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

Lecture - 04
Bayesian Learning- I

Hello welcome to the NPTEL Online Certification Course on Deep Learning. You recall in the previous class, what we did is we discussed about the different descriptors or the features that can be extracted from a given signal, whether the signal is a visual signal like an image or what we can see around the world or it can also be a voice signal like speech signal. And the applications are in the object recognition, object classification, understanding the world, speech identification, speaker identification, then speech to text conversion and many such applications.

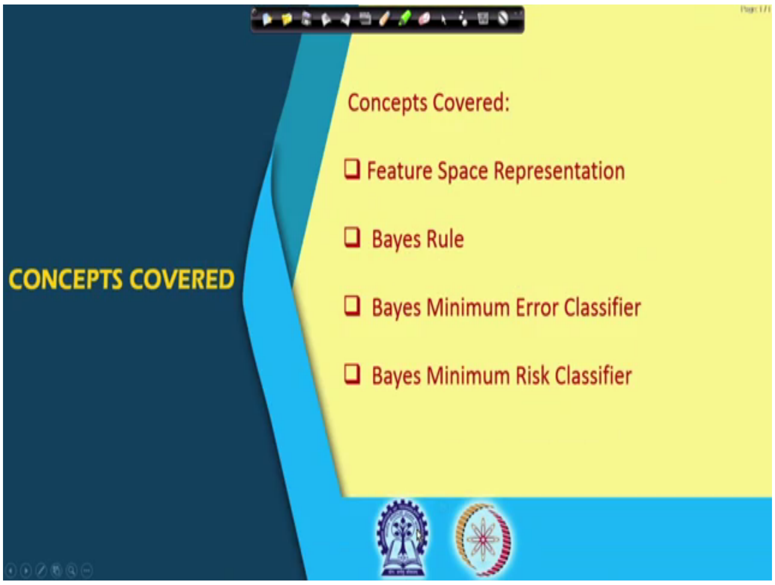
So, what we talked about in the previous two lectures is that given any such signal, we can extract the descriptors or the features from that signal which represent that particular signal, whether it is visual or auditory signals. And once you represent a signal using a number of features or a number of descriptors, those features can be represented in the form of a vector. So, if I extract say three features from any given signal, the features may be obtained from the shape or may be obtained from the region like its intensity, its color or its texture.

So, those three different features; if I put in the form of a vector, it becomes a vector in one dimension having three different components. Or in other words the signal is represented by a point in a three dimensional feature space or by a vector in a three dimensional feature space. So, in general for a given signal, you extract a number of such features.

So, you extract d number of such features or d number of such descriptors and put those descriptors in a particular order to get a vector having d number of elements or a d dimensional vector. So, once I have a d dimension d dimensional vector so, effectively what I am doing is I am transforming my signal into a feature space or a vector space where in that d dimensional vector space, our signal will be represented by a particular vector. And once I present the signal by a particular vector which is nothing, but a point in my d dimensional space and if I have multiple number of such signals, then for each of those signals; I will have a corresponding point or a corresponding vector in a d dimensional space.

So, once I have this feature representation or vector representation, then just to measure whether two different signals are similar or two different signals are dissimilar, I can simply try to find out what is the distance between those two vectors or what is the difference of those two vectors. So, if the difference is very high or the distance between two vectors is very high; in that case, I can infer that the signals are not similar. Whereas, if the distance is very small; that means, that points are very close to each other, then I can infer that those two signals are very close or they are similar.

(Refer Slide Time: 03:55)



CONCEPTS COVERED

Concepts Covered:

❑ Feature Space Representation

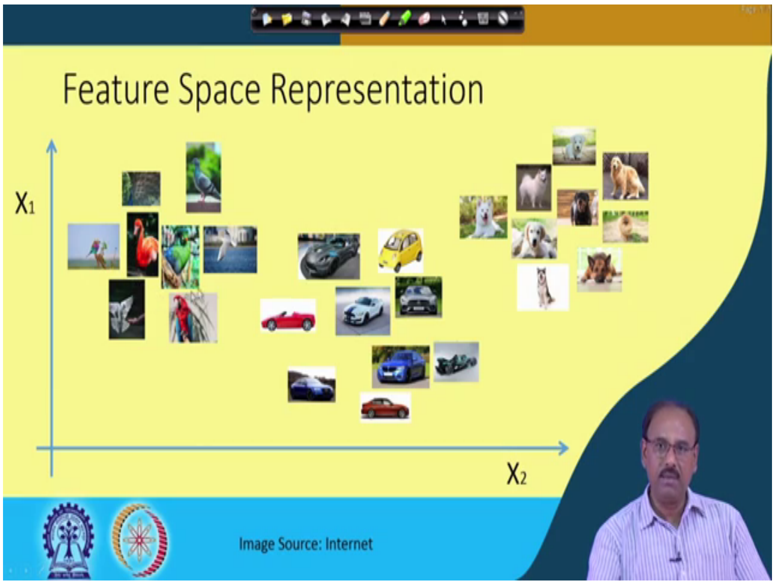
❑ Bayes Rule

❑ Bayes Minimum Error Classifier

❑ Bayes Minimum Risk Classifier

So, in today’s lecture what we will discuss is we will have the concept of features based representation of any given signal. Then to understand or to recognize those signals, we will discuss about Bayes rule. Then we will have two types of classifiers or two types of recognizers; one is Bayes minimum error classifier, the other one is based minimum risk classifier. So, let us see what this feature space representation means?

(Refer Slide Time: 04:31)



Feature Space Representation

Image Source: Internet

So, for the time being let us assume that a signal and to demonstrate this, we will mostly take visual signals. So, let us assume that the visual signal is represented by a two dimensional feature vector having components X 1and X 2 or in other words, we have extracted; let us assume that we have extracted only two features; one feature may be from the shape of the object, the other feature might have been obtained from the intensity of that particular object. So, had two different features; one is feature X 1, the other one is feature X 2.

So, given these two features, now let us assume that I have visual signals or have images or birds, I have images of cars, I have images of dogs and so on. So, if I get one image of a bird as we said that this image of this signal will be represented by a vector or a point in my feature space. So; obviously, here you can see that to represent that image of a bird, just two dimensional feature vector is not sufficient because the birds have different shapes, the birds have different colors, they have different intensity levels, a bird may be obtained from various orientations; it might have been observed from various orientations.

So, accordingly I have to have a large number of features to describe the bird, but since I cannot show any dimension on a paper of more than two. So, I am taking a simplistic case that I am assuming that this bird is represented by a two dimensional feature vector. So, given that I if I take one picture, that picture is represented over here in my feature vector space which is given by X 1 and X 2. If I take another image of some other bird, you find that here; this bird or this picture is also placed at a vector location which is very close to the previous location of the bird.

But you notice that these two locations of these two vectors are not identical because the images are not identical. Similarly, if I take another one, I will again get another vector which is close to the previous two vectors. Now if I take image of a car, you will find that the in a car is represented by a vector over here which is far away from the vector representation of the birds.

And now if I find out the distance or the difference of this vector representing a car and a vector representing a bird, the modulus of the vector difference will be quite high or in other words the distance between these two will be quite high, which is again an indication that these two images are not similar or a car or a bird neither in terms of shape and nor in terms of color, they are similar.

Again if I take a second car, I get a vector over here vector representation, but again it is not identical to the previous location of the car; the reason being the images are not identical. I can have higher I can have cars of various colors, of various shapes, of various sizes; the car images may have been obtained from various orientations.

So, accordingly their appearance will change and the descriptors or the features that we compute that will also be changed. If I take image of a dog, you will find that a dog may be represented by a vector somehow in this location. If I take another dog, this dog is also represented by a vector in this location in my feature space.

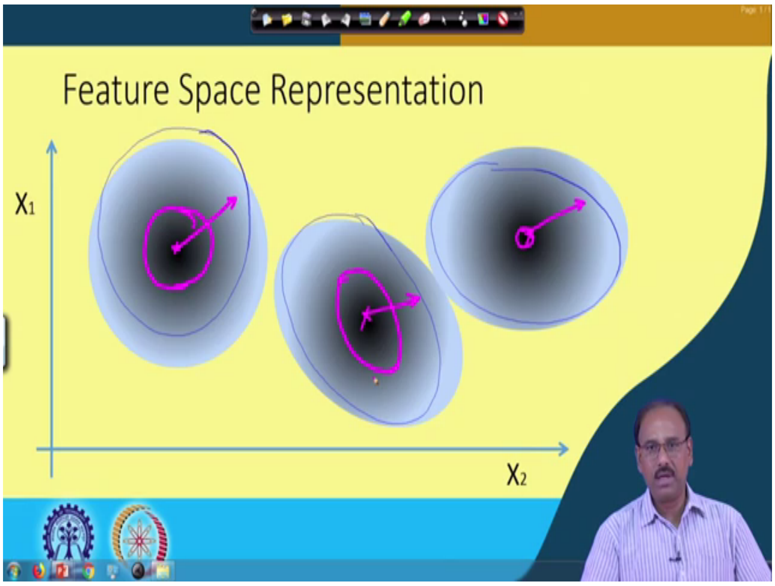
So, like this and again you find that the distance between the first dog and that is and the second dog, this distance is smaller than the distance between the dog and the bird and the distance between dog and the car which again indicates that this first image the image of the first dog and the second dog image, they are very similar. Whereas, this dog image and the car image or the dog image and the bird, they are dissimilar.

So, I can continue like this and you find that for a large number of images of birds or a large number of images of cars or a large number of images of dogs, each of these vectors they represent a cluster of vectors. So, all the images of dogs that forms a cluster, all the images of cars that forms a cluster in my vector space, all the images of birds form another cluster in my vector space or feature space. And obviously, as I said that all of them will not be represented by the same vector; all these bird images are not represented by the same vector. But they are because of variation in the images I can have images of different birds of different colors, they may have different intensity values.

So, there will always be a variation in the vector representation. Similarly there will always be a variation in the vector representation of cars, there will always be a variation in the vector representation of the dogs. So, as a result what we get is given images of the signals belonging to I can call that this is a class of signals ok. Similarly this is another class of signals which represents the class car, this is another class of signals which represent from the class dog right.

So, you will find that the vectors coming from a particular class that forms a cluster of feature vectors or a cluster of points in my feature space. The same is true in case of bird, the same is true in case of cars, the same is true in case of dogs. And naturally there is a distribution of these points in the feature space right. So, what we have; if I put in the other way every such image represented by a point or by a vector in the feature space, I can put this image in this form.

(Refer Slide Time: 11:05)



Feature Space Representation

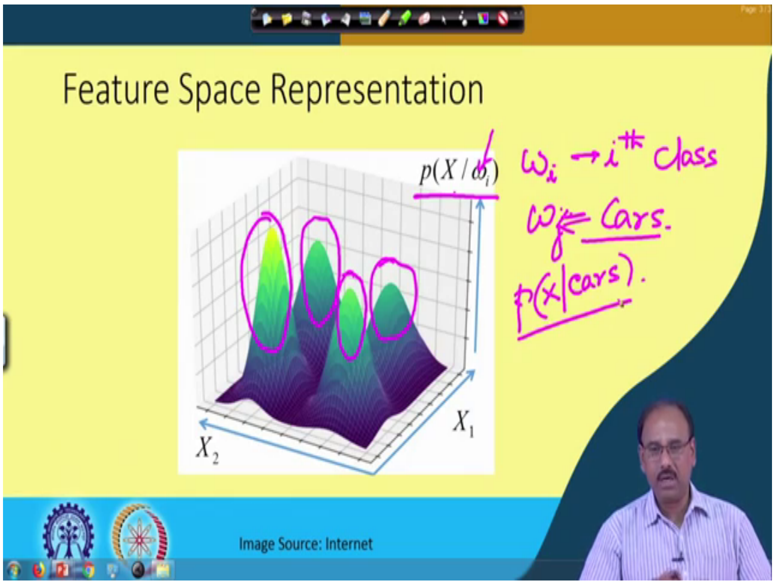
So, here you find that over here this represents one cluster of points similarly this represents another cluster of points or cluster of feature vectors, this represents another cluster of points. And when I have these different cluster points, every clustered is represented by a distribution and that distribution ah; if you look at these figures, the density of the feature vectors at the center is very high. As you move away from the center as the distance between the center and the other point increases, you find that the density goes on it goes on decreasing.

Similarly over here at the center the density is high which is actually mean location; mean of the cluster of vectors. And as I move away from the center, the density goes on reducing. Similarly in this case at the center, the density is high as you move away from the center; the density goes on reducing.

So, this indicates that all this feature vectors follow a particular probability density function. And in this particular case in this diagram as I have shown it in a two dimensional feature space, this probability density is a elliptic distribution. In one case, it is circular in other two cases, those are elliptic. If I go for three dimensional feature space, then this will be represented by spherical distribution or ellipsoidal distribution. When the dimension becomes even more than 3, which I cannot possibly draw on a two dimensional plane, I represented those as or term those as hyper spherical or hyper ellipsoidal.

So, what of the spheres or the ellipsoids mean? See for example, here if I take a contour of equal probability density functions that becomes a circle. Similarly in this case; if I take a contour of equal probability density values, it becomes an ellipse and as I move away from the center, the value of the probability density goes on reducing. So, that is what is meant by these and different diagrams.

(Refer Slide Time: 13:52)



Feature Space Representation

$\omega_i \rightarrow i^{th}$ class

ω_k = Cars.

$p(x|cars)$.

Image Source: Internet

Now, if I go further in a three dimensional representation, those densities can be represented by surface plots. So, as you see over here, this one may represent the probability density or distribution of the vectors coming from a particular class of objects. This class may be the class of birds, this distribution may be coming from a class of dogs, this distribution may be coming from a class of cars, this distribution may be coming from a class of bicycles and so on.

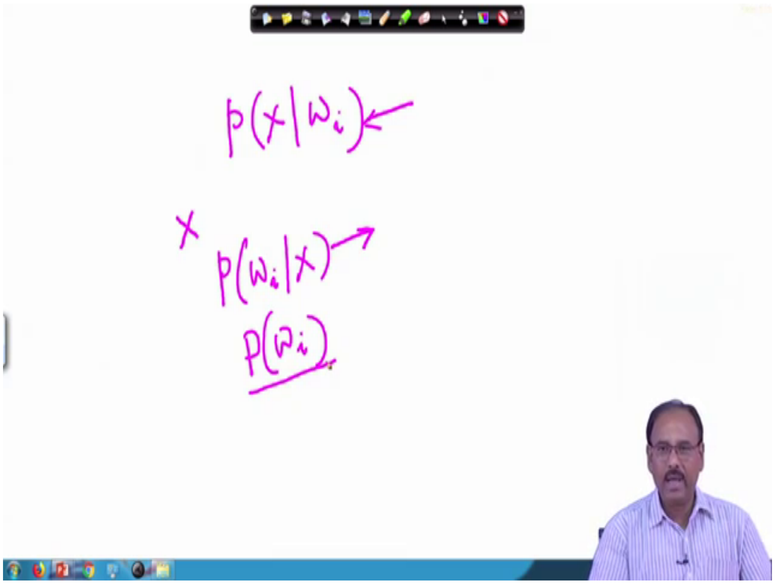
So, this diagram through this diagram what I wanted to show that the distribution that you get is dependent on the class. So, if I take any value of X so, what I get is that what is if I take any feature vector from a class of say dogs, what is the probable value of X? If I take a feature vector from a class of cars, what is the probable value of that feature vector X? So, that is the distribution which is given by this figure. So, or in other words what I get is I get a class conditional probability density function which is known as p of X given omega I, this omega i, I represent as i’th class.

So, as I said that this class may be class of dogs, this class may be class of birds, this class may be class of cars and so on. So, what I have is a class conditional probability density function. And how do you get this class conditional probability density function? This is a part of training. What I do is I take a large number of images belonging to bird and for all those images, I compute what is the feature vector X. And then I find out what is the distribution of those feature vectors X in my feature space. And because all these feature vectors I obtain from images from the class birds so, this distribution is for the feature vectors coming from the class bird.

So, I get p of the feature vector X given omega i where this particular omega i is nothing, but class birds. Similarly, if I take a large number of images from of cars find out the corresponding feature vectors and find out the distribution of all those feature vectors again in the feature space. So, in this case, the class of cars I write as say class omega j which indicates cars. So, the distribution of the feature vectors that I get in this particular case is nothing, but the probability density of the feature vector X given the class as cars.

So, these are all the class conditional probability density functions which I have to compute from a large number of signals or large number of images with examples that we are taking from a particular class. So, I represent this as this class conditional probability density function and once I have this, then given an unknown image. If I want to recognize that what that image is, then I have to make use of this class conditional probability density function to find out that given a signal which class it belongs to that is given by class conditional probability density function p of omega j or p of omega i.

(Refer Slide Time: 18:14)



$p(x|\omega_i) \leftarrow$

x

$p(\omega_i|x) \rightarrow$

$\underline{p(\omega_i)}$

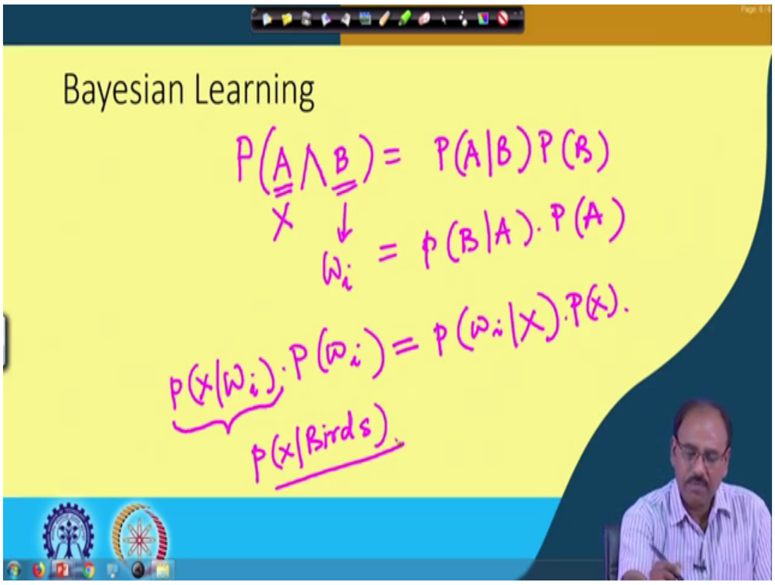
So, what I have is p of X given omega i or p of X given omega j. What X is feature vector and omega i or omega j is the class. During classification what I wanted to have is or what I need to do is that given an unknown feature vector, I have to identify whether this unknown feature vector is coming from a bird or this unknown feature vector corresponds to a car or this unknown feature vector corresponds to a dog. So, what I need to compute is what is p of omega i given X.

So, what I have in during training is p of X given omega i and what I need to compute is p of omega i given X. In addition to this there is another concept. So, this is what is known as class conditional probability density function and this is what is known as a posterior probability density function. In addition to this there is another concept of a priori probability of classes; that means, what is the probability that in a given domain, a sample will belong to the class omega i or p of omega which is the a priori probability.

So, as an example just to give you an example that how this may occur. See if I go to a forest and I see an object then after looking at the object; if I have to infer, whether that object or that animal or whatever I am seeing in the forest should be a car or it should be a bird? And it is quite obvious that if I am visiting a forest, it is more likely to see a bird than to see a car because you cannot expect a car to enter a forest until and unless there is a motorable road whereas, the birds are abundant in the forest.

So, given any object the a priori probability or the probability that object of the what I am seeing will be a bird. If I have to compare between bird and a car that the a priori probability that it will be a bird is more than the probability that will be that it will be a car and that is what is this p omega i or a priori probability. Now given this I want to go for how the classification of the objects of the signals can be done given the feature vectors belonging to different objects.

(Refer Slide Time: 21:14)



Bayesian Learning

$$P(\underset{X}{\underline{\underline{A}}} \wedge \underset{\underset{\omega_i}{\downarrow}}{\underline{\underline{B}}}) = P(A|B)P(B)$$

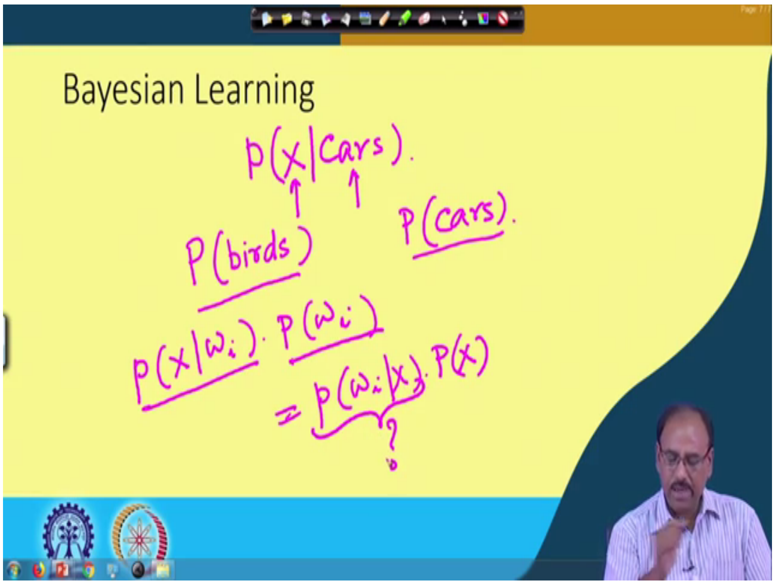
$$= p(B|A).P(A)$$

$$\underbrace{p(x|\omega_i)}_{\underline{p(x|\text{Birds})}}.P(\omega_i) = p(\omega_i|x).P(x).$$

So, for that what I will have is Bayesian learning. So, before we go for what is Bayesian learning, let us just recapitulate probability rule that the probability says that if I have two events say event A and event B. The probability of event A and B can be probability of a given B into probability of B which is also same as probability of B given A into p of A. So, the Bayes rule or the Bayes decision, Bayes decision theory is based on this probability rule. So, in this case we can say that this X is my feature vector; A is my feature vector X and B is our class omega i.

So, what does it mean? That if I have been given a feature vector of an object or of an image belonging to class omega i, then what is the probability that both of them occur together? So, which is nothing, but p of X given omega i into p of omega i which is the a priori probability of class omega i and that is also same as p of omega i given X into a priori probability p of X. So, you recall that what I said in just few minutes ago that through training what I compute is p of X given omega i. That means, if I take a large number of images of birds and I compute the distribution of the feature vectors computed from all those bird images, then actually what I am computing is p of X given the class of objects as parts.

(Refer Slide Time: 24:06)



Bayesian Learning

$p(x|\text{cars})$.

$P(\text{birds})$ $\quad$ $P(\text{cars})$.

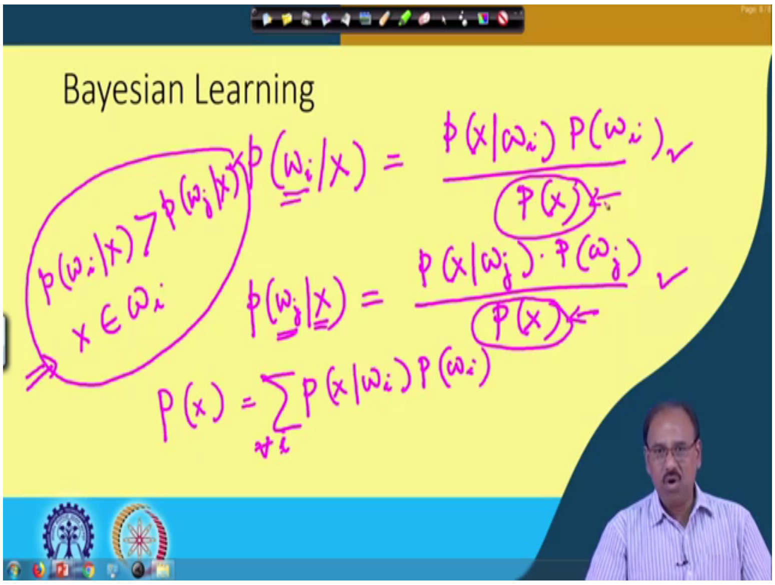
$$p(x|\omega_i).P(\omega_i)$$

$$= \underbrace{p(\omega_i|x)}_{?}.P(x)$$

Similarly, if I take a large number of images of cars and compute the feature vectors from those images and find out what its distribution, I compute what is the distribution of those feature vectors computed from images of cars; what I am actually computing is the class conditional probability p of X given cars. What this X is the feature vector and cars are my class or class of objects. So, what I computed is basically, class conditional probability density function. In addition to that in that particular situation or in a given situation, I also have the a priori probability of birds and I also have a priori probability of cars.

So, what does it mean? If my domain is a forest; that means, I have visited a forest and there what is the probability that an object that you see is a car or it is a bird, then those are the a priori probability is p of birds and p of cars. So, given these two, now using that probability theory I can say that p of X given omega i into p of omega i is equal to p of omega i given X into p of X. So, this p of X given omega i, there is the class conditional probability that I have computed p of omega i is the a priori probability. But what I need to compute is p of omega i given X that is a posterior probability.

(Refer Slide Time: 25:52)



Bayesian Learning

$$p(\omega_i|x) > p(\omega_j|x) \Rightarrow x \in \omega_i$$

$$p(\omega_i/x) = \frac{p(x|\omega_i)\,P(\omega_i)}{\boxed{P(x)}} \checkmark$$

$$p(\omega_j|x) = \frac{p(x|\omega_j)\cdot P(\omega_j)}{\boxed{P(x)}} \checkmark$$

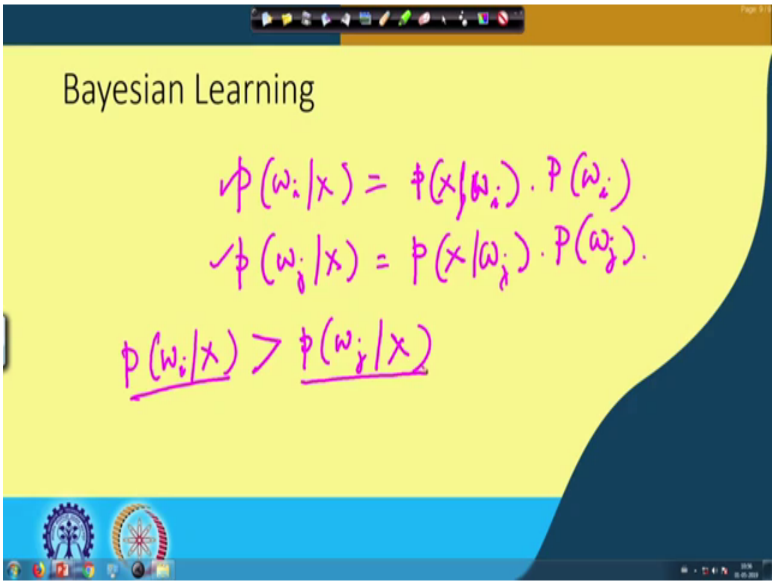
$$P(x) = \sum_{\forall i} p(x|\omega_i)\,P(\omega_i)$$

So, this computation p of omega i given X is nothing, but p of X given omega i into p of omega i upon p of X. So, if I compute the same; suppose omega i is the class of birds and I have another class omega j which is the class of cars this is nothing, but p of X given omega j into p of omega j that is a priori probability upon p of X where what is this p of X? P of X is nothing, but p of X given omega i into a priori probability omega i, take the summation over all I. And if you look at these two expressions p of omega i given X and p of omega j given X, you find that p of X appears in the denominator of both of these expressions.

So, here what I need to do is I have to compute this a posterior probabilities p of omega i given X; that means, given X what is the probability that if this X belongs to class omega i or given X what is the probability that X belongs to class omega j. So, obviously, out of these two whichever is more I will classify or I will associate this input vector X to that corresponding class. So, if I find that p of omega i given X is greater than p of omega j given X then; obviously, my interpretation will be that X belongs to class omega i. This is the interpretation that I have.

So, given these two expressions of p of omega i given X and p of omega j given X, you will find that this p X appears in the denominator both of them. So, this p X does not contribute anything in discrimination.

(Refer Slide Time: 28:24)



Bayesian Learning

$$✓\,p(\omega_i|x) = p(x|\omega_i)\cdot P(\omega_i)$$

$$✓\,p(\omega_j|x) = p(x|\omega_j)\cdot P(\omega_j).$$

$$\underline{p(\omega_i|x)} > \underline{p(\omega_j|x)}$$

So, I can simply compute p of omega i given X is equal to p of X given omega i into a priori probability omega i and p of omega j given X is equal to p of X given omega j into a priori probability p of omega j. So, I can simply compute this which is an approximation of p of omega i given a X because I have ignored the denominator which is p of X. And out of these two whichever is larger, I will associate my input vector to that workers morning class.

And you find that given an X; if I have p of omega i given X to be greater than p of omega j given X. I still have; so, I am deciding that X is associated with class omega i, but I still have a finite probability that X may belongs to cos omega j and that is what is my probability of header. But I am taking decision in favor of that particular class for which a posterior probability is more and my probability of error is the probability of X belonging to the other class.

So, in this particular case, the probability of error is the probability p of omega j given X which is minimum of the two. So, this particular classifier that we have discussed today is what is known as Bayes minimum error classifier. So, in this lecture what we have done is we have talked about the vector representation of the different signals and then, we had talked about the Bayes rule and discussed about the Bayes minimum error classifier.

So, in the next class we will talk about Bayes minimum risk classifier and we will continue further.

Thank you.