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-19

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and applications. I have been discussing the concept of generative classifiers. In case of the generative models, it is assumed that the samples of training data of a class come from a probability density function that is the class conditional density. So, if I have the information of class conditional density, I can easily determine the discriminate function. And with the help of this discriminate function, I can determine the decision boundary between the classes. So, this is about the generative classifiers.

In case of the discriminative classifiers, I do not need the information of class conditional density. So, arbitrarily I can fix a decision boundary between the classes and this decision boundary is characterized by the weight vector. After this, I have to update the weight vectors by some optimization techniques to get the best decision boundary between the classes. So, this is the fundamental concept of the discriminative classifiers.

So, some of the examples like MLP, multi layer perceptrons, support vector machine. So, these are some examples of discriminative classifiers. So, today I am going to discuss about the concept of the discriminative classifiers and also the concept of support vector machine. So, let us start this class. So, the classifier taxonomy.

So, if you remember in my first classes, I discuss these two types of classifiers. One is the generative classifiers and another one is the discriminative classifiers. So, already I told you in case of the generative classifier, the samples of training data of a class assumed to come from a probability density function that is the class conditional PDF. So, in my discussion I showed for normal density or for the Gaussian density, I can determine the discriminative function and after this I can determine the decision boundary between the classes. So, the types of generative classifiers already I have discussed about this.

One is the parametric approach. So, parametric another one is the non-parametric. So, in case of the parametric classifier, I know the class conditional density that is the density from is known, but I do not know about the values of the parameters. So, for example, I considered the Gaussian density.

So, I have that information,

but I do not know the values of the parameters for the Gaussian density.

So, that is the mean vector and the covariance matrix. So, that I can determine from the algorithms like the maximum likelihood estimation or the Bayesian estimation. In case of the non-parametric, I can directly estimate the density and I have explained two important concepts. One is the Parzen window technique and another one is the k nearest neighbor technique. So, this in case of the generative classifiers, you can see I can determine the discriminative function and I can find the decision boundary between the classes.

Suppose these are my samples belonging to a class omega 1 and I have some samples belonging to another class omega 2 and I can determine the decision boundary between the classes. So, this is the decision boundary between the classes. So, it is a two dimensional feature space. So, I will be getting a line between these two classes. So, these are two dimensional feature space x 1 and x 2 these two features.

So, I will be getting a line. If I consider a three dimensional space, I will be getting a plane and if I consider high dimensional case, then I will be getting a hyper plane. So, this is the decision boundary between the classes in case of the generative classifiers. So, in case of the discriminative classifiers, we do not need the information of the class conditional density. So, what we can consider, we can start with initial weight that define the decision boundary.

So, I can write here. So, what I need to do start with initial weights that define the decision surface. So, after this I have to update the weights based on the norm optimization criterion. So, I have to apply some iterative algorithms and finally, I will be getting the best decision boundary between the classes. So, in this case, I can give some examples like MLP.

MLP is the multi layer perceptron or maybe we can consider the single layer perceptron. Another example is the support vector machine. So, these are some examples of the discriminative classifiers. So, the concept is start with initial weights that define the decision boundary between the classes. After this I have to update the weights based on some optimization criterion and I have to apply some iterative algorithms.

So, finally, I will be getting the best decision boundary between the classes. So, pictorially this can be shown what is the concept of the discriminative classifier. So, let us move to the next slide. So, first I have discriminative classifier. Suppose I have some samples and suppose the level is suppose 1 and similarly I am considering another class.

So, suppose the level is 0. So now suppose I am fixing a decision boundary something like this. So, if I consider this decision boundary, you can see here, this is the position 1 and this is the position in the decision boundary. So, initially we are considering this is the decision boundary So, here you can see this sample that is misclassified and also this sample and this sample they will be misclassified. So, after this what I have to do I have to update the weights and I have to find the next position of the decision boundary.

So, suppose after some iteration, I am getting the decision boundary between the classes. So, suppose it is in the position number 2. So, if I consider this decision boundary that is the number 2, you can see this decision boundary can perfectly classify the samples belonging to two classes. So, like this I have to apply some optimization technique to get the best decision boundary between the classes and I can show some of the examples of linearly separable data. So, suppose if I consider some samples like this, you can see this is a linearly separable case and similarly if I consider these are the samples for a particular class and if I draw the decision boundary.

So, this is one example of the linearly separable data and what about the non-linearly separable data. So, pictorially I can show you suppose I have some samples like this that is for the class omega 1 and suppose I have some samples belonging to another class. So, for this it is not possible to draw a single decision boundary. So, maybe we can consider the decision boundary like this. So, one decision boundary is not possible in this case.

So, I can consider like this the two decision boundaries between the samples of two classes. So, this is one example of non-linearly separable data. So, move to the next slide. So, suppose I have a two dimensional feature space. So, features are x 1 and x 2 and I have some samples belonging to a class omega 1 and omega 2.

So, these are the samples belonging to the class omega 1 and I have some samples belonging to another class omega 2. So, these are the samples for the class omega 2 and I can draw a decision boundary between the classes. So, this is the decision boundary between the classes. So, in this case I will be getting a separating line because we are considering a two dimensional feature space. So, I will be getting a separating line.

So, what is the equation of this line? The equation of this line is

$w_1 x_1 + w_2 x_2 + b = 0.$

This is the equation of the separating line between the classes and corresponding to the class omega 1. So, what is my decision rule? It is w 1 x 1 plus w 2 x 2 plus the bias is b is greater than 0. Then I have to consider the class omega 1 and corresponding to the class omega 2 the condition is w 1 x 1 plus w 2 x 2 plus b less than 0. So, you can see based on these conditions I can take a classification decision.

So, you can see we are showing the separating line between the classes. So, this is separating line in two dimensional feature space. So, now let us discuss about the concept of the discriminate function. So, already I discussed about the concept of the discriminate function and based on the discriminate function I can take a classification decision. So, let us move to the next slide.

So, suppose I have a Feature vector. Feature vector is x. x can be assigned to the class omega 1 that is x can be assigned to the class I can say suppose omega i if the condition is g i x that is the discriminate function for the class omega i is greater than g j x for all i which are not equal to j. So, if I consider for two category case and this I can write like this g x is nothing, but the difference between g 1 x and g 2 x g 1 x minus g 2 x. So, I have to decide the class omega 1 if g x is greater than 0 otherwise I have to decide the class omega 2 and based on this discriminate function we have designed the minimum error rate classifier.

So, based on this discriminate function I have discussed the concept of the minimum error rate classifier. So, for minimum error rate classifier we determine g x like this probability of omega 1 given x minus probability of omega 2 given x. So, this we have determined if you remember in my first some classes I discussed the concept of the discriminate function and also I discussed the concept of the minimum error rate classifier. So, this is a fundamental concept of the discriminate function. So, based on this discriminate function you can see I am showing some of the decision boundaries for different different classifiers the first one is the nearest neighbor classifier.

So, you can see a first case you have seen the decision boundary between the classes and it is a non-linear decision boundary between the classes that is corresponding to the nearest neighbor classifier. The concept of decision tree I will be explaining later on, but in case of the decision tree you can see how I am getting the linear decision boundary between the classes. And in case of the linear discriminate function you can see I am getting the linear decision boundary between the classes. So, g x is equal to W transpose x plus b. So, W is the weight vector, x is the input feature vector and b is the bias.

And if I consider non-linear function you can see I am getting the non-linear decision boundary between the classes. So, here I am showing the decision boundary for different different cases one is for the nearest neighbor one is for the decision tree one is for the linear discriminate function and I am considering also the non-linear functions. So, now let us consider g x is a linear function and expression for the linear discriminate function is W transpose x plus b. So, you can see I am showing one decision boundary between the classes. So, here you can see this is the decision boundary.

So, it is a two dimensional feature space. So, I will be getting a line. So, if I consider the high dimensional feature space. So, I will be getting a hyper plane in the feature space.

So, here you can see I am considering one unit vector that is n

that is a normal vector of the hyper plane.

So, this unit length unit length normal vector of the hyper plane. So, we are just considering this decision boundary as a hyper plane of the hyper plane. So, the unit linked normal vector n is equal to the weight vector W divided by W norm. So, you can find a normal vector of the hyper plane and you can see the equation of the hyper plane is W transpose x plus b is equal to 0 and I can decide two classes based on these two inequality conditions. So, corresponding to the first class the condition is

$W^T x + b > 0$ and corresponding to the second class the condition is $W^T x + b < 0$.

So, my case is how to find the base decision boundary between the classes because for perfect classification I have to determine the best decision boundary between the classes because we have to minimize the error rate. Move to the next slide you can see I am considering number of decision boundaries number 1 number 2 number 3 number 4. So, like this I can consider infinite number of decision boundaries and out of this which one is the base decision boundary. So, that my error rate is minimum. So, in this example I have shown the number of decision boundaries between these two classes 1 2 3 4 like this I may have infinite number of decision boundaries and

out of which one is the best decision boundary that I have to determine.

So, that means the question is how would you classify the sample points using a linear discriminate function in order to minimize the error rate and corresponding to this question the answer is infinite number of decision boundaries infinite number of answers. So, you can see I have infinite number of answers. So, which one is the best decision boundary in order to minimize the error rate that I have to define. So, move to the next slide. So, in this figure also I have shown the same classification problem two classes we are considering and we have shown the decision boundary.

So, this is the decision boundary. Now, I am defining one term the term is the margin. So, the margin is defined as the width that the boundary could be increased before hitting the data point. So, the concept is like this. So, you can see this is the decision boundary I am increasing the width of the decision boundary and like this I am increasing and I am stopping just before hitting the data point.

So, here you can see this point these are very critical points.

So, just I am expanding or I am increasing the width of the decision boundary and I am stopping just before hitting these data points. So, based on this I can define the margin. Could be increased by before hitting a data point. So, you can see I am expanding the decision boundary and just this decision boundary is touching the data point that is shown by this arrow. So, here in this case I have three sample points near the boundary and I am expanding my decision boundary up to that points.

So, which one is the best decision boundary? So, the linear discriminate function with the maximum margin is the best. So, the margin should be so the linear discriminate function with maximum margin is the best so now I have to show some mathematical techniques how to find the best margin or how to find the best decision boundary between the classes. So, for this we are considering given a set of data points. So, suppose I have some data points x i that is a data point corresponding to the class y i.

So, all these are level points 1, 2 up to n. So, what is the output? Output is for y i is equal to plus 1 that means I am considering the level for the class 1 is plus 1 the condition should be W transpose x i plus b should be greater than 0 and the second condition is corresponding to the second class it is y i should be equal to minus 1 and the condition is W transpose x i plus b less than 0. So, these two conditions I am obtaining. So, I have the samples the samples are x i is i is equal to 1, 2 up to n and I have considered two classes. So, for the first class y i is equal to plus 1 and for the second class y i is equal to minus 1. So, for y i is equal to plus 1 the condition is W transpose x i plus b greater than 0 and for the second class the output is y i is equal to minus 1.

So, the condition is W transpose x i plus b less than 0. So, with a scale transformation on both W and b these two equations I can write like this for y i is equal to plus 1 the condition is W transpose x i plus b. So, earlier it was 0 now I am putting is 1. So, greater than equal to 1 and y i is equal to minus 1 the condition is W transpose x i plus b less than minus 1.

So, this condition I am getting. So, it is obtained from this one I am doing a scale transformation on both W and b. So, that means the second set of equations they are equivalent to the first set of equations. So, now I have to define the margin because the width of the margin is quite important. So, move to the next slide. So, here we are considering these data points if you see these data points these are called the support vectors the arrow points these are the support vectors I am increasing my decision boundary and I am stopping just before hitting these data points.

So, these data points are the support vectors. So, beyond this I cannot increase my decision boundary. So, my decision boundary is just passing the support vectors. So, here support vectors I have shown like x plus and x minus. So, in one side it is x plus another side is x minus.

So, one is x plus another one is x minus. So, corresponding to the support vectors how to write this equation. So, corresponding to the support vectors W transpose x plus b is equal to 1 and similarly W transpose x minus plus b is equal to minus 1. So, I can write these two equations corresponding to the support vector x plus and x minus.

So, I think you understand the concept of the support vectors the concept I am explaining it again I am increasing the decision boundary.

So, that it just passes the support vectors. So, beyond that I cannot increase my decision boundary. So, based on these two equations I can define the margin. The margin with m the difference between x plus and x minus and we have to consider the direction that is given by the unit vector n. So, unit vector n already I have shown in my figure also in this figure also you can see the unit vector n. So, this n now can be determined like this x plus minus x minus this x plus and x minus they are support vectors this unit vector is

$\frac{W}{|W|} = \frac{2}{|W|}$.

So, that is the width of the margin how to get this margin width I can give one example. Suppose if I consider two parallel lines. So, equation is suppose ax plus by plus c1 is equal to 0 and another equation is ax plus by plus c2 is equal to 0. So, two parallel lines we are considering what is the distance between two parallel lines the distance between these two parallel lines from the coordinate geometry c2 minus c1 divided by root over a square plus b square. So, similarly here if I consider the equations of this hyperplane

$W1\, x1 + W2\, x2 + b - 1$.

So, we obtain this equation I have shown this equation in my previous slide. So, W1 x1 plus W2 x2 plus b plus 1 is equal to 0. So, corresponding to two lines the distance again I can determine distance will be b minus 1 minus b minus 1 and this is W1 square plus W2 square. So, I will be getting 2 by W norm. So, here you can see this is very similar to this and actually we have obtained this one from this.

So, you can see the width of the margin is 2 divided by W norm and in this figure already I have shown the support vectors. So, that means the points which lie on the canonical hyperplane are called the support vectors. So, what is the support vector the points which lie on the canonical hyperplane are called support vectors. Now the condition of this linear classifier I have to maximize the width of the margin because a linear discriminate function or a linear classifier with the maximum margin is the best. So, that is why I have to maximize the quantity 2 divided by W norm that is the margin width.

So, what is the formulation for this? So, move to the next slide the formulation is so, I have to maximize 2 divided by W norm that is the margin width such that yi should be equal to 1 that is the output corresponding to the class omega 1. The condition is W transpose xi plus b greater than equal to 1 and for a second class yi is equal to minus 1 and the condition is W transpose xi plus b less than equal to minus 1. So, whenever W transpose xi plus b is equal to 1 that is actually the equation of the decision boundary. So, I have to maximize 2 divided by W norm this condition is equivalent to minimize 1 by 2 W norm whole square. So, instead of this previous condition maximize 2 divided by W norm we can consider minimize 1 by 2 W norm square.

So, the condition will be same. So, I can say such that yi is equal to plus 1 and same condition we are considering W transpose xi plus b greater than equal to 1 and yi is equal to minus 1 corresponding to W transpose xi plus b less than equal to minus 1. So, from maximization condition to minimization condition I am obtaining like this. So, instead of maximizing I am considering minimizing this one. So, the final formulation is minimize 1 by 2 W norm square such that the condition is y W transpose xi plus b greater than equal to 1.

So, this condition we are considering. So, for solution of this I have to consider the optimization techniques. So, maybe we can consider the Lagrangian function. So, let us move to the next slide. So, solving the optimization problem the optimization problem is minimize 1 by 2 W norm square subject to the condition yi. So, corresponding to this we are considering this Lagrangian function technique minimize the Lagrangian function minimize Lp is a it is a function of W, W is the weight vector b is the bias and alpha i is the Lagrangian multiplier.

$$L_P(\underline{w}, b, \alpha_i) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i(y_i(\underline{w}^T\underline{x}_i + b) - 1)$$

So, it is 1 by 2 W norm square that is summation from i is equal to n we considered n number of samples alpha i is the Lagrangian multiplier. So, you have to see the mathematics book what is the technique for solution of the optimization problem because we are considering linear constraints. So, we can consider the Lagrangian function for solution of this problem W transpose xi plus b minus 1 subject to the condition alpha i greater than equal to 0. So, here alpha i, i is equal to 1, 2 up to n this is Lagrangian multiplier function. So, we are solving this optimization problem by considering this Lagrangian function because we have the linear constraints.

So, we have to minimize Lp that is the Lagrangian function. So, move to the next slide. So, again I am writing minimize this Lagrangian function Lp W b and alpha i 1 by 2 W norm square minus summation from i is equal to 1 to n alpha i yi W transpose xi plus b minus 1 subject to the condition this Lagrangian multiplier should be greater than equal to 0. So, for this what we are considering we are doing the differentiation of Lp with respect to the weight vector W and equating it to 0. So, corresponding to this if I do the differentiation I will be getting the weight vector W is equal to summation from i is equal to 1 to n alpha i yi xi.

And similarly if I differentiate this Lp with respect to the bias b equating it to 0.

So, I will be getting summation i is equal to 1 to n alpha i yi is equal to 0. So, that means I am getting the solution for the weight vector W. So, this is the solution for the weight vector W. So, now I am considering this Lagrangian dual problem better to move to the next slide. So, what we have considered minimize I am writing it again the Lagrangian function is Lp W b equal to 1 by 2 W norm square minus summation over i is equal to 1 to n alpha i yi W transpose xi plus b minus 1. This condition we have considered subject to the condition alpha i greater than equal to 0.

So, the Lagrangian dual problem I can obtain like this. So, maximize this Lagrangian dual problem maximize the minimization problem is converted into maximization problem maximize alpha i i is equal to 1 to n minus 1 by 2 summation i is equal to 1 to n summation j is equal to 1 to n alpha i alpha j yi yj xi transpose xi. And subject to the condition alpha i greater than equal to 0 and another condition is summation i is equal to 1 to n alpha i yi is equal to 0. So, we obtain this condition. So, how to get this condition mathematically I can show you. So, Lp was 1 by 2 W norm square minus summation alpha i yi W transpose xi plus b minus 1 we have this condition and also we have derived W is equal to weight vector is equal to alpha i yi xi.

So, i is equal to 1 to n and also we have another condition i is equal to 1 to n alpha i yi is equal to 0 we have this two conditions. So, putting the value of this W I will be getting 1 by 2 summation alpha i yi xi dot is a dot product alpha j yj xj minus summation alpha i yi xi dot alpha j yj xj plus alpha i. So, you can see alpha i yi is equal to 0 you have this condition. So, if I consider this condition then you will be getting summation alpha i summation alpha i alpha j yi yj and we have the dot product xi dot xj.

So, which can be written like this xi transpose xj also. So, I will be getting this expression from this expression actually I am getting this one because corresponding to this Lp if I want to maximize Lp I have to maximize summation alpha i. What is the advantage of this dual form? So, the advantage is it expresses the optimization criterion as a inner product of the patterns xi. So, it is a inner product between xi and xj. So, that means xi dot xj so that I can write like this xi transpose xj. So, that is the importance of the dual form this optimization criterion can be expressed as inner products of the patterns xi.

So, that is the advantage of the dual form of this Lagrangian formulation. Now considering the solution of the optimization problem. So, I have to maximize summation alpha i, i is equal to 1 to n. So, what is the solution for this? So, we are considering one condition that is called Karush Kuhn Tucker that is called the KKT criteria. That is actually the first derivative test for solution for the non-linear optimization. So, this KKT condition we are considering and based on this KKT condition I can write alpha i from the previous slide yi w transposed xi plus b minus 1 should be equal to 0.

So, this is the condition and for support vectors that is only the support vectors have alpha i not equal to 0. So, this point is very important that means for the support vectors alpha i is not equal to 0 that means the Lagrangian multiplier is not equal to 0 for the support vectors. So, the solution as the form because earlier we obtained the solution for the w. So, w is nothing but summation from i is equal to n alpha i yi xi.

So, that is the solution for the weight vector and that is equal to alpha i yi xi and in this case i only we are considering only the support vectors.

The support vectors are more important the rest of the points are not important suppose if I consider other points these are the other points if I move the other points anywhere then it will not affect the solution. So, one important point is the support vectors are more important. So, for a support vector alpha i is not equal to 0. So, the solution depends on the support vectors. So, if any other patterns with alpha i equal to 0 are moved around they do not affect the solution of the separating hyper plane.

So, that is the concept. So, the data points with non-zero Lagrangian multiplier lie on the hyper plane and they are the support vectors and they are the most informative points in the data set. So, you can see I have shown the support vectors here. So, this x these are the support vectors I have shown by arrows. So, x plus and nx minus. So, they lie on the hyper plane and corresponding to the support vector the Lagrangian multiplier alpha i they are not equal to 0.

So, the data points with non-zero Lagrangian multiplier lie on the hyper plane and they are the support vectors and I can say they are the most informative points in the data set.

So, if any other points or if any other patterns with alpha i equal to 0 are moved around they do not affect the solution of the separating hyper plane. So, corresponding to this you can see I am obtaining the expression for the W and also we can get the bias B from yi W transpose xi plus B minus 1 is equal to 0 where xi is the support vector. So, this KKT complementary condition we are considering and actually this product of Lagrange multiplier and the inequality constraints that is actually the KKT complementary condition. So, in the KKT condition you can see it is nothing but the product of the Lagrange multiplier alpha i and the inequality constraints I can write this is very important for active constraints the solution satisfies yi W transposed xi plus B minus 1 is equal to 0 then alpha i will be greater than equal to 0 that is for the active constraints otherwise in active constraints alpha i is equal to 0.

So, this point is important for active constraints the solution satisfied yi into W transpose xi plus B minus 1 is equal to 0 then alpha i should be greater than equal to 0 otherwise the inactive constraints will be alpha i is equal to 0. So, you can see here the importance of the support vectors the support vectors are x plus and x minus they lie in the hyper plane and corresponding to these support vectors the Lagrange multipliers will be not equal to 0. So, based on this weight vector because we have derived the expression for the weight vector I can determine the linear discriminant function. So, move to the next slide.

So, the linear discriminant function that is the Z x is nothing but W transpose x plus B. So, W already we have determined this alpha i xi transposed x plus B. So, i belongs to all the support vectors Sb means support vectors. So, here you can see in this expression for the Z x. So, I am writing here Z x also. So, in the expression of the Z x you can see it depends on the dot product between the test point X and the support vector xi.

So, xi is the support vector here we are considering this 1 this is nothing but the dot product between the test point x and a support vector xi. So, that is nothing but the dot product dot product between the test point x and the support vector support vector xi. So, gx we can compute like this. So, one important point is whenever we solve this optimization problem what we need to compute we need to compute the dot product between xi and xj. So, what point is important here during the solution of the optimization problem we considered the dot product between xi and xj between all pairs of training points.

That means we considered the dot product between the dot product is xi transpose and xj between all pairs of training points. So, you can see this optimization problem is nothing but we are computing the dot product between xi and xj between all pairs of training points. So, for classification what we have to consider I have to consider the discriminate function and mainly I have to consider the dot product between the test point x and the support vector xi. So, based on this I can take a classification decision. So, in this class I introduced the concept of the discriminative classifier and after this I explained the concept of the support vector machine.

In the support vector machine one important point is the concept of the Lagrange's multipliers. For the support vectors the Lagrange's multipliers will be not equal to 0 and these are the most informative points in the data set. And after this I have shown how to determine the weight vector and based on this weight vector I can determine the discriminate function. So, based on this discriminate function I can take a classification decision. So, this is the fundamental concept of the support vector machine.

In my next class I will be explaining the concept of non-linear support vector machine. So, let me stop here today. Thank you.