

Deep Learning
Prof. Prabir Kumar Biswas
Department Of Electronics And Electrical Communication Engineering
Indian Institute of Technology, Kharagpur

Lecture – 05
Bayesian Learning – II

Hello, welcome to the NPTEL Online Certification Course on Deep Learning. You remember that in the previous lecture we have talked about the feature distribution in the feature vector space and we have shown that this distribution because the different objects that we get from the same class all of them may not be identical. The reason that they may not be identical is that the variation among the objects whether it is shape or color or texture or illumination or orientation whatever.

So, when you get multiple instances of the objects belonging to the same class it is hardly possible that the feature vectors that we compute given two instances of the objects belonging to the same class those feature vectors will be identical. And because of this variation when I have large number of objects belonging to a particular class and I compute the feature vectors of all those different objects belonging to the same class all these feature vectors are not identical rather in the feature space they will form a sort of distribution.

(Refer Slide Time: 01:48)



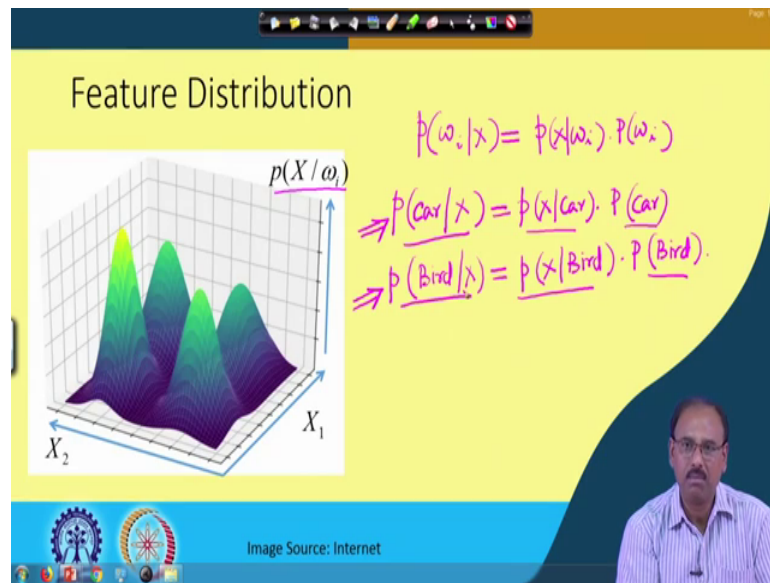
So, in today's class in the previous class what we have talked about this feature space representation, then we have talked about the Bayes rules and then we have also talked about the Bayes minimum error classifier. So, today we will further analyze the remaining part of Bayes minimum error classifier, then we will go to what is known as Bayes minimum risk classifier, and I will also try to discuss whether there is any relation between Bayes minimum error classifier and Bayes minimum risk classifier. We will try to see that.

(Refer Slide Time: 02:28)



So, this is what we have shown in the previous class that is given images from the same class, but multiple number of images belonging to the same class over here. So, we had considered three different classes the class of birds, the class of dogs and the class of cars. You will find that all these representations of these objects in the feature space that forms a cluster or our distribution.

(Refer Slide Time: 03:00)



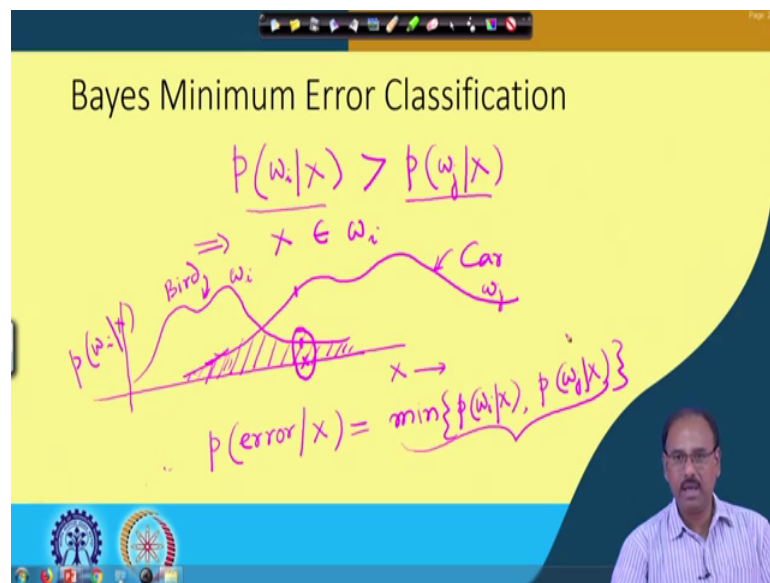
So, we had shown in the other figure in 3D how this distribution looks like and we have said that because we are computing the distribution of the feature vectors belong from objects belonging to a particular class. So, this tells us what is the class conditional probability density function or p of X given ω_i where ω_i is the class and X is the feature vector that we have computed.

So, experimentally by collecting large number of objects from different classes I compute this p of X , given ω_i and at the same time I also have an a priori probability that is what is the probability of occurrence of class ω_i . So, I had two concepts p of X given ω_i and the a priori probability p of ω_i and from this for classification or for recognition the what I have to compute is what is p of ω_i given X and we have shown using Bayes rule that this is nothing, but p of X given ω_i into a priori probability ω_i .

So, given two classes the birds and cars, I have to compute given an unknown feature vector what is p of say car given X and I also have to compute the probability of what is p of bird given the same X and this p of car given X is nothing, but p of X given car which you have already computed through experiments, that is our class conditional probability density multiplied by the a priori probability what is p of car. Similarly, in this case we will compute what is p of X given bird into a priori probability what is p of bird?

So, these are my class conditional probability densities, these are the a priori probabilities and from this we compute the posterior probability p of car given X and p of bird given X . So, out of these two whichever is more for a given unknown X which are vector. So, if I find that p of car given X is greater than p of bird given X then my inference will be that this feature vector X belongs to a car or the object from which this feature vector X has been computed that object is nothing, but a car nothing, but a car.

(Refer Slide Time: 06:04)



So, what we are computing is p of ω_i given X if it is greater than p of ω_j given X then my interpretation is X belongs to ω_i . If p of ω_j given X is greater than p of ω_i given X then my interpretation will be the other ways that is X belongs to ω_j . Now, even in this case you find that if I plot these two in one dimension say p of ω_i given X maybe I plot something like this. So, here what I have is my feature vector X and in this time direction what I have is p of ω_i given x .

So, if my ω_i is bird, then I can have this sort of a posterior probability density whereas for car maybe p of car given X is having some density something like this. So, this is my say ω_i this is my ω_j . So, given any unknown X , say X vector is somewhere over here you find that for this X p of car given X is more than p of bird given X . So, as a result I am deciding that this X belongs to car; it does not belong to

bird. But, still there is a finite probability that X may belong to bird also and that is what is my probability of error.

So, in this case what is the probability of error that I have? The probability of error is nothing, but the minimum of the two. So, p error if I put it write it like this p of error given X is nothing, but minimum of p of omega i given X and p of omega j given X. So, as I am taking the decision which minimizes the error because it is minimum of these two quantities p of omega i given X and p of omega j given X.

So, the classifier that we design is what is known as Bayes minimum error classifier and what is the total error in this case? Total error of classification is nothing, but the error or the area under this curve. So, which is nothing, but if I integrate this area minimum of these two over X varying from minus infinity to infinity, then what is what I get is the total error of this classification rule. That is given by this minimum error classifier ok.

(Refer Slide Time: 09:57)

The slide is titled "Bayes Minimum Risk Classification" and features handwritten mathematical notations in purple ink. The notations are grouped by a large curly brace on the left:

$$\left\{ \begin{array}{l} \omega_i : i = 1 \dots C \\ \alpha_j : j = 1 \dots K \\ X : \rightarrow d\text{-dimensional Feature Vector} \\ \lambda(\alpha_i / \omega_j) \end{array} \right.$$

In the bottom right corner of the slide, there is a small video inset showing a man with glasses and a mustache, wearing a light-colored shirt, speaking.

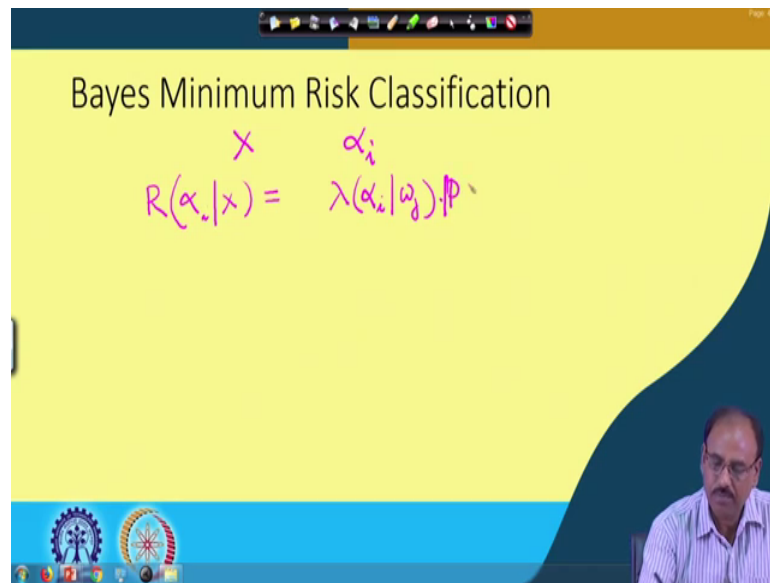
Given this, now I go to the next type of classifier which is Bayes minimum risk classifier. So, Bayes minimum risk classifier is more general from Bayes minimum error classifier. In the sense that in case of Bayes minimum error classifier we have considered only the classes. See, if I have C number of classes then I have classes omega i; i varying from 1 to C. So, I take if I take C number of classes omega i where this I varies from 1 to C, if I have C number of classes.

In case of Bayes minimum risk classifier we consider this ω_i to be the states of nature which are nothing but, classes for our classification problem and we also have a set of actions say α_i where or α_j set of actions α_j ; where j varies from say 1 to capital K . So, I have C number of two states of nature, I have K number of actions α_i , as before I consider the feature X to be d -dimensional feature vector. But, what makes Bayes minimum risk classified more general than Bayes minimum error classified is introduction of a loss function.

So, this loss function is introduced loss function λ which is α_i given ω_j ; that means, if I take an action α_i where the true state of nature is ω_j or the true class is ω_j then the loss that we incur is $\lambda \alpha_i$ given ω_j . So, given this the male Bayes minimum risk classifier works in this fashion. In our classification problem our problem is that given an unknown feature vector X we told earlier that any input signal now we will consider to be a feature vector in my feature space.

So, my classification problem is that given any unknown feature vector X I have to compute or I have to predict to which class ω_i that feature vector belongs. So, that is my classification problem. In case of Bayes minimum risk classifier it is assumed that for every such action you take or for every such prediction that X belongs to a particular class, you have a risk involved in it. So, you compute that risk the value of the risk for every decision that you are taking and the decision that gives you the minimum risk you have to take the corresponding decision. So, it is like this.

(Refer Slide Time: 13:27)



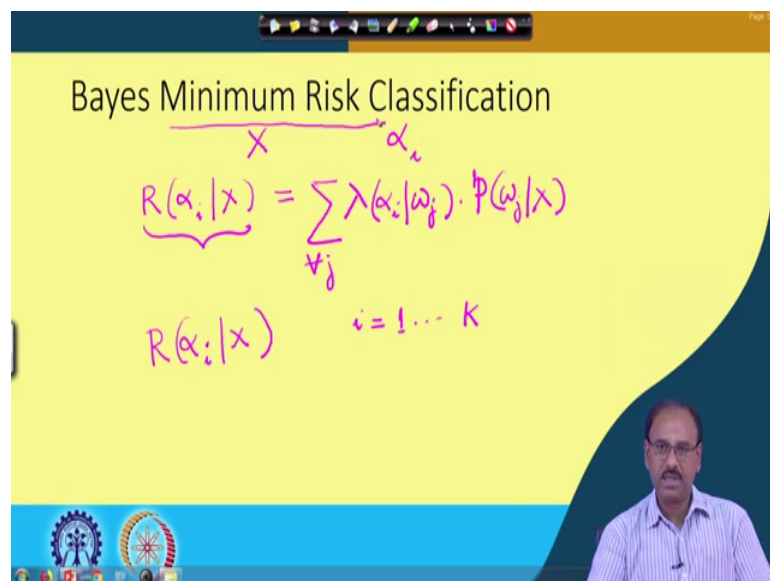
Bayes Minimum Risk Classification

$$R(\alpha_i | X) = \lambda(\alpha_i | \omega_j) \cdot P$$

The slide features a yellow background with a dark blue curved shape on the right side. At the bottom, there is a blue bar containing several logos, including the Indian Institute of Technology (IIT) logo. A small video inset in the bottom right corner shows a man with glasses and a mustache, wearing a light blue shirt, speaking.

So, again I assume that I have a given feature vector X and on this X , I take an action α_i . So, the risk involved in taking action α_i given X can be computed as if the true state of nature is say ω_j then we said that we incur a loss $\lambda(\alpha_i | \omega_j)$ because α_i is the action that I am taking whereas true state of nature is ω_j , it should have been ω_j . So, for while taking this action, I incur some loss. So, that is my loss function $\lambda(\alpha_i | \omega_j)$ given $\lambda(\alpha_i | \omega_j)$ into what is the probability P of sorry, let me rewrite.

(Refer Slide Time: 14:25)



Bayes Minimum Risk Classification

$$R(\alpha_i | X) = \sum_{\forall j} \lambda(\alpha_i | \omega_j) \cdot P(\omega_j | X)$$

$R(\alpha_i | X) \quad i = 1, \dots, K$

The slide features a yellow background with a dark blue curved shape on the right side. At the bottom, there is a blue bar containing several logos, including the Indian Institute of Technology (IIT) logo. A small video inset in the bottom right corner shows a man with glasses and a mustache, wearing a light blue shirt, speaking.

So, I have been given vector X and I take an action α_i . So, I compute the risk R of α_i given X which is nothing, but the risk involved for taking action α_i if the true state of nature is ω_j multiplied by a posterior probability p of ω_j given X . And you take the sum of this over all j because I am taking action α_i may be the actual true state of nature is ω_1 the signal actually belongs to class ω_1 .

So, for that what is the risk? It may actually belong to class ω_2 for that what is the risk? And likewise it may actually belong to class ω_C and therefore, they for that what is the risk? And if I add all these risks then I get the overall risk for taking action α_i given my feature vector X right. So, this I have to take. So, I have to compute this R of α_i given X for all i , we said that we have K number of actions so, for all i varying from 1 to K .

And out of all these for whichever R of α_i given X is minimum I have to take that corresponding action. So, that is why it is based minimum risk classification that is I want to take that particular action for which my risk is minimum unlike in case of Bayes minimum error classification. There you decided X to belong to a particular class which minimized your error. So, now, it is minimization of risk. Now, let us see whether I can establish any relation between this minimum risk classification and minimum error classification.

(Refer Slide Time: 16:57)

Bayes Minimum Risk Classification

$$\lambda(\alpha_i | \omega_j) = \lambda_{ij}$$

$$\left. \begin{array}{l} \lambda_{ij} = 0 \quad i=j \\ \lambda_{ij} = 1 \quad i \neq j \end{array} \right\}$$

$$R(\alpha_i | X) = \sum_{\forall j} \lambda_{ij} \cdot P(\omega_j | X)$$

$$= \sum_{j \neq i} P(\omega_j | X) = 1 - P(\omega_i | X)$$

Let us assume that we have an one-zero loss function that is I assume that λ_{ij} given ω_j , let me represent this in short as λ_{ij} . So, if i is equal to j ; that means, I am correcting I am taking the correct action it belongs to class ω_j and I am saying that and I am my decision is it belongs to class ω_j and it actually belongs to class ω_j . So, in that case the loss function the loss that i incur is 0.

So, I assume that this λ_{ij} the loss function is equal to 0, if i is equal to j and I also assume that this is equal to 1 if i is not equal to j . So, that is for every incorrect action you incorporate our unity loss and which is same for all incorrect decisions. But, if your decision is correct; obviously, you are not incurring any loss. So, λ_{ij} is equal to 0. So, given this now if I compute the risk involved in taking an action α_i given X which we have said that this is nothing, but λ_{ij} that is λ_{ij} of α_i given ω_j into $P(\omega_j | X)$ and some of this over all j .

And now coming taking these values of λ_{ij} you find that we have defined that wherever i is equal to j λ_{ij} is equal to 0, and wherever i is not equal to j λ_{ij} is equal to 1. So, this expression simply becomes $P(\omega_j | X)$ take the summation over all j not equal to i because where even j is not equal to i λ_{ij} is 1 and whenever j is equal to i λ_{ij} is equal to 0. So, I have to take this summation over all j , where j is not equal to i for all of them λ_{ij} is 1.

And this is nothing but $1 - P(\omega_i | X)$. Now, see that what we wanted is in Bayes minimum risk classification, I want to take that particular action α_i for which $R(\alpha_i | X)$ is minimum. In Bayes minimum risk classification we wanted to classify X to that particular class for which $P(\omega_i | X)$ is maximum. Now, if you look at this expression that $R(\alpha_i | X)$ is equal to $1 - P(\omega_i | X)$.

Obviously, you can find out you can check that $R(\alpha_i | X)$ will be maximum or $R(\alpha_i | X)$ will be minimum when $P(\omega_i | X)$ is maximum because $R(\alpha_i | X)$ is nothing, but $1 - P(\omega_i | X)$. So, wherever $R(\alpha_i | X)$ is minimum $P(\omega_i | X)$ is maximum.

So, in this particular case when my loss function is one-zero loss function or zero-one loss function that is for every correct decision I assume that I incur a loss or I do not incur any loss that is the loss function value is 0, for every incorrect decision I incur

unity loss. So, under that condition my Bayes minimum risk classifier and Bayes minimum error classifier both of them are same. But, in general for taking wrong decisions for different types of wrong decisions my loss function will not be same right. So, we will discuss about that more later.

(Refer Slide Time: 21:51)

$$R(\alpha_i | X) = \sum_{\omega_j} \lambda(\alpha_i | \omega_j) \cdot P(\omega_j | X)$$

α_i ω_j X

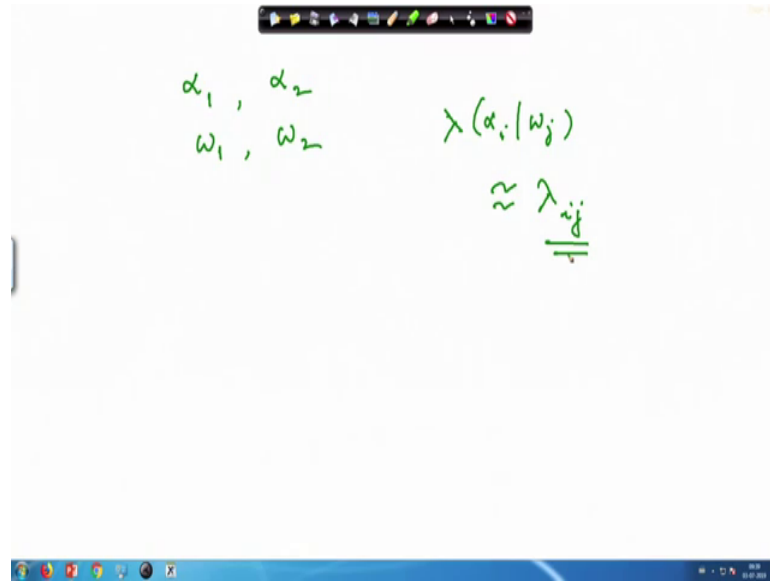
K

$\alpha_i : i = 1, \dots, K$

Now, let us consider case of two classes. So, we are discussing about the risk function and we have defined that if you take an action say alpha i where the two state of nature is say omega j then for and given input vector X the risk involved is given by R of alpha i given X which is nothing, but lambda alpha i given omega j into P of omega j given X and you take the sum of this over all j. So, that is the total risk for taking an action alpha i given an input which are vector x.

So, if I have say K number of such actions where alpha i, i varying from 1 to K; so, for each of this i I have to compute what is the risk function and I have to take that particular action for which the risk involved is minimum. So, that is what Bayes minimum risk classifier says.

(Refer Slide Time: 23:13)



So, now let us take two class problem where I assume that I have two actions given as alpha 1 and alpha 2 and I have two states of nature or two classes given by omega 1 and omega 2. So, let us compute try to find out that what is the risk involved if I take an action alpha i or what is the risk involved if I take an action alpha 2. So, alpha 1 and alpha 2, I want to compute the two risk functions and the loss function that we have defined lambda alpha i given omega j for simplicity I write this in the form of lambda ij.

So, the risk or the loss function for taking an action alpha i if the true state of nature is omega j which is lambda alpha i given omega j, I represent this as lambda ij. So, given this now I will have two risk functions involved, one is for taking action alpha i the other one is for taking action alpha j.

(Refer Slide Time: 24:25)

$$\begin{aligned}
 R(\alpha_1|X) &= \lambda_{11} P(\omega_1|X) + \lambda_{12} P(\omega_2|X) \\
 R(\alpha_2|X) &= \lambda_{21} P(\omega_1|X) + \lambda_{22} P(\omega_2|X) \\
 R(\alpha_1|X) &< R(\alpha_2|X) \\
 \Rightarrow \lambda_{11} P(\omega_1|X) + \lambda_{12} P(\omega_2|X) &< \lambda_{21} P(\omega_1|X) + \lambda_{22} P(\omega_2|X) \\
 \Rightarrow (\lambda_{21} - \lambda_{11}) \cdot P(\omega_1|X) &> (\lambda_{12} - \lambda_{22}) P(\omega_2|X)
 \end{aligned}$$

So, I have to compute R of alpha 1 given omega 1 and I also have to compute R of alpha 2 given sorry R of alpha 1 given X and I also have to compute R of alpha 2 given X. So, this R of alpha 1 given X is nothing, but lambda 11 P of omega 1 given X plus lambda 12 P of omega 2 given X. So, here you find that lambda 11 is nothing, but lambda alpha 1 given omega 1, similarly lambda 12 is nothing, but lambda alpha 1 given omega 2 and in the same form I can write all of R of alpha 2 given X as lambda 21 P of omega 1 given X plus lambda 22 P of omega 2 given X.

So, given these two risk values, my decision will be in favor of action alpha 1 or deciding that input X belongs to class omega 1 is when I find that R of alpha 1 given X is less than R of alpha 2 given X. That is the risk involved in taking action alpha 1 is less than the risk involved in taking action alpha 2.

So, if I put this bringing or using the risk values from here I have to have lambda 11 P omega 1 given X plus lambda 12 P omega 2 given X this has to be less than lambda 21 P of omega 1 given X plus lambda 22 P of omega 2 given X. Or I can derive this as lambda 21 minus lambda 11 P of omega 1 given X has to be greater than lambda 12 minus lambda 22 P of omega 2 given X right.

So, this is the condition that has to be satisfied that is lambda 12 minus lambda 21 minus lambda 11 into P of omega 1 given X has to be greater than lambda 12 minus lambda 22 P of omega 2 given X. So, let me just rewrite this in the form as let me or refresh this.

(Refer Slide Time: 28:11)

$$\begin{aligned} & \lambda_{11} P(\omega_1|X) + \lambda_{12} P(\omega_2|X) < \lambda_{21} P(\omega_1|X) + \lambda_{22} P(\omega_2|X) \\ \Rightarrow & (\lambda_{21} - \lambda_{11}) P(\omega_1|X) > (\lambda_{12} - \lambda_{22}) P(\omega_2|X) \\ \Rightarrow & \frac{P(\omega_1|X)}{P(\omega_2|X)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \quad \left\{ \begin{array}{l} P(\omega_1|X) = P(X|\omega_1) \cdot P(\omega_1) \\ P(\omega_2|X) = P(X|\omega_2) \cdot P(\omega_2) \end{array} \right. \\ \Rightarrow & \frac{P(X|\omega_1)}{P(X|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \cdot \frac{P(\omega_2)}{P(\omega_1)} \end{aligned}$$

So, my condition was that $\lambda_{11} P(\omega_1|X) + \lambda_{12} P(\omega_2|X)$ has to be less than $\lambda_{21} P(\omega_1|X) + \lambda_{22} P(\omega_2|X)$. So, which we have rewritten in the form $(\lambda_{21} - \lambda_{11}) P(\omega_1|X)$ has to be greater than $(\lambda_{12} - \lambda_{22}) P(\omega_2|X)$. So, this is the condition that has to be satisfied for taking a decision in favor of class ω_1 or for taking action α_1 .

So, I can also rewrite this in the form $\frac{P(\omega_1|X)}{P(\omega_2|X)}$ to be greater than $\frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}}$. And you remember that λ_{12} is the loss function for taking an action α_1 when the actual true of nature is ω_2 and λ_{22} is the loss function involved when you are taking action α_2 when the true class of nature is ω_2 . So, naturally λ_{22} has to be less than λ_{12} . Similarly, λ_{11} will also be less than λ_{21} .

So, both these quantities both in the numerator and denominator on the right hand side both of these quantities are positive. And again you remember from the Bayes rule that $P(\omega_1|X)$ which is the posterior probability I can write this as $P(\omega_1|X)$ as $\frac{P(X|\omega_1) \cdot P(\omega_1)}{P(X)}$ which is the class conditional probability into the a priori probability $P(\omega_1)$. Similarly, $P(\omega_2|X)$ can also be written as $\frac{P(X|\omega_2) \cdot P(\omega_2)}{P(X)}$.

So, using this now this expression can be written as $P(X|\omega_1) \geq P(X|\omega_2)$ has to be greater than $\lambda_{12} - \lambda_{22}$ upon $\lambda_{21} - \lambda_{11}$ into $P(\omega_2)$ upon $P(\omega_1)$. So, now, considering this $P(X|\omega_1)$ to be a function of ω_1 , $P(X|\omega_1)$ gives me the likelihood value and accordingly $P(X|\omega_1) \geq P(X|\omega_2)$ that gives me the likelihood ratio.

So, this expression that in order to take an action in favor of class ω_1 which is $P(X|\omega_1) \geq P(X|\omega_2)$ has to be greater than $\lambda_{12} - \lambda_{22}$ upon $\lambda_{21} - \lambda_{11}$ into $P(\omega_2)$ given ω_1 . This condition has to be true for taking an action in favor of class ω_1 and that is what comes from the Bayes minimum risk classification rule. So, going by that you find that on the right hand side of this expression of this inequality that is $\lambda_{12} - \lambda_{22}$ upon $\lambda_{21} - \lambda_{11}$ into $P(\omega_2)$ upon $P(\omega_1)$ this is independent of X .

So, a favorable decision in favor of class ω_1 can be that if the likelihood ratio is greater than certain threshold, where the threshold is given by this. The thresholds are in terms of the loss functions and the a priori probabilities. So, we can say that if the likelihood ratio is above then this threshold then we take an action in favor of class ω_1 . So, we will continue this discussion further.

So, in today's lecture what we have discussed about, we have recapitulated our previous lectures content that is featured representation of a given signal; then we have talked about the Bayes theory and the Bayes minimum error classifier; Bayes minimum risk classifier and we have also tried to establish that what is the relation between Bayes minimum error classification and Bayes minimum risk classification.

Thank you.