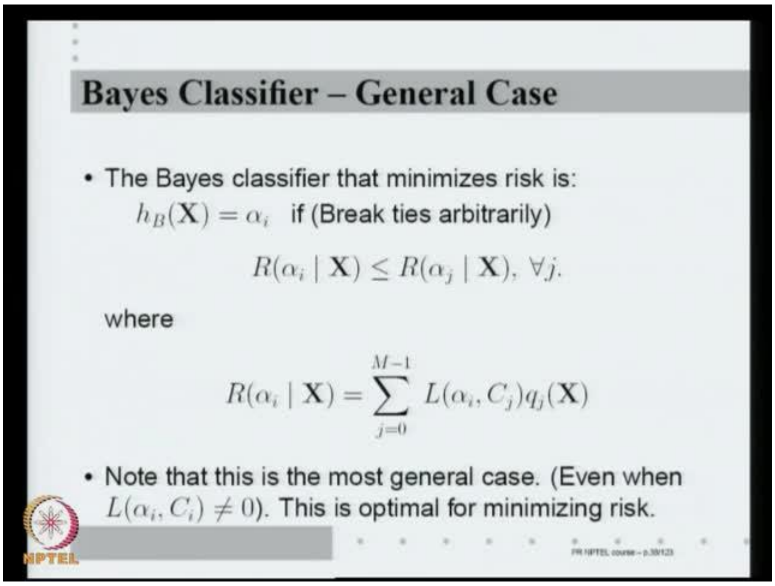
- This is the M-class classifier for 0–1 loss function. Minimizes probability of misclassification.

PR NPTEL course – p.36/123

So, what does this mean for the M-class case and 0-1 loss function, my Bayes classifier is for the M-class case 0-1 loss function. We just now shown that r alpha i given X is 1 minus q i X, where the Bayes classifier is simply is if 1 minus q i X is less than 1 minus q j X for all j, then put X in alpha i. One minus q i less than q j is same as q i greater than q j. So, even for M-class case if we have 0-1 loss function all that Bayes classifier does is assigned X to that class, which has the highest posterior probability. So, this is very straight forward we seen that in two class that is what we have been doing, if you have a 0-1 loss function, you ask whichever class has the highest posterior probability put it there.

So, even for M-class case if we use a 0-1 loss function what what our general optimal Bayes classifier tells us is that we have to assign X to the class with highest posterior probability. As I mentioned in the first lecture, this is the most obvious thing to do, for given X if I can calculate the probability that X comes from class 1 X comes from class 2 X come from class 3 M so on. Then whichever probability is higher I should put X there. So, our intuition is correct and that happens to be the optimal Bayes classifier. This is the M-class classifier for 0-1 loss function which minimizes the probability of misclassification.

(Refer Slide Time: 26:44)



Bayes Classifier – General Case

- The Bayes classifier that minimizes risk is:

  $h_B(\mathbf{X}) = \alpha_i$ if (Break ties arbitrarily)

$$R(\alpha_i \mid \mathbf{X}) \leq R(\alpha_j \mid \mathbf{X}), \forall j.$$

where

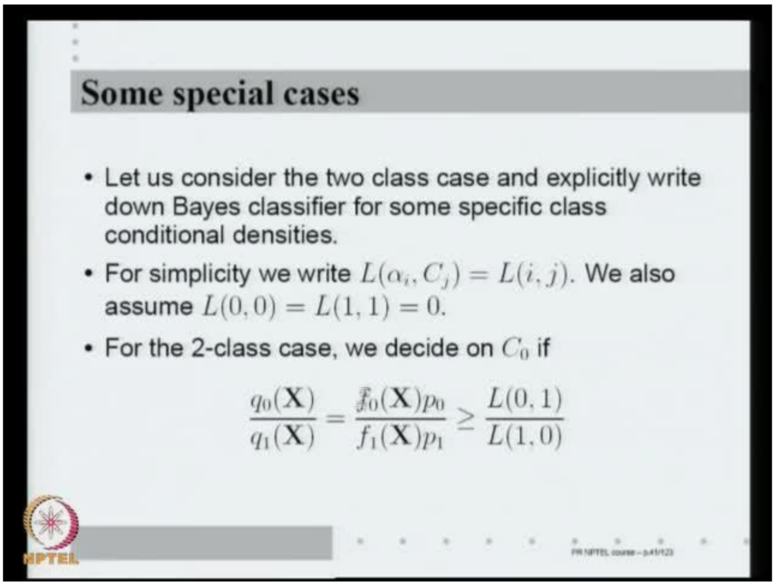
$$R(\alpha_i \mid \mathbf{X}) = \sum_{j=0}^{M-1} L(\alpha_i, C_j) q_j(\mathbf{X})$$

- Note that this is the most general case. (Even when $L(\alpha_i, C_i) \neq 0$). This is optimal for minimizing risk.

These two special cases not withstanding this is the general, we have already derived this. So, the Bayes classifier that minimizes risk, under any general loss function is that given an X the Bayes classifier assigns alpha i to X that is h B of X is alpha i. If R alpha i given X less than or equal to R alpha j given X for all j, where all alpha i given X is given by j is equal to 0 time minus 1, L alpha C j, q j X. So, if I know all the posterior probability if you can somehow calculate all the posterior probabilities given X, then I can calculate R alpha i given X. And hence, once I can calculate E alpha i given X for all alpha i, I know which action to take that is how I will actually implement the Bayes classifier. So, to implement the Bayes classifier the statistical information need to know is the the posterior probabilities of course, I need to know the loss function.

Then give me any X for each of the alpha X I will calculate this expression which gives me R alpha i given X. Then I will I will find out for which alpha is the minimum and that is the class into which I will put X. I hope you understand the only reason we we we are writing alpha for classifier output since, C j for the two class is than in the loss function R given we know which is which otherwise, alphas are same as c j's for us here. So, I can implement then general Bayes classifier if I know all the posterior probabilities. So, this is the most general case where are not even assuming that R L of alpha comma C j is not equal to 0. Even if there is not equal to 0 even then we can calculate based on this expression. So, this is the optimal Bayes classifier for minimising risk for any general loss function.

(Refer Slide Time: 28:31)



Some special cases

- Let us consider the two class case and explicitly write down Bayes classifier for some specific class conditional densities.
- For simplicity we write $L(\alpha_i, C_j) = L(i, j)$. We also assume $L(0, 0) = L(1, 1) = 0$.
- For the 2-class case, we decide on C_0 if

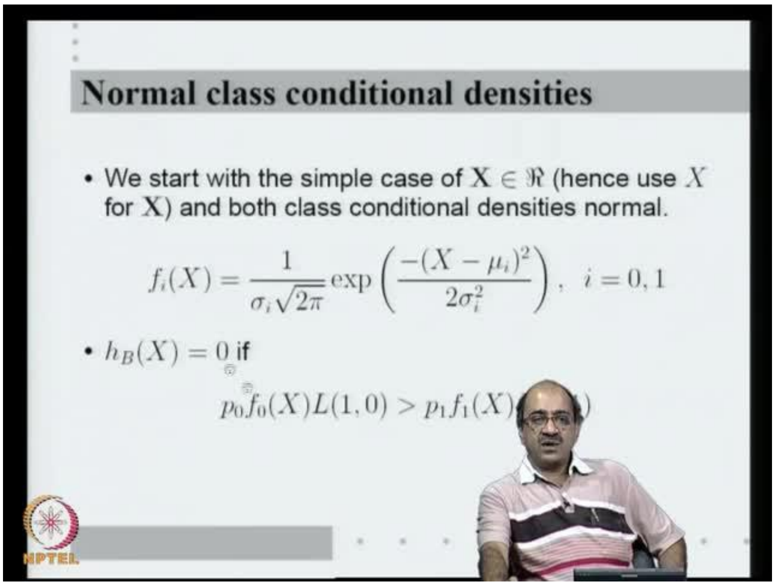
$$\frac{q_0(\mathbf{X})}{q_1(\mathbf{X})} = \frac{f_0(\mathbf{X})p_0}{f_1(\mathbf{X})p_1} \geq \frac{L(0, 1)}{L(1, 0)}$$

PR NPTEL course – p.41/123

Now, what we will do is we consider the two class case, take some specific class conditional densities and try to look at what the Bayes classifier looks like. That gives us some more insight into how to calculate Bayes classifier in specific instances. So, from now on instead of writing L alpha a comma C j we will simply write as L i comma j by now, we have got induced to understanding that the first argument is the classifier output and second argument is the class label. So, we will write L alpha i comma c j's is L i, j and we also assume that the correct classification is 0 loss.

So, for your two class case this is the Bayes classifier, we know that q 0 by q 1 by Bayes theorem is same as f 0 X p 0 by f 1 X p 1. Where f 0 is the class conditional density for class 0 and f 1 is the class conditional density for class 1. p 0 is the prior probability of a class 0, p 1 is the prior probability of a class 1, q j by q 1 is same as f 0 by p 0 by f 1, p 1. So, for the two class case we decide on class C 0 if f 0, p 0 by f 1, p 1 is greater than L 0, 1 by L 1, 0. So, given the loss function values, given the class conditional density prior probabilities, we can always calculate. But now, we are going to do is you assuming specific functional form for f 0 and f 1, we will actually crunch this expression to see how such a classifier looks like.

(Refer Slide Time: 30:03)



Normal class conditional densities

- We start with the simple case of $\mathbf{X} \in \Re$ (hence use X for $\mathbf{X}$) and both class conditional densities normal.

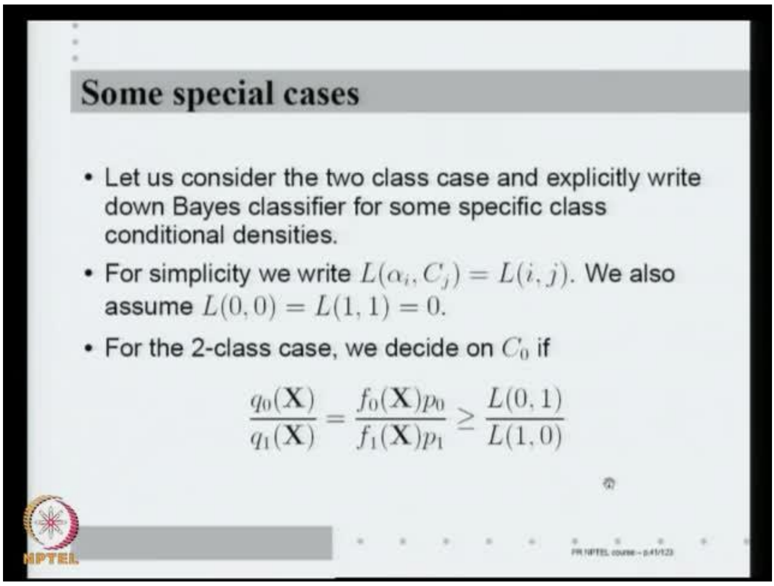
$$f_i(X) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(\frac{-(X-\mu_i)^2}{2\sigma_i^2}\right), \quad i = 0, 1$$

- $h_B(X) = 0$ if

$$p_0 f_0(X) L(1,0) > p_1 f_1(X)$$

So, we will start with this simple case we will assume that the feature vector is one dimensional that is a single day number. That only one feature and assume that both class conditional densities are normal. So, f i of X is 1 by sigma a root 2 by exponential minus X minus mu whole square by 2 sigma a square, this is the one dimensional normal density. So, f 0 is normal with mean mu 0 and variance sigma 0 and similarly, f 1 is normal with mu n and variance sigma 1.

(Refer Slide Time: 28:31)



Some special cases

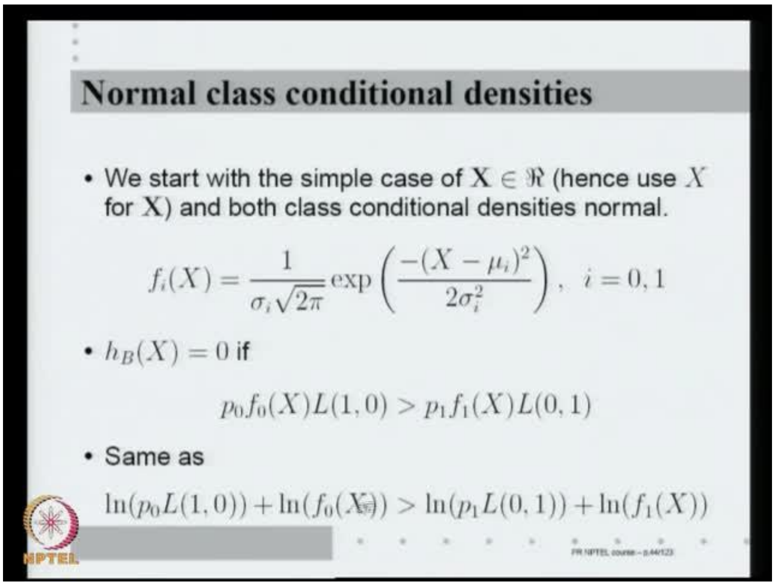
- Let us consider the two class case and explicitly write down Bayes classifier for some specific class conditional densities.
- For simplicity we write $L(\alpha_i, C_j) = L(i,j)$. We also assume $L(0,0) = L(1,1) = 0$.
- For the 2-class case, we decide on C_0 if

$$\frac{q_0(\mathbf{X})}{q_1(\mathbf{X})} = \frac{f_0(\mathbf{X})p_0}{f_1(\mathbf{X})p_1} \geq \frac{L(0,1)}{L(1,0)}$$

PR NPTEL course – p.41/123

.

(Refer Slide Time: 30:03)



Normal class conditional densities

- We start with the simple case of $\mathbf{X} \in \Re$ (hence use X for $\mathbf{X}$) and both class conditional densities normal.

$$f_i(X) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(\frac{-(X-\mu_i)^2}{2\sigma_i^2}\right), \quad i = 0, 1$$

- $h_B(X) = 0$ if

$$p_0 f_0(X) L(1,0) > p_1 f_1(X) L(0,1)$$

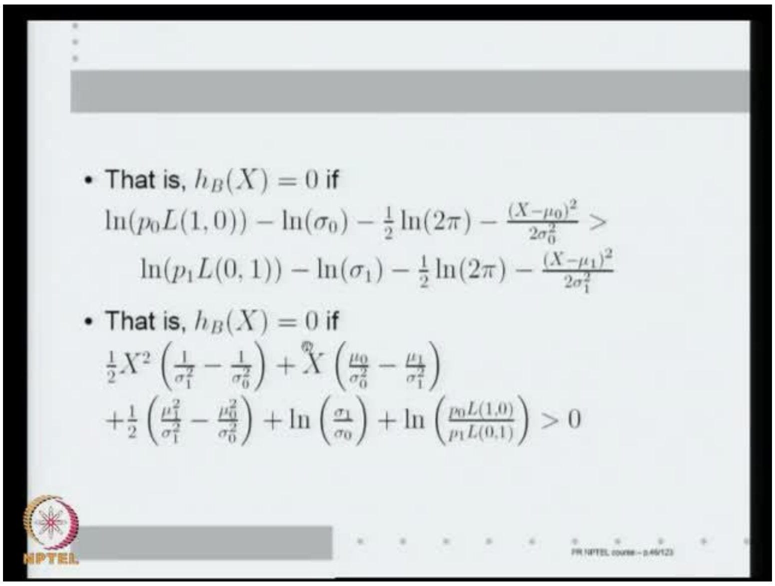
- Same as

$$\ln(p_0 L(1,0)) + \ln(f_0(X)) > \ln(p_1 L(0,1)) + \ln(f_1(X))$$

PR NPTEL course – p.44/123

Now, what is that we have to have h B X is 0 if p 0, f 0 as you can see from here what it means is p 0, f 0, L 1, 0 is greater than p 1, f 1, L 0, 1. So, p 0, f 0, L 1, 0 is greater than p 1, f 1, L 0, 1. Since, l n or log is a monotherm function this will be same as if I take log on both sides l n of p 0, L 0, 1 those are constants I will take out separately l n of f plus l n of f 0 is greater than l n of p 1, L 0, 1 plus l n of f 1. What did I gain by choosing l n because the class conditional density involves an exponential by taking l n I will get a simple linear quadratic expression in X.

(Refer Slide Time: 31:33)



- That is, $h_B(X) = 0$ if

$$\ln(p_0 L(1,0)) - \ln(\sigma_0) - \tfrac{1}{2}\ln(2\pi) - \tfrac{(X-\mu_0)^2}{2\sigma_0^2} >$$
$$\ln(p_1 L(0,1)) - \ln(\sigma_1) - \tfrac{1}{2}\ln(2\pi) - \tfrac{(X-\mu_1)^2}{2\sigma_1^2}$$

- That is, $h_B(X) = 0$ if

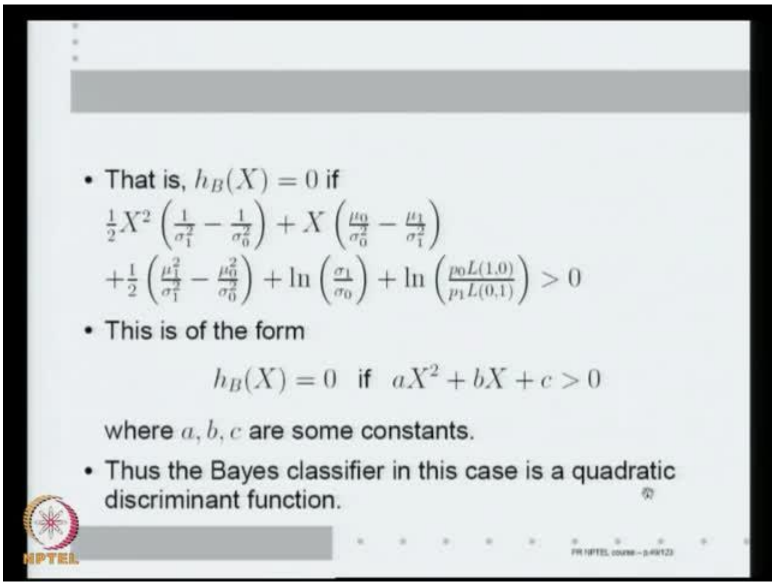
$$\tfrac{1}{2}X^2\left(\tfrac{1}{\sigma_1^2} - \tfrac{1}{\sigma_0^2}\right) + X\left(\tfrac{\mu_0}{\sigma_0^2} - \tfrac{\mu_1}{\sigma_1^2}\right)$$
$$+\tfrac{1}{2}\left(\tfrac{\mu_1^2}{\sigma_1^2} - \tfrac{\mu_0^2}{\sigma_0^2}\right) + \ln\left(\tfrac{\sigma_1}{\sigma_0}\right) + \ln\left(\tfrac{p_0 L(1,0)}{p_1 L(0,1)}\right) > 0$$

PR NPTEL course – p.46/123

So, let us explicitly write this and see what it looks. So, we want l n p 0, L 0, 1 plus l n f 0, what will l n f 0 given me l n of this l n of this the constant term is minus l n sigma sigma 0 and minus half l n 2 pi. And this will give me minus X minus mu 0 by 2 sigma i whole square, because l n cancels exponential. So, that is what I will get, this is the earlier constant term minus l n sigma 0 minus half l n 2 pi minus X minus mu 2 whole square by 2 0 sigma whole square. Similarly, in the right hand side where essentially where mu 0 came, now mu 0 comes where sigma 0 came or sigma 1 comes.

So, if X satisfies this expression, then I will put X in class 0, now this expression can be further crunched. So, I got an X square term here X square term from here the coefficient of X terms will be say let me write the next expression. So, I have half if I take out there is an X square term from here and X square term from here. Here I will get 1 by sigma 1 squared, here I get 1 by sigma 0 squared. So, I am bringing everything in the left side will be half X square into 1 by sigma 1 square minus 1 by sigma 0 square. The X term will be I get the two will cancel, I will get X mu 0 by sigma 0 square here and X mu 1 by sigma 1 square here. So, I will get X into mu 0 by sigma 0 square minus mu 1 by sigma square and then all the remaining constants. All these constants plus what I get from here, mu 0 square by sigma 0 square and mu 1 square by sigma 1 square.

(Refer Slide Time: 33:26)



- That is, $h_B(X) = 0$ if
  $\frac{1}{2}X^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) + X\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right)$
  $+\frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_0^2}{\sigma_0^2}\right) + \ln\left(\frac{\sigma_1}{\sigma_0}\right) + \ln\left(\frac{p_0 L(1,0)}{p_1 L(0,1)}\right) > 0$
- This is of the form

$$h_B(X) = 0 \quad \text{if} \quad aX^2 + bX + c > 0$$

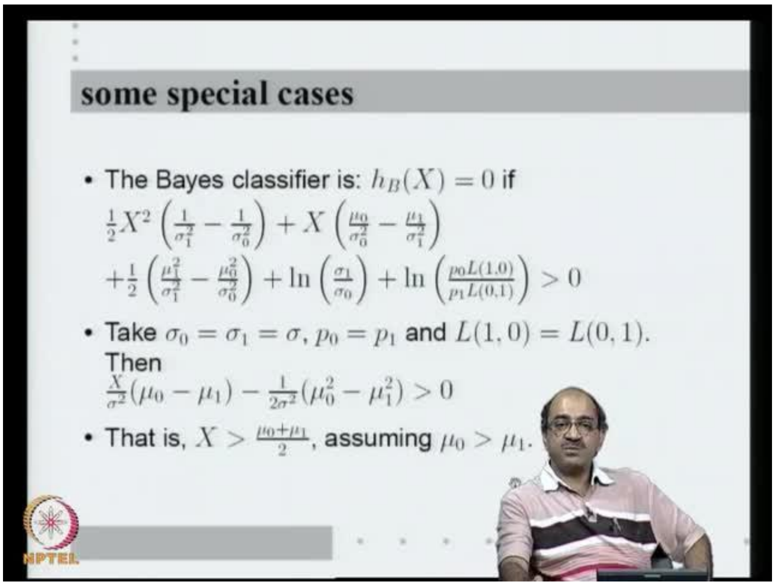
where a, b, c are some constants.

- Thus the Bayes classifier in this case is a quadratic discriminant function.

So, if X satisfies this expression then we say we will the Bayes classifier will put X in class 0. So, I have rewritten this expression here, so what we have done by algebra to show that for the one dimensional normal case, both class conditional densities are normal. Then Bayes classifier I will put X in class 0, if X satisfies expression. What is this expression? This expression is simply a quadratic expression, what it means is that I am saying h B X equals 0, if a square plus b X plus c is greater than 0, where a b and c are some constants.

Constant that depend on the underlying class conditional densities, they depend on mu 0 and sigma 0 and also on the loss function and prior probabilities. But essentially as a function of X it is nothing but a quadratic. So, my best classifier now turns out to be h B X equals to 0, if a X square plus b X plus c greater than 0. What does that mean, in this case the Bayes classifier is a quadratic discriminant function is essentially depends on the quadratic discriminant function. So, for one dimensional both class conditional normal the optimum Bayes classifier is nothing but a quadratic discriminant function.

(Refer Slide Time: 34:28)



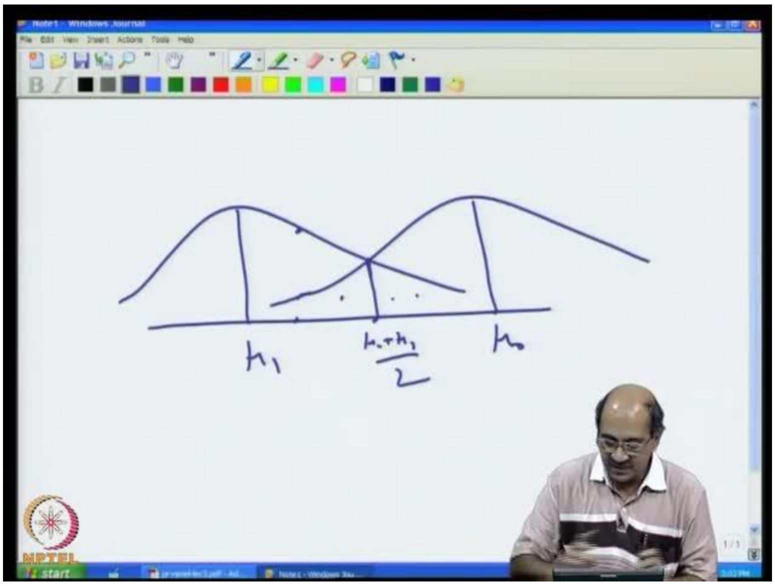
some special cases

- The Bayes classifier is: $h_B(X) = 0$ if $\frac{1}{2}X^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) + X\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right)$ $+\frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_0^2}{\sigma_0^2}\right) + \ln\left(\frac{\sigma_1}{\sigma_0}\right) + \ln\left(\frac{p_0 L(1,0)}{p_1 L(0,1)}\right) > 0$
- Take $\sigma_0 = \sigma_1 = \sigma$, $p_0 = p_1$ and $L(1,0) = L(0,1)$. Then $\frac{X}{\sigma^2}(\mu_0 - \mu_1) - \frac{1}{2\sigma^2}(\mu_0^2 - \mu_1^2) > 0$
- That is, $X > \frac{\mu_0+\mu_1}{2}$, assuming $\mu_0 > \mu_1$.

Actually in special cases it may become linear. So, this is the general expression Bayes classifier has since, X to 0 class, if this this expression this quadratic expression greater than 0. Now, let us take a special case where the two conditional densities are such that there variances are same that is sigma 0 equal to sigma 1 unless assume that the priors are also same and let us say we are on 0-1 cross function. So, that L 1, 0 is equal to L 0, 1. Now, when sigma 0 is equal to sigma 1, the X square term drops off, once the X square term drops off what I have is a linear function X that means no Bayes classifier becomes a linear discriminant function. X by sigma square into mu 0 minus mu 1 all the others have 1 expect this term, there is 1 by 2 sigma square into mu 0 square minus mu 1 square.

So, in this special case if the two conditional densities are normal with same variance and priors are same under 0-1 loss function. Then essentially X has to satisfy this linear expression, now if I assume mu 0 is greater than mu 1 I can cancel the mu 0 minus mu 1 common factor without changing the sign of the inequality. So, that is if X is greater than mu 0 plus mu 1 by 2 then you put in class 0 or otherwise put in class 1, here we we assume mu 0 greater than mu 1.

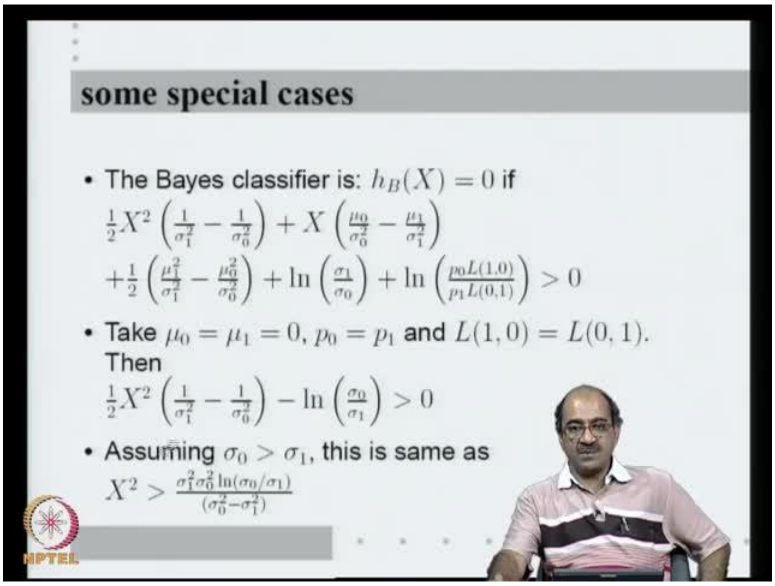
(Refer Slide Time: 35:59)



Let us, we can actually see that this is this is quite a intuitively clear picture let us draw the two normal densities, we assumed mu 0 greater than mu 1. So, that is mu 0 that is mu 1 both have the same variance. So, where would they cut they will cut midway between the two means. Now, what is the Bayes classifier because the pairs are same angle and the 0-1 loss function, all it means is whichever conditional density is higher at any given h. So, if I am write this X at this X the f 0 X at this much value, f 1 X at this much value. So, because f 1 is more than f 0 I will put this plus 1.

So, when when both class conditionals are normal and variance are same essentially Bayes classifier means, till this threshold it will be put in the left class, after this threshold it will be put in the right class. So, that is why if the classifier base upper classifier turned out to be if X is less than mu 0. If X is less than mu 0 plus mu 1 by 2 then put it in the mu 1 class otherwise, put it in the mu 0 class. As you can see this from this picture this is the same thing that we showed about the error integral being same. The error integral is the area of these two these two type these these two sides. So, as we would expect because variances are same essentially the curves will cut midway between mu 0 and mu 1. So, till mu 0 plus mu 1 by 2 it should be one class otherwise, it should be other class.

(Refer Slide Time: 38:18)

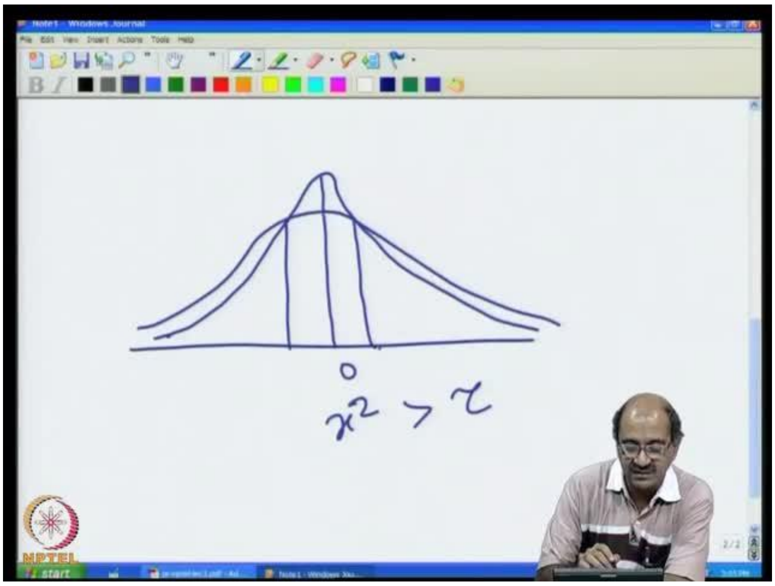


some special cases

- The Bayes classifier is: $h_B(X) = 0$ if $\frac{1}{2}X^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) + X\left(\frac{\mu_0}{\sigma_0^2} - \frac{\mu_1}{\sigma_1^2}\right)$ $+\frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_0^2}{\sigma_0^2}\right) + \ln\left(\frac{\sigma_1}{\sigma_0}\right) + \ln\left(\frac{p_0 L(1,0)}{p_1 L(0,1)}\right) > 0$
- Take $\mu_0 = \mu_1 = 0$, $p_0 = p_1$ and $L(1,0) = L(0,1)$. Then $\frac{1}{2}X^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) - \ln\left(\frac{\sigma_0}{\sigma_1}\right) > 0$
- Assuming $\sigma_0 > \sigma_1$, this is same as $X^2 > \frac{\sigma_1^2\sigma_0^2 \ln(\sigma_0/\sigma_1)}{(\sigma_0^2 - \sigma_1^2)}$

That is what we got X greater than mu 0 plus mu 1 by 2 put it in the C 0 otherwise put it in C 1. So, we can say that this is a very intuitive classifier, so the Bayes classifier gives you what will intuitively think is the right one. Let us take one more special case, once again this is my general classifier. Now, instead of assuming that these variances are same, now let us assume means are same and variances are different that is assumed mu 0 is equal to mu 1 equal to 0. Once again curves equal pairs and 0-1 loss function. Now, what happens is the actual linear term drops out, only the quaternary term is left. This half X square into 1 by sigma square minus sigma 0 square plus minus ln sigma 0 by sigma 1. So, I can crunch it so it essentially means that X is greater than something, then X square is greater than something then you put in class 0. I am assuming sigma 0 as greater than sigma 1.

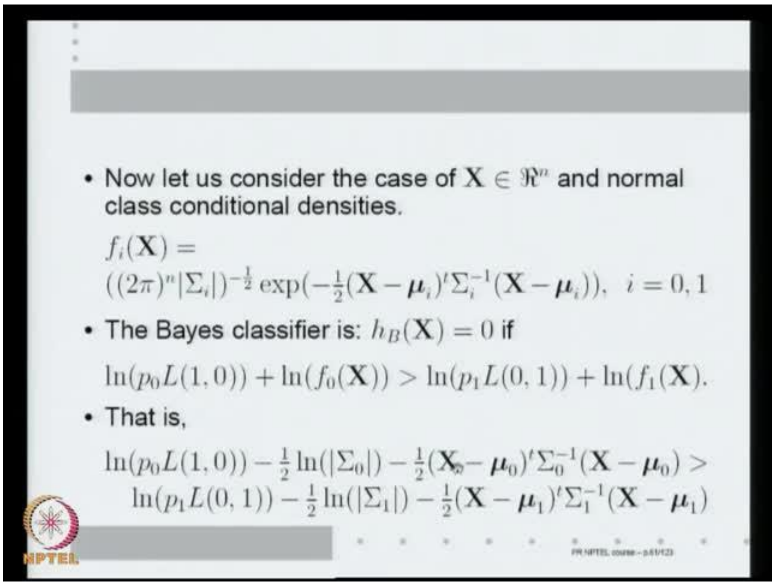
(Refer Slide Time: 39:16)



Once again we can see that this is intuitively clear, whatever we are assuming now both the means are 0, 1 classifier has large mu, another classifier has large larger variance, another classifier has another class conditional density has smaller variance. Because, both the variances are same we assumed that to be 0 the one that has larger variance, will go up at 0. So it will cut the other one on either side symmetrically, so from this point to this point when X is between these two this class conditional density is larger, at all the other places other class conditional density is larger.

So, we will expect that for we will our final classifier should be something like if X square is greater than some tau. Then I should put in the smaller variance class, in the larger variance class, if X square is less than tau I should put in the other class. So, h between some minus something to plus something it will be in this smaller variance class and outside of that it will be in the larger variance class. That is exactly what we got.

(Refer Slide Time: 40:52)



- Now let us consider the case of $\mathbf{X} \in \Re^n$ and normal class conditional densities.

$$f_i(\mathbf{X}) = ((2\pi)^n|\Sigma_i|)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)^t\Sigma_i^{-1}(\mathbf{X}-\boldsymbol{\mu}_i)), \quad i=0,1$$

- The Bayes classifier is: $h_B(\mathbf{X}) = 0$ if

$$\ln(p_0L(1,0)) + \ln(f_0(\mathbf{X})) > \ln(p_1L(0,1)) + \ln(f_1(\mathbf{X})).$$

- That is,

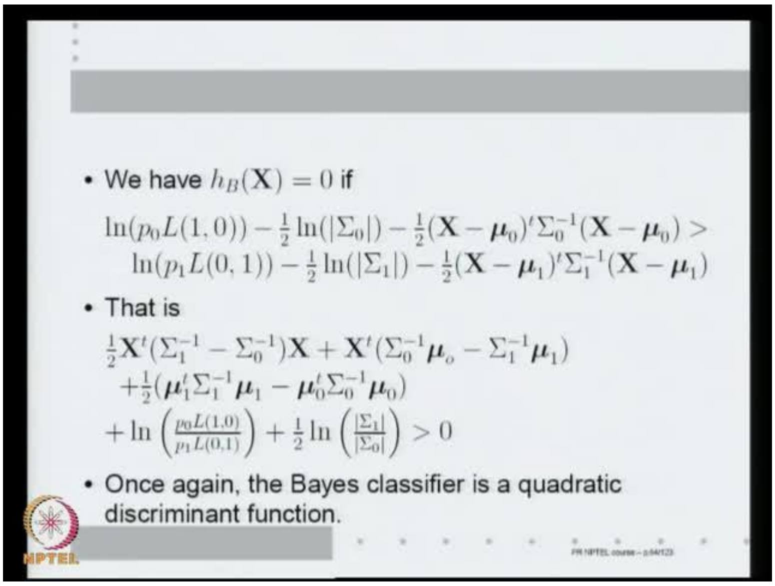
$$\ln(p_0L(1,0)) - \tfrac{1}{2}\ln(|\Sigma_0|) - \tfrac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_0)^t\Sigma_0^{-1}(\mathbf{X}-\boldsymbol{\mu}_0) >$$
$$\ln(p_1L(0,1)) - \tfrac{1}{2}\ln(|\Sigma_1|) - \tfrac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_1)^t\Sigma_1^{-1}(\mathbf{X}-\boldsymbol{\mu}_1)$$

PR NPTEL course – p.61/123

We are assuming sigma 0 is greater than sigma 1. So, if X square is greater than something then we put in the larger variance class, so once again an intuitively very clear case. Now, let us consider let us become more ambitious and consider the n dimensional case, once again normal class conditional densities. Now, that we have done for one dimensions we can do it a little quicker, we m n dimensional class normal density is given by this. The joint density will be 2 pi to the power n by 2 pi to the power n by 2 sorry 2 pi to the power n into a the determinant of sigma a whole to the power minus half and exponential the quadratic X minus mu i transpose sigma inverse X minus mu i.

So, mu 0 and mu 1 as earlier or the means and sigma 1 0 1 sigma 1 at the covariance matrices. And now of course, X is a vector so on mu 0, 1, mu 1 and sigma i will be a n by n matrix. So, sigma both sigma 0 and sigma 1 will be n by n matrices, they are the covariance matrices. Once again this is my Bayes classifier it puts X in 0, if this expression is satisfied once again taking (( )) help me because the exponential will go away. So, what will be l n of f 0, there will be l n of constant plus I get a quadratic form minus half X minus mu 0 transpose sigma 0, inverse X minus mu 0 and similarly, in the L and f 1 side. So, that is what I will get, I got the quadratic form with mu 0 once and mu 1, once it is like the old case this is a quadratic form. So, there will be an X square term there will be there is a X transpose some matrix into X term.

(Refer Slide Time: 42:31)

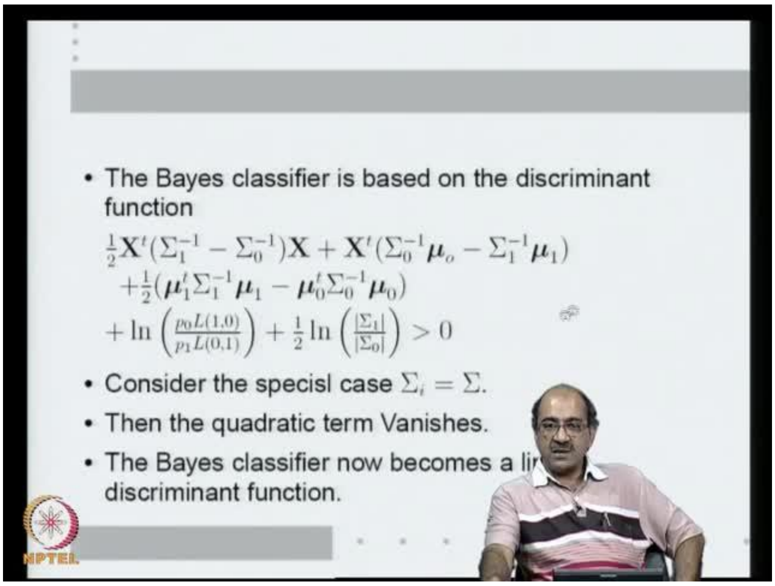


- We have $h_B(\mathbf{X}) = 0$ if
$$\ln(p_0 L(1,0)) - \frac{1}{2}\ln(|\Sigma_0|) - \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_0)^t \Sigma_0^{-1}(\mathbf{X} - \boldsymbol{\mu}_0) > \ln(p_1 L(0,1)) - \frac{1}{2}\ln(|\Sigma_1|) - \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_1)^t \Sigma_1^{-1}(\mathbf{X} - \boldsymbol{\mu}_1)$$
- That is
$$\frac{1}{2}\mathbf{X}^t(\Sigma_1^{-1} - \Sigma_0^{-1})\mathbf{X} + \mathbf{X}^t(\Sigma_0^{-1}\boldsymbol{\mu}_o - \Sigma_1^{-1}\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{\mu}_1^t\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^t\Sigma_0^{-1}\boldsymbol{\mu}_0) + \ln\left(\frac{p_0 L(1,0)}{p_1 L(0,1)}\right) + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) > 0$$
- Once again, the Bayes classifier is a quadratic discriminant function.

PR NPTEL course – p.54/123

So, there will be a linear term in X and a constant term if I bring all of them to one side this is what I will get. There will be X transpose sigma n inverse minus sigma 0 inverse X, X transpose sigma 0 inverse mu 0 minus sigma 0 inverse mu 1 plus the constant. And you can remember in the in the scalar case you got half X square into 1 by sigma 1 square minus 1 by sigma 0 square. Now, it becomes a vector case that 1 by has become inverse of the covariance matrix and this becomes a quadratic form otherwise it is exactly same thing.

(Refer Slide Time: 43:20)

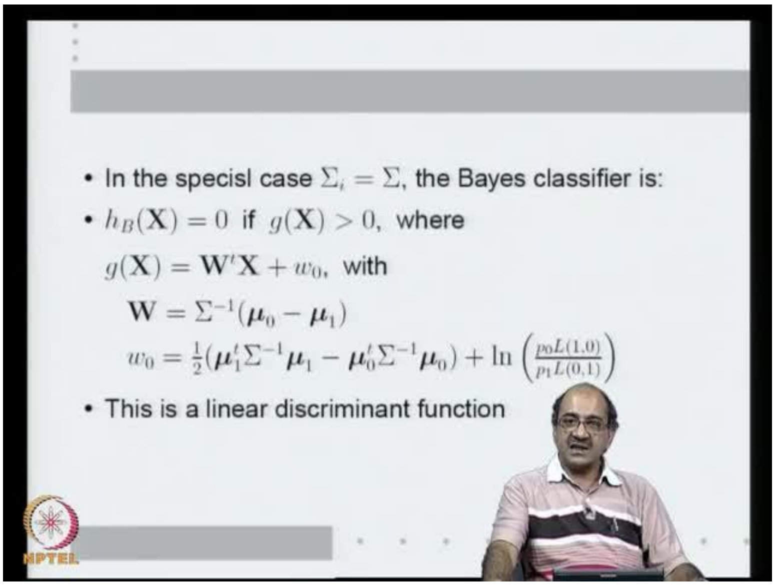


- The Bayes classifier is based on the discriminant function
$$\frac{1}{2}\mathbf{X}^t(\Sigma_1^{-1} - \Sigma_0^{-1})\mathbf{X} + \mathbf{X}^t(\Sigma_0^{-1}\boldsymbol{\mu}_o - \Sigma_1^{-1}\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{\mu}_1^t\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^t\Sigma_0^{-1}\boldsymbol{\mu}_0) + \ln\left(\frac{p_0 L(1,0)}{p_1 L(0,1)}\right) + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) > 0$$
- Consider the specisl case $\Sigma_i = \Sigma$.
- Then the quadratic term Vanishes.
- The Bayes classifier now becomes a li discriminant function.

So, once again Bayes classifier is a quadratic discriminant function. So, whether in one dimensions or n dimensions, if the class conditional density is such normal then whether or not the feature vector is for any dimension the feature vector optimum Bayes classifier is a quadratic discriminant function. So, Bayes classifier is a quadratic discriminant function once again we can do the same special case if I take sigma is equal to sigma then the quadratic term drops of, when the quadratic term drops of have some X transpose into some constant plus some constant greater than 0 then class 1. So, you will be like a linear discriminant function.

On the quadratic term vanishes the Bayes classifier now becomes a linear discriminant function. So, in the special case of sigma is equal to sigma the Bayes classifier becomes h B X equals to 0 g X is greater than 0, where g X is w transpose X plus w 0. What is so once the quadratic term drops out this is w, w transpose X plus whatever is left is w 0.

(Refer Slide Time: 44:06)



- In the specisl case $\Sigma_i = \Sigma$, the Bayes classifier is:
- $h_B(\mathbf{X}) = 0$ if $g(\mathbf{X}) > 0$, where

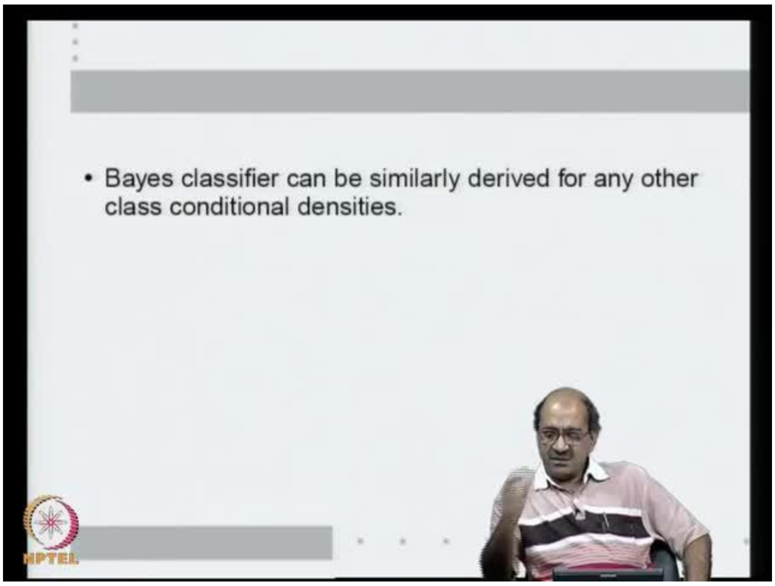
  $g(\mathbf{X}) = \mathbf{W}^t\mathbf{X} + w_0$, with

  $\mathbf{W} = \Sigma^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$

  $w_0 = \frac{1}{2}(\boldsymbol{\mu}_1^t\Sigma^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^t\Sigma^{-1}\boldsymbol{\mu}_0) + \ln\left(\frac{p_0L(1,0)}{p_1L(0,1)}\right)$
- This is a linear discriminant function

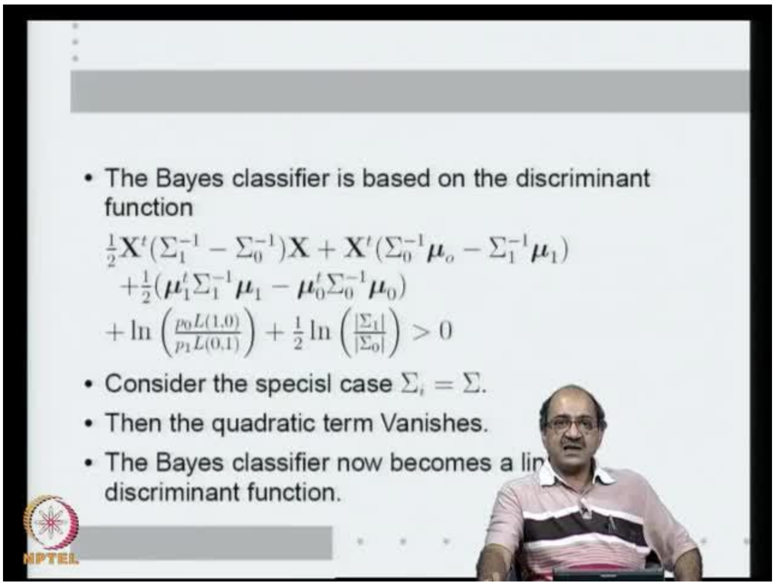
So, it becomes W transpose X where W is sigma inverse mu 0 minus mu 1 and w 0 is all the remaining terms, this is the linear discriminant function. So, both in one dimensional n dimensional case, if the covariance of the two classes are same Bayes (( )) classifies the linear discriminant function. Otherwise, in the general case it is a quadratic discriminant function.

(Refer Slide Time: 44:26)



- Bayes classifier can be similarly derived for any other class conditional densities.

(Refer Slide Time: 44:41)

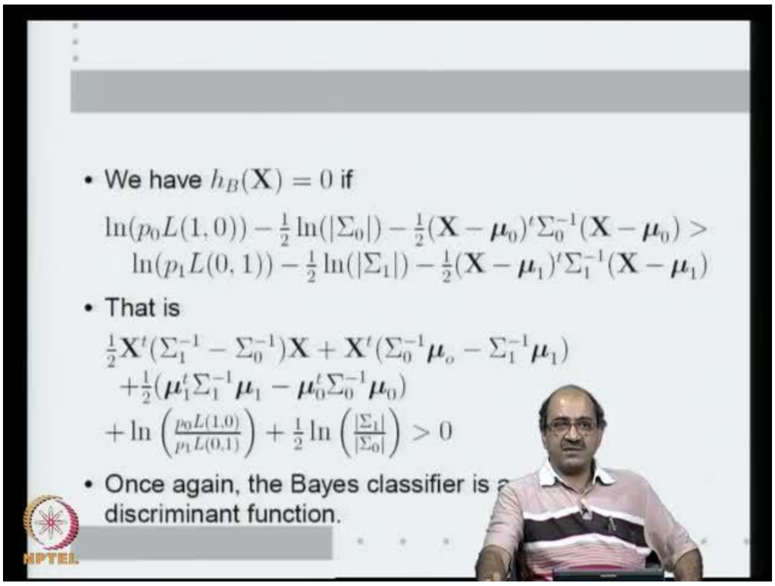


- The Bayes classifier is based on the discriminant function

$$\tfrac{1}{2}\mathbf{X}^t(\Sigma_1^{-1} - \Sigma_0^{-1})\mathbf{X} + \mathbf{X}^t(\Sigma_0^{-1}\boldsymbol{\mu}_o - \Sigma_1^{-1}\boldsymbol{\mu}_1)$$
$$+\tfrac{1}{2}(\boldsymbol{\mu}_1^t\Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^t\Sigma_0^{-1}\boldsymbol{\mu}_0)$$
$$+\ln\left(\frac{p_0L(1,0)}{p_1L(0,1)}\right) + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) > 0$$

- Consider the specisl case $\Sigma_i = \Sigma$.
- Then the quadratic term Vanishes.
- The Bayes classifier now becomes a lin discriminant function.

(Refer Slide Time: 44:44)



- We have $h_B(\mathbf{X}) = 0$ if

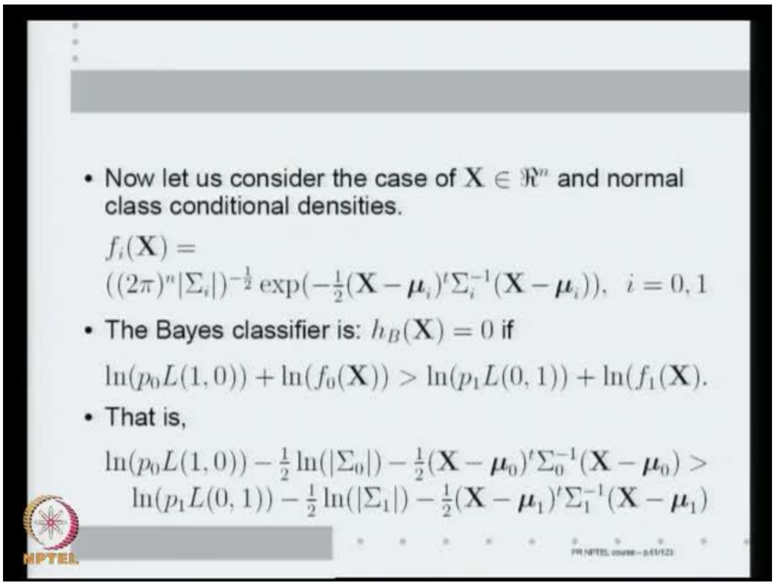
$$\ln(p_0 L(1,0)) - \tfrac{1}{2}\ln(|\Sigma_0|) - \tfrac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_0)^t \Sigma_0^{-1}(\mathbf{X} - \boldsymbol{\mu}_0) > \ln(p_1 L(0,1)) - \tfrac{1}{2}\ln(|\Sigma_1|) - \tfrac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_1)^t \Sigma_1^{-1}(\mathbf{X} - \boldsymbol{\mu}_1)$$

- That is

$$\tfrac{1}{2}\mathbf{X}^t(\Sigma_1^{-1} - \Sigma_0^{-1})\mathbf{X} + \mathbf{X}^t(\Sigma_0^{-1}\boldsymbol{\mu}_o - \Sigma_1^{-1}\boldsymbol{\mu}_1) + \tfrac{1}{2}(\boldsymbol{\mu}_1^t \Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^t \Sigma_0^{-1}\boldsymbol{\mu}_0) + \ln\left(\frac{p_0 L(1,0)}{p_1 L(0,1)}\right) + \tfrac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) > 0$$

- Once again, the Bayes classifier is a discriminant function.

(Refer Slide Time: 44:48)



- Now let us consider the case of $\mathbf{X} \in \Re^n$ and normal class conditional densities.

$$f_i(\mathbf{X}) = ((2\pi)^n|\Sigma_i|)^{-\frac{1}{2}} \exp(-\tfrac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_i)^t \Sigma_i^{-1}(\mathbf{X} - \boldsymbol{\mu}_i)), \quad i = 0, 1$$

- The Bayes classifier is: $h_B(\mathbf{X}) = 0$ if

$$\ln(p_0 L(1,0)) + \ln(f_0(\mathbf{X})) > \ln(p_1 L(0,1)) + \ln(f_1(\mathbf{X}).$$

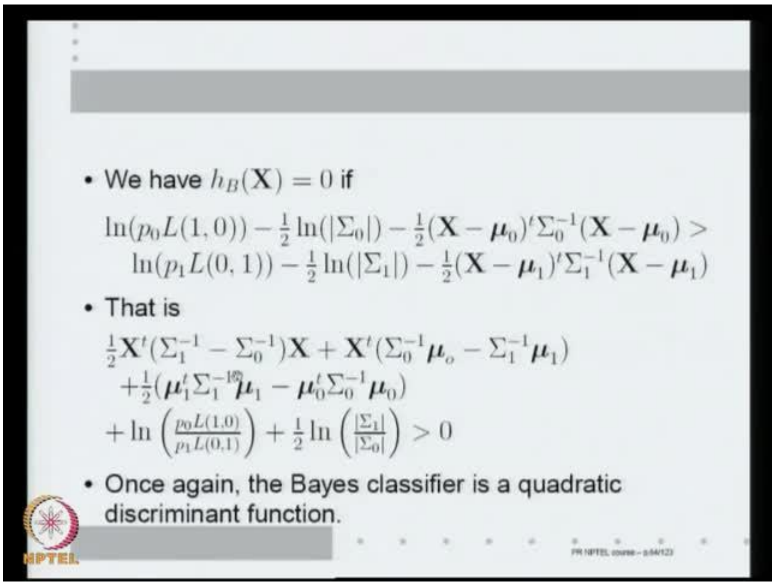
- That is,

$$\ln(p_0 L(1,0)) - \tfrac{1}{2}\ln(|\Sigma_0|) - \tfrac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_0)^t \Sigma_0^{-1}(\mathbf{X} - \boldsymbol{\mu}_0) > \ln(p_1 L(0,1)) - \tfrac{1}{2}\ln(|\Sigma_1|) - \tfrac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_1)^t \Sigma_1^{-1}(\mathbf{X} - \boldsymbol{\mu}_1)$$

PR NPTEL course – p.61/123

So, in this way so we have we have seen examples of both one dimensions and n dimensions only normal normal density you say considered. But these exercise see basically what is that we have done, if you look at all our derivations we start with a assumed class conditional density. Then we say we use our basic derived Bayes classifier h B X equal to 0, if this is greater than this. And you plug in the expression for f 0and f 1 and then with algebra you can ask, what kind of expression is this in X?

(Refer Slide Time: 45:26)



- We have $h_B(\mathbf{X}) = 0$ if

$$\ln(p_0 L(1,0)) - \frac{1}{2}\ln(|\Sigma_0|) - \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_0)^t \Sigma_0^{-1}(\mathbf{X} - \boldsymbol{\mu}_0) > \ln(p_1 L(0,1)) - \frac{1}{2}\ln(|\Sigma_1|) - \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu}_1)^t \Sigma_1^{-1}(\mathbf{X} - \boldsymbol{\mu}_1)$$

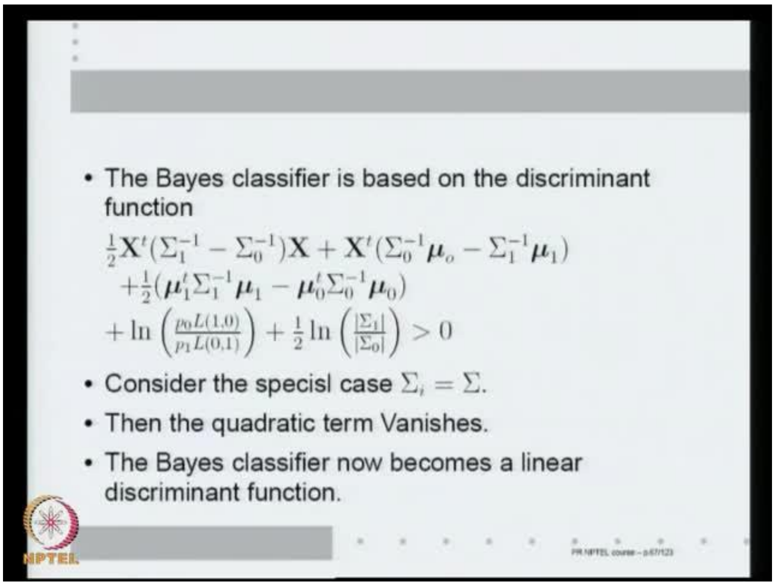
- That is

$$\frac{1}{2}\mathbf{X}^t(\Sigma_1^{-1} - \Sigma_0^{-1})\mathbf{X} + \mathbf{X}^t(\Sigma_0^{-1}\boldsymbol{\mu}_o - \Sigma_1^{-1}\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{\mu}_1^t \Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^t \Sigma_0^{-1}\boldsymbol{\mu}_0) + \ln\left(\frac{p_0 L(1,0)}{p_1 L(0,1)}\right) + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) > 0$$

- Once again, the Bayes classifier is a quadratic discriminant function.

PR NPTEL course – p.64/123

(Refer Slide Time: 45:29)



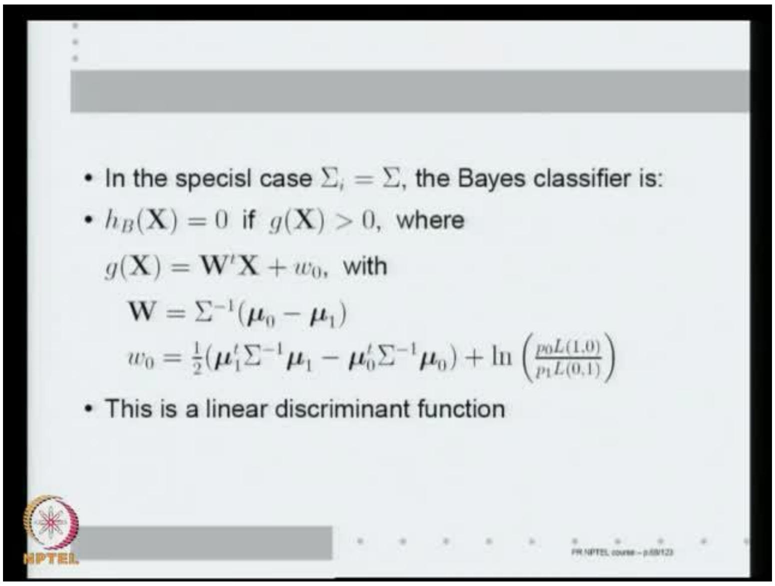
- The Bayes classifier is based on the discriminant function

$$\frac{1}{2}\mathbf{X}^t(\Sigma_1^{-1} - \Sigma_0^{-1})\mathbf{X} + \mathbf{X}^t(\Sigma_0^{-1}\boldsymbol{\mu}_o - \Sigma_1^{-1}\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{\mu}_1^t \Sigma_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^t \Sigma_0^{-1}\boldsymbol{\mu}_0) + \ln\left(\frac{p_0 L(1,0)}{p_1 L(0,1)}\right) + \frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) > 0$$

- Consider the specisl case $\Sigma_i = \Sigma$.
- Then the quadratic term Vanishes.
- The Bayes classifier now becomes a linear discriminant function.

PR NPTEL course – p.67/123

(Refer Slide Time: 45:32)



- In the specisl case $\Sigma_i = \Sigma$, the Bayes classifier is:
- $h_B(\mathbf{X}) = 0$ if $g(\mathbf{X}) > 0$, where

  $g(\mathbf{X}) = \mathbf{W}^t\mathbf{X} + w_0$, with

  $$\mathbf{W} = \Sigma^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$$

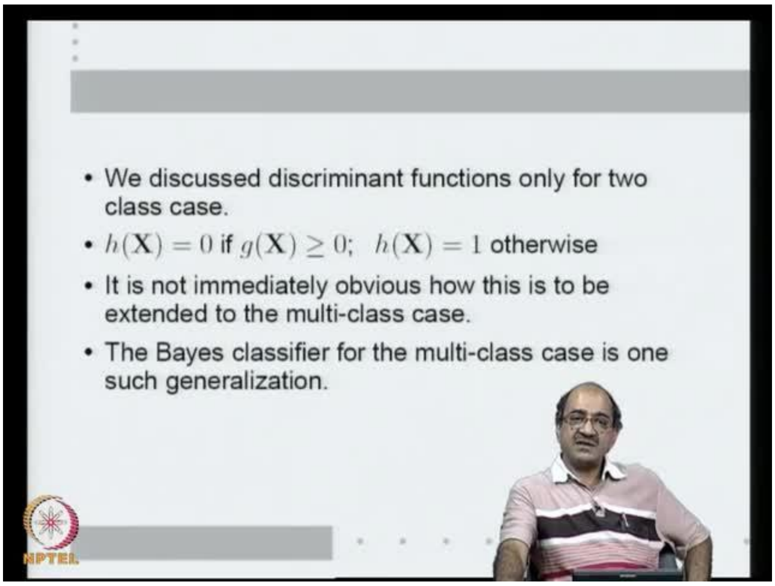
  $$w_0 = \frac{1}{2}(\boldsymbol{\mu}_1^t\Sigma^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^t\Sigma^{-1}\boldsymbol{\mu}_0) + \ln\left(\frac{p_0 L(1,0)}{p_1 L(0,1)}\right)$$
- This is a linear discriminant function

PR NPTEL course – p.69/123

It could be a quadratic expression, it could be a cubic expression, it could be alinear expression, that is how you showed that Bayes classifier in different cases, is either a quadratic discriminant function or a linear discriminant function and so on. So, Bayes classifier in the same way can be derived for many other class conditional densities. Later on may be you will see a few more examples, but I hope the the basic method is clear now. Depending on the nature of the densities the final expression can be complicated of course, here their exponential something. So, when I take l n everything turned out to be just a quadratic form.

In some other density may be some of the expressions can become a little more complicated, but the point is given all the class conditional densities and prior probabilities and knowing the loss function. One can easily decide on the class of any given feature vector, we can actually analytically crunch and get a very nice simple expression. Of course, we could have actually calculated f 0 X, f 1 X and plugged it with that equation, but that involves finding exponentials. By doing all this algebra we finally, just evaluating either quadratic form in X or simply a linear function of X, which is more simpler than evaluating the quadratic form, then putting in one exponential evaluating, another quadratic form putting in another exponential, that is computationally more expensive.

(Refer Slide Time: 47:27)



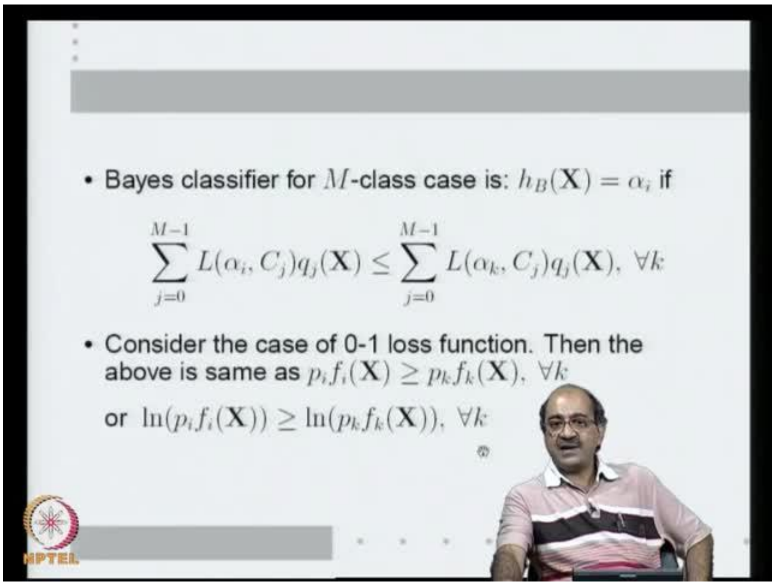
- We discussed discriminant functions only for two class case.
- $h(\mathbf{X}) = 0$ if $g(\mathbf{X}) \geq 0$; $h(\mathbf{X}) = 1$ otherwise
- It is not immediately obvious how this is to be extended to the multi-class case.
- The Bayes classifier for the multi-class case is one such generalization.

So, what all this algebra is given is that we can actually find the final form of how to decide on the class. Of course, we can always calculate p 0, f 0, p 1, f 1, no matter what f 0, f 1 was and then we we we can get these Bayes classifier. What this example showed is in specific cases, cracking this can be simpler than actually evaluating f 0 and f 1 at an X, given full statistical information Bayes classifier optimal and we can always implemented. Because, the next question is who will give the full statistical information, as I said earlier we will see later on in our coming lectures, how we can estimate the class conditional densities and hence, posterior probabilities given the training data that we will see later on.

Now, this another thing that we can see from what we derived today, so far in in the beginning of this class I have described discriminant functions. But we discussed it only for two class case, with saying discriminant function based classifier is h B X is 0, if some g X is greater than 0, where g is the discriminant function.

Now, if I multiple classes how will the discriminant function idea is to be extended to multiple classes. When there are only two classes sign depending on the sign of one function, I can say which classifier it is then the discriminant function based classifier is very clear. But if multiple classes I do not know how how to extend it I extend the concept of discriminant function to multiple classes. The Bayes optimal classifier form that we got we will immediately give us one idea of how we can extend discriminant function to multiple classes. So, here it goes, the two classes discriminant function as we seen is this form h X equals 0 of g X greater than or equal to 0 otherwise, h X equals to 1.

(Refer Slide Time: 48:48)



- Bayes classifier for M-class case is: $h_B(\mathbf{X}) = \alpha_i$ if

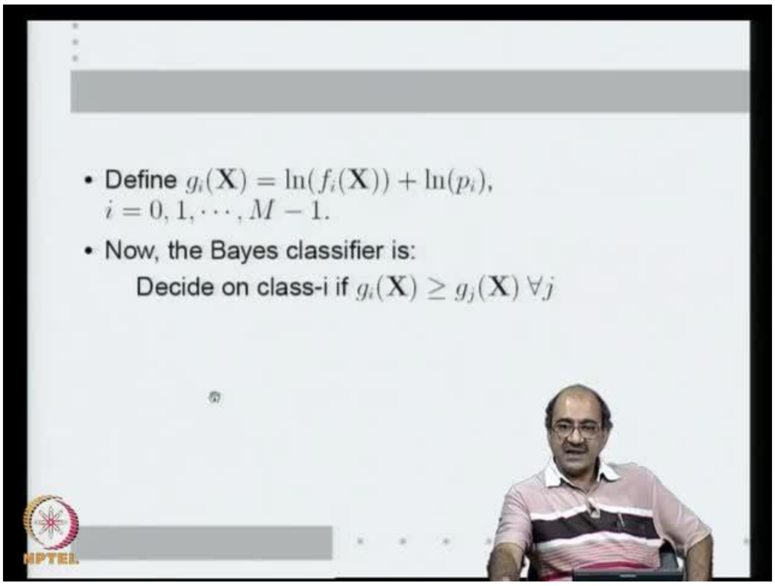
$$\sum_{j=0}^{M-1} L(\alpha_i, C_j) q_j(\mathbf{X}) \le \sum_{j=0}^{M-1} L(\alpha_k, C_j) q_j(\mathbf{X}), \ \forall k$$

- Consider the case of 0-1 loss function. Then the above is same as $p_i f_i(\mathbf{X}) \ge p_k f_k(\mathbf{X}), \ \forall k$

  or $\ln(p_i f_i(\mathbf{X})) \ge \ln(p_k f_k(\mathbf{X})), \ \forall k$

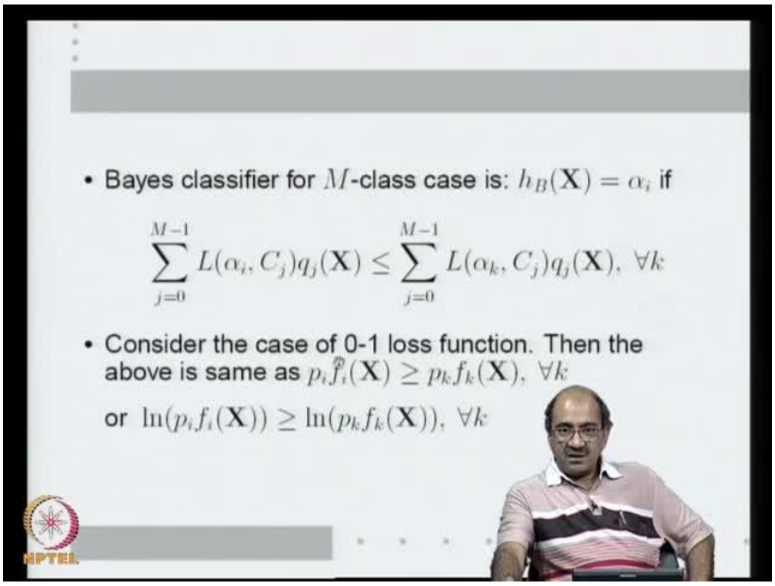
So, as I said the Bayes classifier for a multi class case is is a generalisation of the discriminant idea to multiple case, we have not seen it like that. So, let us see it like that, this is the Bayes classifier for multiple case, this is the most general Bayes classifier. But if you can say say 0-1 loss function then this thing is same as p i, f i greater than p k, f k, we seen that essentially we would say p i greater than p k for all k, then I have put in class i. So, essentially in the 0-1 case, we generally classify turns out to be p i, f i greater than p k, f k for all k, which is same as l n of p i, f i X greater than l n of p k, f k X for all k. So, this is the general Bayes classifier for 0-1 loss function. I can view it as a discriminant function based classifier as follows; define g of X as l n of f i X plus l n, p i. What is that I have to do actually p a, l n p i, f i greater than l n p k, f k.

(Refer Slide Time: 49:50)



- Define $g_i(\mathbf{X}) = \ln(f_i(\mathbf{X})) + \ln(p_i)$, $i = 0, 1, \cdots, M-1$.
- Now, the Bayes classifier is: Decide on class-i if $g_i(\mathbf{X}) \geq g_j(\mathbf{X}) \; \forall j$

(Refer Slide Time: 50:16)



- Bayes classifier for M-class case is: $h_B(\mathbf{X}) = \alpha_i$ if
$$\sum_{j=0}^{M-1} L(\alpha_i, C_j) q_j(\mathbf{X}) \leq \sum_{j=0}^{M-1} L(\alpha_k, C_j) q_j(\mathbf{X}), \; \forall k$$
- Consider the case of 0-1 loss function. Then the above is same as $p_i f_i(\mathbf{X}) \geq p_k f_k(\mathbf{X}), \; \forall k$ or $\ln(p_i f_i(\mathbf{X})) \geq \ln(p_k f_k(\mathbf{X})), \; \forall k$

So, this is l n p i plus l n f i, where the l n p k plus l n f k. So, I am saying define that as a g i of X, so g i of X is l n f i X plus l n p i. So, I have now M such functions instead of one function I have M g i functions. Now, what did the Bayes classifier saying now you decide and class i if g i X is greater than g j X for all j, two in the in the two class case I am saying the discriminant based classifier is h X equals to 0, if g X greater than 0 for where g is a single function that is called the discriminant function.

Now, the Bayes classifier for 0-1 loss function case transferred to be there are some M functions g i X, each g i X is given in terms of the class conditional densities, but that is of no concerned was right now. Essentially there are M functions g i and the base classifier is you decide on class i, if the i th value functions at X is greater than value any other functions value at X, if g i of X is greater than g j of X for all j, then you decide on class i.

So, this is a general way in which I can use discriminant functions for the M-class case. So, while two class cases only one discriminant function, for the M-class case I am actually have M functions. And I will make my decision by saying which function have the highest value of course, as in the Bayes classifier I have to break ties arbitrarily. But except for breaking tiles arbitrarily, this is the generic form for discriminant function based on M-class case. We will come to learn in different functions directly later on, at that time we will see how to learn all these functions. But this is a generalisation discriminant function based classifier to the M-class case.

So, let me sum up what we have done this class. We have derived the base optimal classifier for M-classes and any general loss function and we have proved we have stated what the Bayes optimal classifier is and then stated. And then proved that it actually minimises the risk among all classifier that have full statistical information and this is the most general case. So, even though we proved earlier for the 0-1 loss function and two class case, this is for a M-class general loss function. And then we seen many other special cases that comes out of it, general loss function two class case, 0-1 loss function for M-class case. For example, 0-1 loss function M-class case gives you the same generic generic view of Bayes classifier at the two class case.

The two class case I calculate the two posterior probabilities and whichever class have higher probability I will put in there. Even in M-class if I have 0-1 loss function, all you have to do is to calculate all the posteriors and whichever posterior probability is higher I will put it in that class. Then we also see how in a specific case one writes on the Bayes classifier, instead of instead of calculating p 0, f 0 on p 1, f 1, p 0, f 0 X in p 1, f 1, X 1 give X, which involves calculation f 0 X and f of X. We have seen in specific cases given the functional form of the class conditional densities, the expression p 0, f 0 is greater than p 1, f 1 X can be simplified into a much simpler function of X to compute.

For example; when the two class conditional densities are normal then this is simply a quadratic expression in X, instead of being exponential functions. And in this special case where the covariance matrices are same is even better it is just a linear function. That means Bayes optimal classifier, when the class conditional densities are normal turns out to be a quadratic discriminant function in the general case and turns out to bea linear discriminant function in the case of equal covariances. Then we also seen that the the general Bayes classifier where the 0-1 loss function tells us a a simple way in which to generalise the idea of discriminant functions. What we will do in the next class I will just given you a... We will also may be do one more example of Bayes classifier to understand how this computation is done, but more importantly now that we know the base classifier.

The next question is to ask can I calculate the error at the base classifier. As it turns out given the class conditional densities obtain the Bayes classifier is very simple as we have seen in the simple in the sense, that can always write the expressions and crunch them. But actually calculating the Bayes the error that the Bayes classifier makes, which is the optimal error for the classification problem. And hence, it is a good number to know what is the attainable probability of correctness, very often calculating Bayes error turns out to be difficult. So, next class we look at some techniques for approximately calculating in (( )). After that what we will see that even when we we have full statistical information, even when we know all the class conditional densities, there is no reason why we should only look for disc minimisation, there can be other objective functions.

So, we look at least a couple of other equally plausible ways of defining what an optimal classifier, look at what those classifiers mean. And hence, see that Bayes classifier will be says optimal is only one of the many different criteria that you can use. However, the reason why we emphasise based classifier so much is that minimising risk is often a very useful practice in machinery and pattern in general. And Bayes classifier is based and minimising the risk and that is the reason based classifier has a very important place. So, we will just look see that the other optimisation criteria, but and then come get back to Bayes classifier and disc minimisation.

Thank you.