Course Name: Machine Learning and Deep learning - Fundamentals and Applications

Professor Name: Prof. M. K. Bhuyan

Department Name: Electronics and Electrical Engineering

Institute Name: Indian Institute of Technology, Guwahati

Week-4

Lecture-16

Welcome to NPTEL online course on machine learning and deep learning fundamentals and applications. In my last few classes, I explained the concept of the discriminate function and also I explained how to take a classification decision with the help of the discriminate function. If I have the information of the class conditional density, then I can easily determine the discriminate function. That means, if I have the information of the parametric form of the probability density function, I can determine the discriminate function. So, in my one example, I considered normal distribution that is the Gaussian distribution and corresponding to this I determine the discriminate function. Also based on the covariance matrix, I have different discriminate functions.

If the covariance matrix for all the classes are same, then I will be getting a linear discriminate function. On the other hand, if the covariance matrix for all the classes are not same, then I will be getting a quadratic discriminate function. Also I explain how to determine the decision boundary between the classes for these cases. Case number one is if I consider the covariance matrix for all the classes are same, that is the case number one and another case I can consider the covariance matrix are not same for all the classes.

So, based on this I can determine the decision boundary between the classes. So, that means, if I have the information of the parametric form of the probability density function, I can determine the discriminate function. So, the parameters we can determine with the help of these two popular algorithms. One is the maximum likelihood estimation technique and another one is the Bayesian estimation technique. That is the parametric form is known, but I have to determine the values of the parameters.

So, for this I have to consider these popular techniques. One is the maximum likelihood estimation technique and another one is the Bayesian estimation technique. Suppose, if I do not have the information of the parametric form of the probability density function, then also I can determine the discriminate function. If the classes are linearly separable,

then with the help of the training samples of different classes, I can determine the discriminate function. So, that concept I am going to explain today.

So, what is the concept? We do not have the information of the parametric form of the probability density function that is the class conditional density. We do not have the information of the probability density function that is the parametric form is not known. In this case, I can determine the discriminate function with the help of the information of the training samples of different classes and we are assuming that the classes are linearly separable. So, that concept I am going to explain today. So, let me begin this class.

So, we are discussing now the linear discriminate function. So, these classes are linearly separable. So, that is why we are considering the linear discriminate function. So, we do not have the information of class conditional density and the classes are linearly separable and we can design a linear discriminate function. So, the linear discriminate function maybe I can write like this gx is the discriminate function, w is the weight vector, x is the input feature vector and w naught is the bias.

$$g(\underline{X}) = \underline{w}^T \underline{x} + w_0$$

this w naught and that is actually the bias or maybe I can consider the threshold weight. So, this x is a d dimensional feature vector. So, this equation this is a linear equation if you see this equation this is the equation of the discriminate function. So, this is a linear equation. Now, based on this discriminate function I can take classification decision.

So, what are the decision I can take. So, if gx that is the discriminate function is greater than 0 then corresponding to this the vector x is assigned to the class omega 1 and suppose gx is less than 0 then the vector x is assigned to the second class. So, we are considering a 2 class problem and suppose the gx is equal to 0 and that is actually the decision boundary. So, this is the equation of the decision boundary. So, that means gx is equal to 0 the vector x can be assigned to either class.

So, and that is the concept of the discriminate function. Now, we have to find the measure of the weight vector. So, that means the problem is we have to find to find the measure of the weight vector W. So, let us move to the next slide. So, if I consider 2 points.

So, I can consider 2 points suppose x1 and x2 on the decision boundary. So, corresponding to these 2 points what will be the equation because corresponding to the decision boundary g1 x is equal to g2 x. So, that means I can write corresponding to the decision boundary

$$\underline{w^T x_1} + w_0 = \underline{w^T x_2} + w_0$$

since x1 and x2 lie on the decision boundary. So, I can write like this.

So, this is the equation of the decision boundary.

So, that means from this equation I can write W transpose x1 minus x2 is equal to 0. So, that is the equation of the decision boundary and here you can see it is the inner product between W and x1 minus x2. So, this vector x1 minus x2 is a vector lying on the decision surface. So, is a vector lying on the decision surface this vector W is orthogonal to any vectors lying on the decision surface since the inner product is 0. So, this is this is actually the inner product if you see this is the inner product the inner product is 0.

So, the meaning is vector W that is the weight vector is orthogonal to any vectors lying on the decision surface since the inner product is 0 for 2 arbitrary vectors here we consider H is actually the hyper plane. So, that is the hyper plane between the classes. Now based on the discriminate function how I can take a classification decision. So, in my Bayesian decision theory already I have explained, but same thing I want to repeat this one here. So, move to the next slide I can show you how you can take a classification decision based on the discriminate function that is the linear discriminate function.

So, suppose we are considering x naught. So, x naught is equal to 1 because we are considering this x naught is equal to 1 and suppose we have x 1 x 2 x d. So, this is x 1 suppose x 2 x d because we are considering the d dimensional Feature vector. So, my Feature vector is x is equal to x 1 x 2 up to x d. So, this is my Feature vector this is the Feature vector x and suppose I am considering this the final classification node is this and corresponding to this I am getting the output is g x.

So, how to determine g x? So, the connection is like this. So, it is connected to this and weight is W naught that is the bias corresponding to this it is connected here and weight is W 1 corresponding to this the weight is W 2 and corresponding to this the weight is W d these are the weights. So, it is a simple linear classifier having d input units and is corresponding to the value of the components of an input vector. So, I am repeating this. So, it is a simple linear classifier having d input units and is corresponding to the values of the component of an input vector and

is input Feature value x i is multiplied by is corresponding with W i.

So, that means the x i is multiplied with W i. So, we have the output unit. So, this is the output unit output unit. So, what is the function of the output unit? The output unit sums all these products and emits a value the value is plus 1 if W transpose x plus W naught is greater than 0. So, what we are considering W transpose x plus W naught actually we are determining with the help of this connection.

So, output units sums all these products and emits a value the plus 1. So, it will give the value 1 plus 1 if W transpose x plus W naught is greater than 0 or it will be minus 1 otherwise. So, you can see that how to take a classification decision based on the discriminate function. The gx will be 1 if W transpose x plus W naught is greater than 0 otherwise it will be minus 1. So, this is how actually we can determine the discriminate function.

Now let us see what is the meaning of the gx. So, what is actually the gx the discriminate function? So, it is actually the algebraic measure of x from H. So, it is a algebraic measure of x, x is the Feature vector from the hyper plane from H, H is the hyper plane. So, it is algebraic measure of x from H. So, actually what is the meaning of this?

So, let us explain this concept.

So, suppose we are considering one hyper plane. So, this is a hyper plane we are considering and we are considering a 3 dimensional Feature space. So, suppose this x 1, x 2 and x 3. So, we are considering the hyper plane, the hyper plane is H and we are considering one arbitrary point here suppose that point is x. So, this is the x is the point we are considering and we are drawing a perpendicular to this plane we are drawing a perpendicular to this plane and put up this perpendicular is suppose x p that is the put up the perpendicular and this length of the perpendicular and

the length of this perpendicular is r suppose.

So, the x the point x is considered and suppose it is the distance from the origin to the point x and from the point x we are drawing a perpendicular to the hyper plane and x p is the put up the perpendicular on H and we are considering the weight vector. So, this is the weight vector that is perpendicular to the plane hyper plane. So, this is my weight vector that is perpendicular to the hyper plane and one distance we are considering the distance from the origin to the hyper plane. So, the distance from the origin to the hyper plane. So, suppose this is the distance of the hyper plane from the origin.

So, that is we can determine W naught divided by W norm. So, that is the distance from the origin to the hyper plane. So, in this figure again I am repeating we are considering a 3 dimensional feature space x 1 x 2 x 3. Now in this case we are considering a hyper plane the hyper plane is H. So, this is the hyper plane is H and we are considering a point x in the right side of the hyper plane and you can see it is a distance from the origin to the x we have shown and from the x we are drawing a perpendicular to the hyper plane.

The length of this perpendicular is R and what is x p? x p is the put up the perpendicular

on H. So, what is the equation of the this decision boundary the equation of the decision boundary is g x is equal to 0. So, already I have mentioned. So, that is the equation of the decision boundary.

So, this is the hyper plane and the equation of the hyper plane is g x is equal to 0.

So, here we are considering this. Now what is x? x I can represent like this x is nothing, but x p plus R W and W norm. So, what is this actually this value actually this is the unit vector perpendicular to H it is the unit vector perpendicular to H. So, that is the meaning of this. So, corresponding to this what is the g x?

g x is nothing, but W transpose x plus W naught.

So, we know this equation. So, now I am just putting the value of x here because x is nothing, but x p plus R and the unit vector. So, it is W transpose W transpose x p plus W naught plus R W transpose W and W norm. So, just I am putting the value of x here. So, if you see this part. So, if you see this part W transpose x p plus W naught this part if I see this is actually 0.

Since the point x p lies on H. So, now this g x we can write like this g x is equal to R W norm I can write like this. So, from the previous equation I can write like this because if you see here this actually this W transpose W is nothing, but W norm square. So, if I put this value then I will be getting g x is equal to R W norm from this the meaning of g x I can write that is the meaning of g x the distance of x from H. So, what is g x distance of the vector x from the hyper plane from H and scale by the weight vector the norm of the weight vector. So, that is the interpretation of the discriminate function.

So, if you see the interpretation of this equation the interpretation is this the distance of x from H and scale by W norm. So, move to the next slide. So, that means in my previous slide I have shown g x is nothing, but the distance R and it is scaled by the norm of the weight vector. So, that is the meaning of the discriminate function.

So, what is R? R is the distance. So, g x divided by W norm. So, that is the meaning of the R and if I consider R is positive suppose if the R is positive that means if the R is positive. So, what is the meaning of this x lies in the positive side of the hyper plane. So, that means the meaning is x lies in the positive side of H and vice versa.

So, that is the meaning of R. So, R we have determined that is R is nothing, but g x divided by W norm. So, this equation it is very similar to the equation of a straight line. So, if you see how actually it is similar this equation of a straight line what is the equation

of a straight line. So, I can write a x plus b y plus c is equal to 0. So, this is the equation of a straight line and

suppose the distance of a point suppose the point is suppose x 1 y 1 to the line.

So, I want to find the distance of the point x 1 y 1 to the line. So, line is represented by the equation a x plus b y plus c is equal to 0. So, what is the distance? So, this distance is nothing,

$$d = \frac{ax_1 + by_1 + C}{\sqrt{a^2 + \sqrt{b^2}}}$$

So, it is a two dimensional case. So, if you see this is very similar to this expression.

So, R is the distance the distance of the point x from the hyper plane and we have considered in the two dimensional case if I consider the two dimensional case the equation of the straight line is given by this and if I consider the distance of the point x 1 y 1 to the line that we can determine like this. So, d is equal to a x 1 plus b y 1 plus c divided by root over a square plus b square. So, this is very similar to the equation of R. So, R is equal to g x divided by W norm.

So, what is the distance of the origin from the hyper plane that already I have shown in the figure.

So, distance of origin from the hyper plane hyper plane is h is nothing, but W norm W naught divided by W norm. So, this W naught is positive when the origin lies on the positive side of the hyper plane and vice versa and if this W naught is equal to 0 that is the bias is equal to 0 that meaning is the hyper plane passes through the origin the hyper plane passes through the origin in this case I will be getting g x is equal to just W transpose x. So, I will be getting the expression for the discriminate function like this if the bias W naught is equal to 0. So, two expressions of the discriminate function in my next slide.

So, one is the linear discriminate function linear discriminate function g x how to write the linear and discriminate function g x.

So, that is nothing, but W naught W naught is the bias and we are considering the d dimensional feature vector. So, W i x i. So, this is the expression for the linear discriminate function and I can give the expression for the quadratic discriminate function. So, what is the expression for the quadratic discriminate function

$$g(\underline{x}) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij} x_i x_j$$

So, you can see W i j is the interconnection between the nodes the nodes already I have shown in my previous figure.

So, this is one equation for quadratic discriminate function. Now, let us see how we can do the classification with the discriminate function. So, I will be discussing two techniques one is one versus all classification and another one is one versus one classification. So, let us discuss about these two techniques of the classification with the help of the discriminate function. So, first I am discussing about one versus all classification this is the first technique and this is called OVA technique one versus all OVA.

So, for OVA we are considering C discriminate function for a C class problem. So, that means we are considering C discriminate function C number of discriminate functions for C classes C number of classes. So, this i th discriminate function generate a decision boundary between the class i and the other classes C minus 1. So, in this figure you can see if I consider the class omega 1 versus all I can consider the decision boundary 1.

So, this is the decision boundary 1. So, in one side it is omega 1 and in another side it is not omega 1 that means other classes. So, in one side it is omega 1 and another side is not omega 1. And similarly if I consider this decision boundary number 2 decision boundary is 2. So, one side is omega 2. So, this is one side is omega 2 and the second side is other side is not omega 2 that is with respect to omega 2 versus all.

And similarly if I consider omega 3. So, corresponding to omega 3 suppose if I consider this is the decision boundary 3 is the decision boundary. So, in one side it is omega 3 and in another side it is not omega 3. And if I consider the decision boundary number 4 you can see. So, in one side it is omega 4 and another side is not omega 4 that means I am comparing one class with all that is why it is 1 versus all. So, if I do this like this then you can see we are having the ambiguous region.

So, this is the ambiguous region that is shown in the pink color. So, what is the procedure? So, the procedure is the I8 discriminant function I8 discriminant function generates a decision boundary between class I and the rest of the classes I8 and the other C minus 1 classes that is the rest of the classes. The test sample is the class having the largest value of the decision function amongst all the C discriminant function. So, based on this procedure I can take a classification decision that means a particular test sample can be assigned to the class having the largest value of the decision function amongst all the C discriminant function. Because we have C number of discriminant functions we have to find the largest one and corresponding to this you can see I have the ambiguous region in the feature space. Now, move to the second technique that is 1 versus 1 classification.

So, move to the next slide. So, the next technique is 1 versus 1 classification technique 1 versus 1 classification technique. So, this is called OVO the previous technique was OVA now it is the OVO. So, for C class problem in this case we have to consider for C class problem we have to consider we have to consider C minus 1 divided by 2 number of discriminant functions. So, we are considering for C class problem we are considering C into C minus 1 divided by 2 number of discriminant function. So, discriminant function Cij it is trained using the samples from the class i and j.

So, this is the discriminant function it is trained by considering the samples from the classes i and j that means containing positive and negative samples. So, it is trained by samples from class i and j that means containing positive and negative samples respectively. So, for i it is the positive sample for j it is the negative samples. So, whenever the decision function value for a test sample is positive from the discriminant function Cij the vote for the class i is incremented by 1. So, whenever the decision function value for a particular test sample is positive from the discriminant function Cij the vote for class i is incremented by 1 otherwise the vote for the class j is increased by 1.

The sample is assigned to the class with a maximum number of votes. So, this is the procedure. So, you can see here this is the 1 versus 1. So, we are considering the class omega 1 in the figure. So, you can see another class is omega 2 and you can see the decision boundary between the classes omega 1 and omega 2. So, number 1 is the decision boundary between the class omega 1 and omega 2.

And number 2 if you see if I consider this is the decision boundary number 2 that is the decision boundary between the classes omega 1 and omega 3. And number 2 and this number 3 that is the decision boundary between the classes omega 3 and omega 4. And number this is number 4 that is the decision boundary between the class omega 2 and omega 4 that is actually we are doing 1 versus 1 classification. And corresponding to this also you can see I have the ambiguous region but the area of the ambiguous region in this case is less than the OVA classification technique.

So, we have shown the decision boundary between the classes. So, number 1 is the decision boundary between omega 1 and omega 2. Number 2 is the decision boundary between omega 1 and omega 3. Number 3 decision boundary is the decision boundary between omega 3 and omega 4. Number 4 decision boundary is the decision boundary between omega 2 and omega 4. So, like this I have all the decision boundaries and just you can see we are just taking the decision based on 1 versus 1 competition.

In both the cases you can see I have the ambiguous region 1 is the 1 versus all and this is the 1 versus 1 classification. So, how to avoid the problem of the ambiguous region? So,

let us move to the next slide. So, how to avoid the problem of ambiguous regions? So, to avoid this problem we are considering C number of linear discriminant functions. For C class problem we are considering C linear discriminant functions. So, we can assign the Feature vector x to the class omega i that assign the Feature vector x to the class omega i if g i x is greater than g j x for all i is not equal to j.

So, that means for C number of classes I have C number of discriminant function and what is the classification rule? The classification rule is the Feature vector x is assigned to the class omega i if g i x is greater than g j x and corresponding to this I can consider a linear classifier. So, what is the linear machine? The linear machine I can consider like this x is my input Feature vector and corresponding to this I am just determining the discriminant function g 1, g 2 all these discriminant function I can determine g c. So, C number of discriminant function and out of this we have to select the maximum one which one is maximum. So, this is a g 1 x we have g 1 x this is g 2 x this is g c x and which one is the maximum discriminant function we have to determine and based on this we can take a classification decision.

This is a linear machine. So, for C number of classes I have C number of discriminant function and with the help of this I can take a classification decision. So, corresponding to this my decision boundaries may be something like this. So, corresponding to this my decision boundaries may be something like this. So, I have the classes. So, this is the class omega 2, this is the class omega 1, this is the class omega 3.

So, this region is R 2, this region is R 1 and this region is R 3. So, you can see the decision boundary produced by the linear machine for a 3 class problem. This is a decision boundary I can have corresponding to the linear machine. So, I can consider this type of decision boundary suppose this is my class omega 3 and the region is R 3. So, this may be this may be omega 1 corresponding region is R 1, this is omega 5 and corresponding region is 5, this is region 2 R 2 and suppose the class is omega 2, this is the class omega 4 and corresponding region is R 4. So, you can see in the first case we have considered the decision boundary for a 3 class problem.

So, in the first case I have shown the decision boundary for a 3 class problem. So, this is a 3 class problem 3 classes. The second one is a decision boundary for a 5 class problem 5 class classes. So, you can see the decision boundary between the classes. So, one is omega 1, one is omega 2, one is omega 3, omega 4, omega 5.

So, you can see these are the decision boundaries. So, with the help of the linear and discriminant function I can do the classification. In this class I have explained the concept of the discriminant function and also I have explained the physical and the geometrical

interpretation of the discriminant function. After this I discussed the concept of the one versus all and one versus one classification techniques. The problem is the presence of the ambiguous region. So, to consider this issue I am considering c number of discriminant functions for c number of classes and I have to pick the largest or the maximum discriminant function and based on this I can take a classification decision.

In my next class I will be explaining the concept of determining the discriminant function with the help of the training. So, that concept I will be explaining in my next class. So, let me stop here today. Thank you.