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-20

Welcome to NPTEL MOOCs course on machine learning and deep learning fundamentals and applications. In my last class, I discussed the concept of linear support vector machine, we considered that the training samples are linearly separable. For designing of the support vector machine, we considered a parameter the parameter is margin width, the width of the margin. And based on this parameter, I have designed the support vector machine. So, I have to solve one optimization problem and while solving this optimization problem, I define the support vectors. Support vectors are most informative points of the training data set.

And based on this support vector, I can determine the weight vector. So, after getting the weight vector I can determine the discriminate function. And with the help of the discriminate function, I can take a classification decision. In the expression of the discriminate function,

you can see I have considered the dot product between the support vectors and the test sample.

So, it is mainly the dot product between the support vectors and the test samples. And based on this we can determine the discriminate function. And after determining the discriminate function like I have explained previously, so we can take a classification decision. Today I am going to discuss about the non-linear support vector machines.

That means the training samples are not linearly separable.

So, for this I have to design the support vector machine. So, let us discuss about the concept of the support vector machine and how it can be extended for non-linear training samples.

That is the training samples are not linearly separable and how to design the support vector machine. So, let us start this class.

So, in my last class I have shown the concept of the large margin linear classifier.

Large margin linear classifier. That is actually the linear support vector machine. So, in my last class I discussed about this. So, we have determined the width of the margin like this the width of the margin that is $m = \frac{2}{||w||}$. So, that already I have explained.

And based on this margin width I have defined the optimization problem. The problem is minimize $\frac{1}{2}||w||^2$ subject to the condition $y_i(w^T x_i + b) \geq 1$. So, we have defined this condition and based on this we have formulated the Lagrangian function. So, the function is: minimize the Lagrangian function $L_p(w, b, \alpha_i)$ that is equal to $\frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i (y_i(w^T x_i + b) - 1)$ subject to the condition $\alpha_i \geq 0$.

So, this α_i is nothing but the Lagrange's multiplier.

After this what we have considered for solution of this we have determined that is we are differentiating L_p that is the Lagrange's function with respect to the weight vector and equating it to 0 and corresponding to this I have obtained the weight vector like this weight vector is nothing but $\sum_{i=1}^{n} \alpha_i y_i x_i$. So, we obtain W like this and another condition also we obtain by differentiating L p with respect to the bias b and equating it to 0. So, we obtain another condition the condition is $\sum_{i=1}^{n} \alpha_i y_i = 0$. So, these two conditions I am getting conditions means w I can determine and another condition is $\sum_{i=1}^{n} \alpha_i y_i = 0$. And for this Lagrangian optimization problem we considered this Lagrangian dual problem.

So, from this actually we have obtained the criteria maximize $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$ and it is a dot product between x_i and x_j. So, that means x_i transpose and x_j. So, these are dot product between x_i and x_j. So, subject to the condition $\alpha_i \geq 0$ and another condition is $\sum_{i=1}^{n} \alpha_i y_i = 0$, it is the Lagrangian dual problem. So, this actually we obtain like this.

So, if I put the values of this w here and considering this condition $\sum_{i=1}^{n} \alpha_i y_i = 0$ then I will be getting this form that is the Lagrangian dual problem I am getting. So, just putting the value of w and this condition in this equation. So, I will be getting this expression. So, in this case I have to maximize this quantity. So, I have to maximize this one α_i.

So, α_i is nothing but the Lagrange's multiplier and already I told you for the support vectors

α_i is not equal to 0. So, that means $\alpha_i \geq 0$ that is for the support vectors and what is the advantage of the dual problem because the dual form we are considering because this optimization criterion can be expressed as linear products of pattern x_i. So, what is the importance of the dual form the importance is that this optimization criterion can be expressed as inner products of the patterns x_i. So, that is the importance of the dual form, after determining this we obtain the discriminate function $g(x)$. So, moving to the next slide because already we have determined the weight vector and from this we can determine the discriminate function that is the linear discriminate function $w^T x + b$.

So, $\sum_{i \in SV} \alpha_i x_i^T x + b$. So, we obtain the expression for the discriminate function $g(x)$ like this. So, here you can see what is actually the meaning of this expression. So, you can see how actually we do the classification you can see the discriminate function is nothing but the dot product between the test point x and the support vector x_i. So, you can see this is nothing but the dot product between the test point x and the support vector x_i and already I told you the support vectors are the most informative points of the training data set and corresponding to these support vectors this Lagrangian multiplier is not equal to 0.

That means for the support vectors the Lagrangian multiplier lie on the hyper planes. So, this is the final expression for the $g(x)$. So, this classification, the classification relies on on a dot product between the test point, test point is x and the support vector, the support vectors x_i. So, also one important point is for solving this optimization problem what actually we need to compute.

We need to compute the dot products x_i and x_j between all pairs of training points.

So, that is also one important point. So, whenever I want to solve the optimization problem I have to compute the dot product between x_i and x_j between all pairs of training points. So, for support vectors this $\alpha_i \neq 0$ that is the Lagrangian multiplier is not equal to 0 and these support vectors lie on the hyper plane and they are the most informative points in the data set. And if I consider other patterns suppose if I consider for other patterns α_i will be equal to 0 and if I moved around they do not affect the solution of the separating hyper planes. So, that means only we can see the support vectors because

they are the most informative points in the data set.

So, if any other points with α_i is equal to 0 are moved around they do not affect the solution for the separating hyper plane. So, that is the fundamental concept of the support vectors. So, let us consider now what will happen suppose if I consider this data is not linearly separable and suppose I have noisy data or maybe the outliers so whether this model can be modified or what is the modification I need to do that point we need to consider. So, let

us move to the next slide. So, now what we are considered suppose this data is not linearly separable.

So, that point also I will be explaining later on for non-linearly separable data I have to consider non-linear support vector machine or maybe something like if I consider noisy data outliers etcetera. So, for this what modifications I need to do. So, already I have defined the support vector machine now I have to consider this point the data is not linearly separable and maybe we can consider the noisy data or the outliers. So, before going to that point that is a non-linearly separable data first I am considering the noisy data or the outliers after this I will be explaining the concept of the support vector machine for the non-linearly separable training data. So, for this case that is case is the noisy data and outliers what we are considering we are considering slack variable,

we are considering slack variables that is x_i.

So, slack variable x_i is considered to allow misclassification of the difficult or the noisy data points. So, in the figure you can see I can show the difficult points. So, this is a difficult point one difficult point and similarly I can show another difficult point and that is the noisy data or noisy sample points or maybe the outliers. So, corresponding to these two points there will be misclassification. So, to consider this misclassification we are considering one slack variable and that slack variable is the x_i that is we are considering to consider the difficult or the noisy data points because

the misclassification take place because of these data points.

So, I have in the figure I have shown these two data points you can see number 1 I can show or number 2. So, these data points are not correctly classified by earlier design support vector machine. So, corresponding to this case my formulation will be, my formulation will be minimize $\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$ such that we are considering $y_i(w^T x_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$. So, this parameter C can be considered as a way to control overfitting. So, I am writing here the parameter C can be viewed as a way to control overfitting.

So, you can see after considering the slack variable ξ we have considered the optimization problem that is the formulation I have shown minimize $\frac{1}{2}||w||^2$ plus another term we are considering because of this noisy points. So, this term we are considering in addition to the earlier term the $\frac{1}{2}||w||^2$ and that is actually this ξ_1 or ξ_2 this is actually some offset we are considering. So, I can consider as offset. So, ξ_1 or ξ_2. So, in the formulation I have included that one and based on this formulation we can also consider the Lagrangian dual

problem.

So, move to the next slide. So, considering this the formulation will be, formulation is Lagrangian dual problem. So, maximize $\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$. So, you can see it is a dot product between x_i and x_j. So, this is the formulation that is the Lagrangian dual problem corresponding to this formulation considering the slack variables. And the condition is such that $0 \leq \alpha_i \leq C$ and also another condition $\sum_{i=1}^{n}\alpha_i y_i = 0$.

So, that is the only modification we have to do if I consider the slack variable x_i and this I have done only for the noisy data or the outliers. But for the non-linearly separable data we have to consider the non-linear support vector machine. So, let us discuss about the non-linear support vector machine. So, this problem or this solution I am considering only for the noisy data or maybe the outliers. So, for non-linearly separable data I have to consider the non-linear support vector machine.

So, move to the next slide. So, now the concept of non-linear support vector machines. So, let us consider the figure number 1 that is the first figure, figure number 1. So, in this case what we are considering in figure number 1, dataset that are linearly separable with noise. So, in a figure number 1 we have considered that one. And in this case we can employ the support vector machine that we have explained earlier.

But if I consider the figure number 2, you can see here in the figure number 2 these samples are not linearly separable. So, it is very difficult to find a decision boundary between the samples of the two classes and these samples are not linearly separable. So, that means figure number 2 I can consider non-linearly separable data that is the figure number 2. So, corresponding to this what is the solution for this? If you see the figure number 3 in figure number 3 what we are considering mapping of data to a high dimensional space. So, just I am doing the mapping of data to the high dimensional space.

So, in the low dimensional space this data is not linearly separable. But after the projection into the high dimensional space these samples will be linearly separable. That means what we are doing this mapping of data to a high dimensional space. So, from the low dimensional space I am mapping into the high dimensional space and that is the fundamental concept of the non-linear support vector machine. So, in my next slide I can show that concept.

So, what is the non-linear support vector machine in the feature space? So, in the figure you can see in this slide just I am doing the mapping the mapping from the low dimensional space into the high dimensional space. So, these are non-linear support vector machine SVM and we are considering the feature space. So, the idea is you can see the original

input space can be mapped to some high dimensional feature space where the training samples is separable. So, in the figure you can see this is the low dimensional space and I am doing the mapping by considering some mapping functions. The mapping of the low dimensional data into the high dimensional space.

So, it is the higher dimensional higher higher dimensional feature space. So, that means I can write the concept here. The concept is the original, the original input space can be mapped to some higher dimensional feature space where the training set is separable. The original input space can be mapped to some high dimensional Feature space where the training set is separable. So, in the figure you can see I am showing the mapping function and with the help of this mapping function I am projecting the low dimensional data into higher dimensional Feature space.

So, this is the concept of the non-linear support vector machine. So, whenever I do the projection into the higher dimensional Feature space the samples will be linearly separable. So, you can see in the figure the original data is not linearly separable but after the projection the projected data will be linearly separable that is the fundamental concept of the non-linear support vector machine. So, mathematically how to consider this case.

So, let us move to the next slide. So, what is the discriminate function now? So, the discriminate function by considering the mapping $g(x) = w^T\phi(x) + b$ that is equal to the $\sum_{i\in SV}\alpha_i\phi(x_i)^T\phi(x) + b$. So, you can see the modification in the discriminate function. So, what modification you can observe here only this term we are considering the mapping. So, earlier it was a dot product between x_i and x now we are considering the mapping function $\phi(x_i)$ transpose $\phi(x)$. Now, in this case no need to know the mapping explicitly because we are only considering the dot product of the feature vectors in both the training and the test.

I repeat this no need to know this mapping explicitly because we only use the dot product of the feature vectors in both the training and the test. So, corresponding to this what we can consider? So, we can consider a kernel function. So, now a kernel function can be defined as a function that corresponds to this dot product of the two feature vectors. So, I can write a kernel function, a kernel function can be defined now defined as a function that corresponds to the dot product of two feature vectors in some expanded Feature space in some expanded Feature space means some high dimensional space in some expanded that is the higher dimensional Feature space. So, that means to consider this one to consider this one we are defining a kernel function what is the kernel function? The kernel function is $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$.

So, to consider this one that dot product we are considering a kernel function the kernel

function is something like this. So, I can give some of the examples of the kernel functions. So, commonly used kernel functions I can give some examples. So, maybe in the next slide I can give some examples. So, some of the commonly used kernel functions maybe we can consider the linear kernels.

So, one example is the linear kernel $K(x_i, x_j) = x_i^T x_j$, this is the example of the linear kernel maybe we can consider the polynomial. So, in the polynomial kernel we can consider $K(x_i, x_j) = (1 + x_i^T x_j)^p$. So, this is the expression for the polynomial kernel and one popular kernel is the Gaussian kernel Gaussian or it is also called the radial basis radial basis function this is called RBF radial basis function kernel this is a very popular kernel. So, the concept of RBF I will be explaining later on whenever I will discuss the concept of the artificial neural network that time I will explain the concept of radial basis function. So, the expression for this $K(x_i, x_j) = exp(\frac{||x_i - x_j||^2}{2\sigma^2})$.

So, this is the popular one that the Gaussian kernel and maybe also we can consider the sigmoid kernel sigmoid function also we can consider. So, these are some examples of kernel functions. So, based on this we have to formulate the Lagrangian dual problem because we are considering the non-linear support vector machine and we have to solve the optimization problem. So, for solution of this optimization problem the formulation is important that already I have discussed in case of the linear support vector machine we have to consider the Lagrangian dual problem. So, move to the next slide that is the non-linear support vector machine support vector machine and we have to consider optimization so, like in case of the linear support vector machine we have the formulation.

This formulation is that is a Lagrangian dual problem the formulation is maximize $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$. So, this is the kernel function we are considering the conditions already I have explained the important condition is such that $\alpha_i \geq 0$ and another condition $\sum_{i=1}^{n} \alpha_i y_i = 0$.

So, from this I can determine the discriminate function. So, discriminate function is $g(x) = \sum_{i \in SV} \alpha_i K(x_i, x) + b$, we are only considering the support vectors because support vectors are the most informative points of the training data set. So, α_i is nothing but the Lagrangian multiplier and we are considering the kernel function k is the kernel function. So, you can see we have to consider x_i, x_i is nothing but the support vector and x is nothing but the test sample plus b. So, we are obtaining the discriminate function. So, this optimization technique is the same technique already we have discussed in case of the linear support vector machine.

So, we can obtain the expression for $g(x)$. So, this is the expression for the $g(x)$ in case of the non-linear support vector machine. So, only we are considering the kernel function k that is between the support vector x_i and the test sample x. So, for this algorithm the non-linear support vector machine. So, what are the steps I need to consider the steps will be like this. So, for this non-linear support vector machine what are the steps.

So, non-linear support vector machines. So, first I have to select the kernel function choose a kernel function. So, generally we considered the radial basis function or that is the Gaussian kernel we can consider number 2 we have to solve the quadratic programming problem that is actually the optimization problem and number 3 because after solution of this non-linear optimization problem we have to determine the weight vector and based on the weight vector we can determine the discriminate function. So, construct the discriminate function from the support vectors. So, these are the steps of the support vector machine and the kernel generally we considered the Gaussian kernel or the radial basis function we considered. So, in this case you can see in the expression for the non-linear support vector machine we obtain $g(x)$ that is the discriminate function that is something like this $w^T\phi(x)$ that is the mapping function plus b.

So, it is equal to $\sum_{i\in SV}\alpha_i\phi(x_i)^T\phi(x) + b$. So, in this expression here you can see we are considering this part that is the mapping function we are considering. So, in the linear support vector machine if you see I considered the dot product between the support vectors and the test sample. In this case we are considering the mapping function the mapping functions are $\phi(x_i)$ and $\phi(x)$, but in this case since we are considering the dot product we do not need to represent the mapping explicitly. So, that means the mapping is not so important because we only considered the dot product.

So, we do not need to represent the mapping explicitly. So, that is the dot product between $\phi(x_i)$ and $\phi(x)$. So, that is why with the help of this dot product we can determine the discriminant function. So, we need not consider to represent the mapping explicitly. So, this is the fundamental concept of the support vector machine. So, briefly I have introduced the concept of the non-linear support vector machine.

In this class I have explained the concept of the non-linear support vector machine I have explained what to do for the non-linearly separable data. So, the concept is if the training samples are not linearly separable in a low dimensional space I have to project them into a higher dimensional feature space. In the low dimensional space the samples are not linearly separable, but in the high dimensional space the samples will be linearly separable. So, I have to define the mapping function and based on this mapping function I have formulated the optimization problem. So, we have defined the Lagrangian dual problem and based on this solution of this we have obtained the weight vector w and after getting

the weight vector w we can determine the discriminant function $g(x)$ and with the help of the discriminant function we can do the classification.

So, this is a fundamental concept of the non-linear support vector machine which is almost similar to the linear support vector machine except the concept of mapping. And for multi-class classification we may extend these principles the principle of the support vector machines also we can consider the principles like one versus one classification or one versus all classification techniques for multi-class classification and that concept already I have explained the concept of one versus one and one versus all classification techniques. So, this is about the support vector machine. So, let me stop here today. Thank you.