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-17

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and applications. In my last class, I discussed the concept of discriminate function and its geometrical interpretation. gx is the discriminate function and we considered linear discriminate function. It is nothing but the distance from the point x to the hyper plane and it is scaled by the norm of the weight vector. So, that is the interpretation of the discriminate function gx. So, I am repeating this it is a distance from the point x to the hyper plane and it is scaled by the norm of the weight vector and that is the geometrical interpretation of the gx the linear discriminate function.

After this I discussed two classification techniques one is one versus all and another one is one versus one. Today I am going to discuss how to determine the best discriminate function. So, that there should not be any misclassification that means all the samples belonging to different classes should be correctly classified. So, whenever I design the discriminate function gx I have to consider the weight vector w is the weight vector.

Now how to design the weight vector which one is the best weight vector I have to determine from the samples of the different classes. So, if I consider two classes suppose omega 1 and omega 2. So, corresponding to omega 1 I have samples and corresponding to omega 2 I have also the training samples. So, with the help of these training samples I have to determine the best weight vector and if I can determine the best weight vector that is actually the best discriminate function. So, that the misclassification will be 0 that means all the samples will be correctly classified.

So, my objective is to determine the best weight vector from the training samples of the classes and that is called the perceptron criterion. So, today I am going to discuss about this concept. So, how to determine the best weight vector from the training samples of the classes. So, beginning I will be considering only two classes and that principle can be extended for multiple classes. So, let us begin this class that perceptron criterion.

So, in my last class I discussed the concept of the discriminate function and we considered linearly separable classes. So, if I consider the linear discriminate function the expression is

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

So, W is the weight vector and W naught is the bias or the threshold weight and x is the d dimensional Feature vector. And based on this discriminate function I can take classification decisions. So, if gx is greater than 0 then I have to assign the Feature vector to the class omega 1 and gx less than 0 that means the x should be assigned to the class omega 2 and gx is equal to 0 actually this is the equation of the decision boundary and corresponding to this I cannot take any decision.

So, we have to find the nature of the weight vector and in the last class also I have shown what is the physical interpretation of this discriminate function. So, we have derived like this gx is equal to R W norm. So, this is the interpretation that means it is the distance of the point x from the hyper plane and it is scaled by the norm of the weight vector that is the interpretation of the discriminate function. So, that is gx is the algebraic measure of x from the decision surface H and also we have determined distance of the hyper plane from the origin. So, the distance of the origin from the hyper plane is H.

So, that expression also we have determined and that is nothing but W naught divided by W norm this W naught that is the bias it is positive when the origin lies on the positive side of the hyper plane and vice versa. And if W naught is equal to 0 suppose if I consider if W naught that is a bias is equal to 0 then what is the meaning of this the hyper plane passes through the origin and corresponding to this the discriminate function will be gx will be W transpose x because the bias is 0 and this form is called the homogeneous form. From the computational point of view we have to consider the homogeneous form. Now, what we have to consider now how to design the weight vector which one is the best weight vector for the best discriminate function because the discriminate function depends on the weight vector. So, I have to design the best weight vector.

So, this weight vector I can design based on the training samples of the classes. So, I have the training samples for different different classes and based on these training samples I have to design the weight vector the weight vector is W. So, let us move to the next slide. So, design of weight vector W. So, how to design the weight vector?

So, for this let us consider the two class problem that is a two category problem

and we are considering linearly separable case.

So, this concept can be extended for multiple classes. So, you know that $g(\underline{X}) = \underline{w}^T \underline{x} + w_0$

. So, we have to convert it to homogeneous form to convert into homogeneous form. So, gx can be written like this A transpose y.

So, this expression I am writing in this form and that is the homogeneous form.

So, in this case y is the augmented Feature vector is y is the augmented Feature vector that means this y is nothing but we have the d dimensional Feature vector. So, x1 x2 up to xd and 1 is appended here. So, this is the augmented Feature vector and the vector is the modified weight vector. So, this vector A this is the modified weight vector. So, in this weight vector all the components of W and also the bias W naught is considered.

So, now how to write gx is equal to A transpose y and we have the modified weight vector. So, in the modified weight vector we are considering all the components of W. So, this is W1 W2 up to Wd and also the bias weight we are considering this is the modified weight vector and after this we are considering augmented Feature vector. So, augmented Feature vector is x1 x2 xd 1. So, this we have considered that is equal to the summation from i is equal to 1 to d Wi xi plus W naught.

So, that is actually equal to W transpose x plus W naught. So, here we are converting the previous equation gx is equal to W transpose x plus W naught into homogeneous from and for this what we are considering the augmented Feature vector we are considering and we are considering the modified weight vector. Now, let us consider about the decision rules. So, move to the next slide. So, what is the decision rules? So, the decision rules will be the A transpose yi is greater than 0 then yi is the modified Feature vector and what is the corresponding original Feature vector the corresponding original Feature vector is xi.

So, xi belongs to the class omega 1. So, I am repeating this if the condition the A transpose yi is greater than 0 then this yi is nothing but the modified Feature vector and what is xi that is the original corresponding Feature vector. So, xi belongs to omega 1 and if it is less than 0 this A transpose yi is less than 0 that means what I have to consider this is the modified Feature vector modified Feature vector is yi xi is the original corresponding Feature vector xi will be assigned to the class omega 2. So, now how to design the weight vector A by considering or by using the knowledge of the samples corresponding to omega 1 and omega 2.

So, I have the samples for class omega 1 I have also the samples for the class omega 2.

So, with the help of the knowledge of the samples corresponding to omega 1 and omega

2 I have to design the weight vector. So, that is the problem. So, we consider augmented samples. So, we are considering suppose n number of samples and these are actually the training samples. So, with the help of these training samples I have to design the weight vector and corresponding to this suppose I have the augmented Feature vector.

So, these are the augmented Feature vectors. Some of the Feature vectors are labeled as omega 1 and some of the Feature vectors are labeled as omega 2. So, that means what is the decision rule now the decision rule is if A transpose yi is greater than 0 that means yi will be assigned to the class omega 1 and if it is less than 0 yi will be assigned to the class omega 2. So, these are two conditions and we have to determine the weight vector A to satisfy these two conditions. Now in this case we are considering two decision criteria.

So, instead of considering two decision criteria whether it is possible to determine only single criterion that is only the single condition we will be considering and if a condition is satisfied then the samples are classified correctly and if it is not satisfied samples are not correctly classified. So, that means instead of considering these two conditions I want to determine only one condition and based on this I have to do the classification. So, if a particular condition is satisfied the samples will be classified correctly and if the condition is not satisfied the samples are not correctly classified. So, that means what we can consider A transpose yi here in this case we are considering two conditions for classification I want to make only one condition. So, A transpose yi is greater than 0 that means we will be considering this condition A transpose yi greater than 0 and

we do not see the label of yi.

So, only we will be considering this condition and that is the single condition and this condition the A transpose yi is equal to 0 that means we cannot decide. So, corresponding to this condition we cannot decide that means the samples may belong to the class omega 1 or they may belong to class omega 2 that we cannot decide. So, we are considering this single condition with the help of the single condition let us determine the classification conditions. So, how to do the classification with this single condition.

So, move to the next slide. So, what I am doing for this single condition samples belonging to the class omega 1 is augmented and take as it is. So, this is for the samples belonging to the class omega 1 and the samples belonging to the class omega 2 is augmented and negated. So, this is for the samples belonging to the class omega 2. So, they are augmented and negated also. So, my samples will be correctly classified irrespective of samples belonging to the class omega 1 or omega 2.

So, now I have to design a criterion function and with the help of this criterion function I have to decide or I have to determine the best weight vector. So, you can see how we are

considering now that for a single criterion what we are considering the samples belonging to the class omega 1 is augmented and take as it is and samples belonging to the class omega 2 is augmented and negated. So, now we are considering one criterion function to design the classifier. So, the criterion function we are considering. So, what is this criterion function? J A is the criterion function we are considering.

So, it has to be minimized for which A is the correct weight. So, I can write here has to be minimized for which A is the best weight vector. That means the weight vector which can classify all the samples correctly and A is the best weight vector which minimize the criterion function. So, I am repeating this that means we are determining the best weight vector. The weight vector which can classify all the samples correctly is the best weight vector which minimize the criterion function J A.

So, to determine the best weight vector I have to consider this criterion function because I have to consider the minimization of the criterion function. So, for this we are considering one algorithm and that algorithm is the gradient descent approach. So, with this approach I want to find the best weight vector and for this I have to minimize this criterion function. So, I have to minimize this criterion function for which A is the best weight vector. That means I want to classify all the samples correctly and corresponding to this A is the best weight vector which can classify all the samples correctly and for this we have to minimize this criterion function.

So, what is the gradient descent approach. So, suppose A k we are considering the value of A in the kth iteration. So, this is the value of A in the kth iteration. This is an iterative algorithm and after the first iteration A is equal to k plus 1 my updation rule is like this A k that is the value of A in the kth iterations A k minus eta and we are taking the gradient of the criterion function. So, this is the equation corresponding to this gradient descent algorithm.

$$\underline{a}(K+1) = \underline{a}(K) - \eta \nabla J(\underline{a}(K))$$

So, in this case this eta is the learning rate. So, let us see what is the physical interpretation of this equation. So, move to the next slide what is the physical interpretation of this and let us explain this condition. So, what we have obtained at k plus 1 iteration A k plus 1 is equal to A k minus eta is the learning rate and we are taking the gradient of A k this is the equation of the gradient descent algorithm.

Suppose I am plotting J A that is the criterion function with respect to A.

So, this criterion function is suppose it is like this it is a minima. So, it is like this and suppose corresponding to the iteration A k. So, this is my value corresponding to A k in the kth iteration that is the value of the criterion function. Now I have to determine the

gradient of this at this point. So, if I determine the gradient of this you can see I will be getting the gradient along this direction.

So, gradient will increase in that direction, but because of this minus sign in this equation if you see the minus sign in this equation here I have to move in the opposite direction. The opposite direction is like this I have to move in the opposite direction because of the negative sign. So, the gradient will increase in the upward direction because I am taking the gradient, but I have to move in the opposite direction. And after this suppose if I consider F k plus 1. So, suppose we are considering this is the value F A k plus 1 iteration that means in the next iteration if I consider eta is very high very large that means the overshooting of the value that means if I consider A k plus 1 is like this.

So, this gradient so it may go to this point if the eta is very large the eta is very large. And after this it may move to this point and again it may move to this point that means overshooting the value and that means the oscillation is taking place if eta is very high. So, if I consider eta is small. So, if I consider eta is very small.

So, that means it will go to this point again it will go to this point it will go to this point and finally it will get the minimum point.

So, minimum point I will be getting. So, if the eta is very very small then the convergence will take time. So, you can see we are moving in the opposite direction of the gradient and if I select the large value of eta then you can see the overshooting is taking place and that is nothing but the oscillation. And if I consider eta is very very small then the convergence will take time. So, this is the convergence for eta which is small and this is the convergence for eta is very high. So, you can see I am showing the conditions for the convergence one is eta is very very small another one is eta is very high and you can see this is the minimum point.

So, we have to find this point the minimum point. So, what we can consider this is actually the steepest descent algorithm or the steepest descent procedure that means we have to follow negative of gradient and the initial arbitrary weight vector we can select. So, initially what we can consider any arbitrary weight vector we can consider for this gradient descent algorithm that is the steepest descent algorithm. So, it is also called the steepest descent procedure that means we have to follow the negative of the gradient. Now let us consider what type of perceptron criterion we can consider for classification of the samples. Now let us consider how to define some criterion function.

So, with the help of this criterion function we can determine the best weight vector. So, one such criterion function is the perceptron criterion. So, that I will be explaining in my next slide. So, move to the next slide.

So, we are considering one criterion function and that is the perceptron criterion.

This perceptron criterion it is actually based on the samples which are not correctly classified. I have to design the classifier which is based on the samples which are not correctly classified. So, this is the perceptron criterion because the samples which are correctly classified we do not have any problem, but the samples which are not correctly classified we have to take care. So, that is why this perceptron criterion it depends on the samples which are not correctly classified. So, what is this perceptron criterion? So, let us consider k as instant and corresponding to this suppose the weight vector is ak.

So, actually k represents the iteration number. So, a is the weight vector and k is the iteration number. So, this ak should correctly classify all the samples. So, ak is the weight vector and ak should correctly classify all the samples so, that means, a transpose k already I told you k is the iteration number that means, if a transpose k y is greater than 0 that means, correctly classified a transpose k y less than 0 or I can say another condition a transpose k y is equal to 0. So, for these two conditions, the samples are not correctly classified.

In fact, the third condition a transpose k y is equal to 0, we cannot take a classification decision.

So, samples may belong to any one of the classes. So, we cannot take any classification decision corresponding to the condition a transpose k y is equal to 0. So, what is my objective? Objective is to find a transpose k y should be greater than 0 for all the samples. That means all the samples belonging to all the classes should be correctly classified. So, we have to find the weight vector ak so that all the samples are correctly classified and we will consider all the samples which are not correctly classified by the weight vector ak because this perceptron criterion it is based on the samples which are not correctly classified. So, I have to give the importance to that samples, the samples which are correctly classified I do not have any problem.

So, the samples which are not correctly classified we are considering that one. So, that means we will consider all the samples which are not correctly classified by the weight vector ak. So, this perceptron criterion function j p a. So, j p means the criterion function and p means the perceptron criterion I can define like this summation minus a transpose y and for all y which are misclassified. So, what we are considering if a particular sample is misclassified so that means I have to modify the weight vector a and if a sample is misclassified a transpose y is negative and minus a transpose y will be always positive or it may be also 0.

So, I am repeating this if a sample is misclassified a transpose y is negative and minus a

transpose y will be always positive or maybe 0 this is the condition. Now I have to take the gradient of this. So, gradient with respect to a j p a.

So, it is equal to summation minus y and for all y which are misclassified.

So, we are taking the gradient with respect to a. So, what is the updation rule for the weight vector because we are designing the best weight vector. So, what is the weight updation rule for the weight vector. So, move to the next slide with updation rule. So, the initial weight vector a 0.

So, we are selecting arbitrarily and at the k plus 1 iteration a k plus 1.

So, a k that is the weight updation rule it at the learning rate and if you see the previously we considered minus. So, minus minus it will be plus now. So, summation y because this minus minus will be plus in the gradient descent algorithm it was minus. So, in the perceptron criterion it is minus summation for all minus that is the misclassified samples. So, this minus minus it will be plus and for all y which are misclassified.

So, this is the weight updation rule. So, we have to consider this rule the weight updation rule. So, if I apply this algorithm we can finally find the weight vector which can classify all the samples correctly. So, we can find the best weight vector which can classify all the samples correctly. So, here this summation for all y which are misclassified this is actually the sum of samples misclassified in the previous iteration.

So, this is the weight updation rule. So, what is the geometrical interpretation of this algorithm? So, what is the geometrical interpretation of this algorithm that is the weight updation rule and the perceptron criterion. So, we are classifying the samples belonging to different classes and we are finding the best weight vectors. So, the illustration I can show in the next slide corresponding to this discussion. So, to the next slide suppose I have a 2 dimensional spatial space. So, these are 2 dimensional feature space x1 and x2 and I have the samples belonging to the class omega 1

and I have some samples belonging to the class omega 2.

So, suppose these are my samples belonging to class omega 1 these are the samples belonging to the class omega 1 and I have some samples belonging to the class omega 2. So, suppose these are my samples belonging to the class omega 2. So, as per this perceptron criterion what we have considered the samples belonging to the class omega 1 is augmented and take as it is and samples belonging to the class omega 2 is augmented and negated. So, these samples belonging to the class omega 2 they are augmented and negated.

So, that means I have to do the negation. So, if I do the negation you can see it will be

something like this. So, samples are negated. So, negated samples corresponding to the class omega 2. So, now this will be the class omega 2 after the negation.

Now I have to find the best decision boundary between these classes. Suppose if I consider this decision boundary. So, corresponding to this decision boundary you can see there is no misclassification. So, all the samples belonging to the class omega 1 are correctly classified and all the samples belonging to the class omega 2 they are also classified. So, suppose if I move this decision boundary to the limiting condition.

So, limiting case will be like this. This is the limiting case. So, beyond this I cannot move because I am moving in this direction in the counterclockwise direction beyond this if I move then there will be misclassification of the nearby samples. So, I cannot move the decision boundary beyond this limiting point or I can say that limiting position beyond this limiting position I cannot move the decision boundary. And corresponding to this what is the weight vector? The weight vector is always perpendicular to the decision boundary. So, this is my weight vector corresponding to this decision boundary.

This is the perpendicular to the decision boundary. Similarly, if I move this boundary in the clockwise direction what is the limiting case? So, which one is the limiting case? I want to show here. The limiting case. So, I can move up to this point and beyond this I cannot move because if I move then what will happen? The misclassification will take place. And corresponding to this what is my weight vector? This is my weight vector that is perpendicular to the decision boundary.

This is perpendicular to this decision boundary. So, my solution region is this. So, within this region I have to find my weight vector. So, this is the weight vector W this is the weight vector and between this I have to find my weight vector and this is the solution region. So, this is the interpretation of the discussion what I have discussed about the steepest descent algorithm and also the concept of the perceptron criterion. So, this is the geometrical interpretation of this. So, I have to find the best weight vector and you can see I have shown the solution region.

So, within this region my weight vector will be available. So, that is the interpretation of the previous discussion about the perceptron criterion and also the steepest descent algorithm. In this class I discussed the concept of determining the best weight vector for a classifier. So, how to determine the best weight vector that means I have to classify all the samples belonging to different classes correctly. The samples belonging to the class omega 1 and the samples belonging to the class omega 2 should be correctly classified.

And for this I am determining the best weight vector. So, I am applying the techniques

like the steepest descent algorithm or the gradient descent algorithm to determine the best weight vector. In the steepest descent algorithm I have to move in the opposite direction of the gradient. After this I discussed the concept of perceptron criterion and with the help of this you can see how we can determine the best weight vector for a particular classifier. So, if I can determine the best weight vector and that corresponding discriminate function I can determine.

So, in my next class also I will be continuing the same concept. So, let me stop here today. Thank you.