Transcriber Name: Saji Paul
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology, Kharagpur

Lecture - 58
SVM (Part - V)

(refer time: 00:31)

Hi guys, so and continuation of our lecture on SVM. Today what we will see as I have mentioned in my last lecture so this transformation data transformation from to make it the non-linear separable data which are not separable linearly. So, there is non-linear data, we try to use linear SVM product, for that what we have done we have we try to use so that we can use the linear SVM for this non-linear separable data.

We have transformed the data from non-linear data to linear data for that what we have used. When transformation was in a input space of say X and that transformation we have changed it to a different input space of a higher dimensional. So, when we have changed it to higher dimensional the computational increases. Computational when we try to come compute the optimization constraint problem.

They are also computational increases when we try to find out the different parameter's values there also the computational when we do the dot product W . x w r x i . x. So, this dot product basically we do if it is an anime student space so it is n multiplication plus an addition, so similar when the dimensional increases the computational increases. So, now how to solve that how to address that.

So, these are two concepts one is the dual formulation of optimization problem and another is that we will introduce a new very intelligent method which is called kernel trick.
(refer time: 01:59)

So, as we have discussed, what are the different issues the first of course there are one issue I forgot to mention that one issue is also mapping, how will I map what mapping procedure I will use as I have shown you if in the whether I will use suppose the dimension is in R 2. So, then dimension of R 2 whether I will keep it in R 2 only or I will use I will transform it to R 3 or I will

transform it to R 4 basically how many attributes do I need to use.

So, based on the number of monomials. So, how many how each monomial's, I will be mapping into an attribute of the higher dimension. So, that is this mapping which mapping exactly there is no hard and fast rule that I will have to use this mapping. So, since there is no such hard and fast rule, we how I will do the mapping that is one of the issues then definitely cost of the mapping dimensionality problem and computational costs. So, these are some of the issues.
(refer time: 03:01)

So, mathematicians have clearly proposed an elegant solution to the above problem. So, that is the dual formation of optimization problem and kernel trick. So, let us see what is that?
(refer time: 03:13)

First, we will see dual formally formulation of optimization problem.
(refer time: 03:18)

So, we have already learned a Lagrangian formulation to find a maximum margin hyperplane is not it. So, we have formulated our problem as a maximization problem, earlier it was a minimization problem 2 / W and which we have changed it to W / 2 and we found it we have formulated as a maximization problem. So, we have learned a Lagrangian formulation to find a maximum margin hyper plane of a linear SVM classifier.

Such, formulation what we have seen such formulation is called the primal form of constraint optimization problem. The formulation what we have done till now, this is also called the primal form of the constraint optimization problem. So, primal form of the Lagrangian optimization problem this was my primal form I have to minimize this $W^2 / 2$ is not it subject to this, is not it.

My difference between the two, what was that my difference between 2 vectors, I have to maximize this is not it so what was this that is 2 / W, so I had to maximize this is the difference. So, this I have formulas as a minimization problem how did I formulas as a minimization problem I have used this, is not it. So, I have to minimize this subject to y i W . x i + b greater equals to 1.
(refer time: 04:51)

So, this was my Lagrangian form, this formulation minimize this subject to this, thus I have written in a Lagrangian form as this earlier I have not used the term p, I have just written L is equals to this this was my Lagrangian form remember we have done this. Now, let I am calling

this this thing actually it is basically that time we did not mention specifically this is called the primal form, this is the primal form.

So, a substrate we can use a subscript p to distinguish that this is a primal from. Why we are using this because there is some other form also that is the dual form. Now, what we will see what is the dual form, now here what is the λ, λ is a dummy variable it is also called lagrangian multiplier. So, L p is called the primal form of lagrangian optimization problem.
(refer time: 05:43)

So, now from this is the primal form, now from this we will have to find out the dual form. Why we will find out the dual form? We will see that. The dual form to find out the dwell from what we have done first as we have seen to minimize this lagrangian what we have done we have done del L / del W del L / del b, from that we have found different expression is not it. So, here del of L p / del W we found this W is equals to this is not it.

Because, what was this W minus this we got 0 so, W is equal to this that is what we have done earlier you can refer to the previous lectures. And this also when we have done del L / del b we got this summation of λ i y i = 0. Now, what I will do is that in my primal form lagrangian, I will be substituting this W value and I will be substitute, basically I will be substituting this value and this value and my I will substitute this in my primal lagrangian form.
(refer time: 06:49)

So, this was my primal lagrangian form. So, first I have done first order differentiation with respect to W and I got this then, I have done first order differentiation with respect to b and I got this. This two I will be using in this expression and there is actually a big one-page calculation. So, I did not show it in this slide so this once we substitute this and once, we will be using this.

In this expression we will get something of this form, where there is no b no W all in terms of λ y and x y is the class and x is my observation. So, when I just use these two equations this one, this equation and this equation in this equation after lots of calculation is as I told you it is a big one page calculation then, we will be finding this so in short I can write it this way and this I am calling it a dual form dual lagrangian form.

Because, primal (()) (08:00) that was my lagrangian form and in that lagrangian form what I have done I have done some sort of substitution, I did not do anything else I have done just some sort of substitution. What I have substituted I have differentiated with respect to W, I have the sensitive with respect to b and whatever I got that I have substituted it in primal form and I got a different expression which I am calling it the dual form, dual form of lagrangian.

So, this is my dual form of lagrangian. So, in my dual form of recognition I do not have W term I do not have b term and this is my this my quadratic term is minus here here, my quadratic term is this is my quadratic term is not it this will give me a quadratic term this is minus here.
(refer time: 08:41)

So, the key difference between the primal L p and the dual L p forms of lagrangian optimization problems are given here. So, there are different differences there are many differences between the L P and L D. So, what are the difference let us see, so L p involves a large number of parameters obviously, our L p has last number it has W it has b it is λ I on the other hand L D involves only L i that is the lagrangian multiplier.

L p is the minimization problem as the quadratic term is positive. L p when we use L p there is a minimization problem because there the quadratic term is positive but however in L D our quadratic term is negative, I have already mentioned there hence it turn outs to be a maximization problem. So, since a quadratic term is negative then, it is no longer a minimization problem basically it turns out to be a maximization problem.

So, L p involves the calculation of W . x, in L p what we do when we find out L p so our calculation is w . x dot product of W and x. But whereas in L D involves a calculation of x i . x j x c we are trying to do dot product of x i and x j, two different observation data. Whereas in L p we have done dot product of W and x if W is n dimensional definitely if x is in n dimensional W is also an x dimensional.

W . x means we will have to how we do calculate a dot product is not it, that you know I do not have to tell you. So, but whereas in L D I will do the calculation of x i and x j. If I do the calculation of x i and x j then, things become easier why the what is when we try to find out a dot product of two vectors x i and x j what we try to actually find out everything, it is there in the slide let us see.
(refer time: 10:45)

So, there are key difference, so the SVM classifier, now once we have used the lagrangian multiplier in a primal form earlier we have seen Lagrangian multiplier we have used and then we have found out the different values of the parameters W b and all. And then given any test data we have to find out it falls in which class, if which class so this is this was what we use remember.

So, del of x means which class it belongs to so this is how we do it W . x + b so x is the test vector. So, if we find it a negative sign then it falls in the class below the hyper plane if it falls in a positive sign, it falls the class above the hyperplane, what may be the above class what may be the below class that accordingly what you have specified. Where is the dual version of the classifier?

So, it was in primal form this was our thing whereas in the dual form we will be using this because, this is my W is not it, this is my W. In dual form I will not be using this W . x + b but instead I will be using this formula x i being the is support vector and assume that there are m support vectors. See the difference, here in the primal form to find out the class of a x we have what we have done we have found out the W value.

We have found out the b value W along which the if it is the n dimensional so W 1 to W n we found out the value and definitely x is n dimensional then only W will also be n dimensional. So, accordingly we found out the different x values also and we do the dot product of this and then we find out the magnitude and the sign of x so as to classify (()) (12:41) to predict it falls in which class.

Now, in the dual form, dual form we no longer do the dot product of W . x but whereas in dual form this is our expression. Here, so this is we do we have a dot product of x i . x where x is my test vector which is x i x i ordered different support vectors, we have seen this here x i as the different for all those support factors λ i is not equal to 0. For vectors which are not support vectors λ i is equal to 0 is not it.

So, this equation becomes zero so here will be only multiplying all those support vectors all those vectors which are support vectors and with those I will do the dot product of my test vector. And if I do the dot product of the stress vector I what difference it will make will see. So, till now we have as I told you for solve the time different issues of the transformation using the linear SVM for the non-linear data, what are the different issues we have seen.

So, now to solve those issues as I mentioned there were two ways how we will solve it then, there is not two ways there is a trick how we will solve this problem the first step of is dual formulation. So, we have seen we have seen this dual formulation now next is the kernel trick. Thus, here this does the L p and L D are equivalent is not it, because already what instead of W I am using this the W value is this we have already found out. So, L p and L D are just equivalent.
(refer time: 14:20)

So, it is just a different form, this is the primal form, this is the dual form now we will see what is the kernel trick.
(refer time: 14:27)

So, we have already learned an idea that the training data which are not linearly separable can be transformed into a higher dimensional feature space such that in higher dimensional transform space a hyper plane can be decided to separate the transform data and hence the original data. So, this was not linearly separable we have transformed into the higher dimensional space.

Once we have transformed into higher dimension space we could use a hyper plane, this is the hyperplane we could use a hyperplane to separate this, so non-linearly separable data because now it has become linearly separable.
(refer time: 15:05)

So, clearly the data on the left of figure this is not linearly separable. So, this is not linearly separable yet if we mapped it to a 3D earlier it was supposed in 2D we have mapped it to the 3D space using φ transformation then with the map data it is possible to have a linear dimension boundary, linear decision boundary that is a hyperplane in 3D space.
(refer time: 15:29)

So, see this example suppose there is a set of data in two dimensional in R 2, we have a set of data which is in two dimensional, let the hyperplane in R takes this form. Then, it is data is in the two dimensions suppose, if I consider a hyperplane because it is these are not linearly separable it is nonlinear separable suppose it takes the hyperplane takes this form quadratic form.

Which is the this is nothing but the equation of an ellipse or maybe you can consider this you can consider this sort of data this is an ellipse. So, now suppose we will be doing Φ transformation. Now suppose φ is a mapping from x that is in R 2 to Z z 1 z 2 z 3 to R 3 in 3D space, I will do the mapping. So, say for this term I have used Z 1, Z 1 = x 1 z 2 is equal say this term, Z 3 is equal to say this term.

There are three nominal's, so for each 3 nominal's each nominal I have used one attribute of higher dimension. So, say z 1 = x 1 $^2$ z 2 = root over 2 x 1 x 2 z 3 = x 2 x 2 $^2$, I have used three different attributes for the different nominal's there were three nominals, so I got this.
(refer time: 17:02)

Now that means when I use this transformation then this has become my equation. So, this is very much a linear equation, is not it this is very much a linear equation so that is that means it is a hyperplane earlier it was a hyperspace. So, now this is very much a hyperplane from which we will have to find out basically the different values. Once you can define the values, we can will be able to classify the test data. This is clearly a linear form in 3D space.

In other words, W . x + b = 0 in R 2 is has mapped into W . z + b dash = 0 in R 3. So, a 2 dimensional it is W . x + b = 0 which we have mapped it to 3 dimensional W . z + b dash if we use the different com intercept that is b dash = 0 that is in R 3, this means the data which are not linearly separable in 2D now R separable in 3D because this is a linear plane. This plane separates the data. Earlier which was separating the data this was separating the data.

It was a hyperspace. A hyperspace was separating the data now what we have we have transformed into a linear plane. This linear plane is now separating the data. So, now we will see the generalization of this formulation which is the key to the kernel trick now we will see what is the kernel tree.
(refer time: 18:38)

See here, so our classifier is earlier our classifier was W . x + b is not it, what is the classifier to given a class X a given an observation at it is false to which class how do I find W . x + b we have seen here, this W . x + b, now it has our thing has become this, this is our classifier now. So, this is my classifier now, I have used a dual form, dual form of the lagrangian. So, this is my classifier now.

So, now this is my classifier in the x dimensional whatever it is my input dimension input dimensional may be 2D 3D over the waves. Now, since I am converting it to an higher dimension I am doing the transformation so as to convert this linear to not sorry non-linear to linear so this transformation I am using say z. So, now I will class for me now if I want to classify given a test that x it falls into which class basically x first.

I will have to convert it to z and that means I will have to find out this z belongs to this class. So, I am now I am telling my classifier z instead of classify x now I can write classify z is not it because, x I will convert it to z. So, classifier z if I convert it this, I can write it in this way x i I got I would have done this φ transformation x also I have done this φ transformation. Earlier, my dot product was between x i . x now my dotproducts is φ . x i to φ . x.

So, what was my dual problem? Since, it was my negative term is the quadratic term is negative

so it is maximized we have already mentioned. So, I will have to maximize this is not it. Once, I maximize this I will find out by maximizing this I will be able to find out my λ value. So, maximize this subject to this constraint we have seen this constraint is not it, when we have found out this expression when we found out this expression then we have seen this constraint.

So, essentially, we will have to maximize this subject to this constraint and when we maximize this subject to this constraint accordingly, we will be able to find out the λ value. Once, we found the λ value then we will be able to do this classification by doing base dot product.
(refer time: 21:14)

Now, question here is how to choose φ the mapping function that is X is mapped to Z, so that linear SVM can be directly applied. So, a breakthrough solution to this problem comes in the form of method known as kernel trick. How to do this mapping? That was our main issue, so that is a one the very good solution is the kernel trick. So, with now what is the kernel trick, we know that a dot product here that I was still talking of this only I told I will discuss later.

So, when we talk of dot product, dot product is what we often try to measure the similarity between two input vector is not it when we do the dot product actually what we have to do we want to find out the similarity between the two vectors. So, this is also we also call it cosine similarity, the angle between the two vectors. So, if the angle between the two vector is 0 then, we can say both the factors are almost similar.

However, if the angle between the vectors is 90 degree, then we can say both the vectors are orthogonal, is not it? So, for here it is written we know that the dot product you know right is what is the dot product what we dotproducts cross product you know all those things, is not it? So, the we know dot product is often regarded as a measure of similarity between two input vectors we call it cosine similarity.

For example, if X and Y are two vectors then we can write it in this way. So, how do we find out the angle from this expression we can find out the angle, is not it? What is cos theta, cos theta is magnitude of X into magnitude of Y / X . y, that will give me cos theta. So, theta will be cos inverse of that what we get, so for similarity between X and Y is measured as a cosine similarity.

So, if theta = 0 then we and then they are more similar the vectors are more similar if the theta is 90 then it is orthogonal as if theta is also 0 then cos theta = 1 then if theta = 90 what will be cos theta? Cos theta will be 0. So, if the cos theta values reduces from 1 to gradually 0 our similarity reduces gradually.
(refer time: 23:42)

So, now analogously if X i and X j are two tuples then X i X j is regarded as measure of similarity between X i and X j, see here when using the classifier what I have used X i . X j I have used the two dot product. Similarly, here also I have used two dotproducts. So, if X i and X j are two tuples then X i . X j is regarded as a measure of similarity between X i and X j.

So, again φ of X i and φ of X j are transformed feature of X j as a respectively, does φ of X i the φ X j is a is also should be regarded as a similarity measure between φ of X i and φ X j. Because, if X i is a vector, definitely φ of X i is also a vector is not it. We just we have transformed it from one dimensional to another dimensional but, it is X i is also a vector to φ of X i is also vector.

So, if X i . X j if you regard is as a measure of similarity then φ of X i . φ of X j also will be regarded as a measure of similarity in the transform space and a different dimension. This is an important revelation and is the basic idea behind a kernel trick. Now, natural equation arises if both measures the similarity then what is the correlation between them. So, now X i . X j this dot product also measures the similarity.

When we try to find out the class of a test data basically, we try to find out this test data how much it is similar to the different support vectors when we are doing the dot product means essentially, we try to find out the state states that is how similar with the support vectors based on the similarity we basically classify it to test data or the other data. So, what we have done for X i . X j X i . X basically that is the test data my test that is X.

We find out the similarity measure between the support vector and the test vector. Here, also it is φ of X i do φ of X j when we do here also, we do the same similarity measure but in a different dimensional in a transform dimension, is not it? So, now what is the relation between that is there now is there any correlation between X i X j and φ X i φ X j. Now let us try to find the answer to this question through an example.
(refer time: 26:11)

So, without loss of generality let us consider a situation stated below. We will be using the same equation which very recently we have used. So, suppose it was from R 2 the input dimensional to input space we have converted into R 3. So, it was x 1 = Z 1 x 2 = Z 2 2 root 2 x 1 x 2 12 is actually the same example which we have used. So, if my x i is this x i 1 x i 2 if my observation given absorption x i is x i 1 x i 2 and x j is this are designed any two vectors in R 2.

Similarly, then my φ of x i will be this and φ of x j will be this is not it nothing but the attributes here the I have three attributes Z 1 Z 2 Z 3, so these are the three attributes in my transform space. So, under two transform function of x i and x j in R 3.
(refer time: 27:12)

So, now let us try to find out the dot product of φ of X i and φ of X j. If I find out the dot product of φ of X i and φ of X j now this is my one vector this is one vector, I cannot, I write it in this way I will just when I have to find out a dot product. So, I have written it in this way, so this is what I got. This is nothing but this is a vector φ of X i is a vector φ of X j is a vector I have written it in this way and I got this.

This again I can write it in this way is not it, this I can write in an a + b $^2$ again, this I can write it in this way and what is this is nothing but X i . X j $^2$. See here that means φ of X i . φ of X j is X i . X j$^2$. Ignore the term whole square, whatever it is basically what I found as φ on X i and φ of X j is somehow related to x i x j that means there is a correlation between φ of X i . φ of X j.

And X i . X j there is some correlation, the square cube whatever it is forget about that just that φ of X i . X j is score related to X i of the example because we have seen when we have done φ of x i . φ of x j ultimately, we got X i . X j $^2$.
(refer time: 28:39)

So, we took reference to the above example this that we have seen we can conclude that φ of X i do φ of X j are correlated to X i . X j. In fact, this we can in general also we can prove that, for any feature vectors and a transform feature vectors. We can take, if we take any picture matters any transformer this is just a simple example, we have seen for any just some feature vectors and is transformed if we see we will be able to prove this there is a correlation between the transform vector as well as the original vector.

And this proof it is not possible to discuss the proof here. More specifically there is a correlation between dotproducts of original data and dotproducts of the transform data. So, based on this above discussion we can write dot product of X i does X j it implies that the dot product of φ of X i of φ of X j it implies this K, this K denotes a function which is more popularly known as kernel function.

Now, what is this kernel function what is this K we will discuss in the next lecture just to give a hint just to understand what is this kernel function. Let me give you an analogy of a very non-technical analogy like suppose, I have here two fruits and I have asked you to compare these two fruits. So, the person whom I will come whom I have asked to compare; this two fruit

suppose he is not a very intelligent person.

So, then what he will compare if both the fruits are red suppose then this both the fruits are red that is once simulate means how does both the fruits are how similar they are if I want to find out. So, he will tell both the future red in colour both the fruits are round size both the fruits have the almost same size. So, this sort of comparison will give but if I ask to compare to a very intelligent person, so on what way the intelligent person will compare.

What she will do is that she will try to find out what are the different values of this food, how what vitamin E has it has B vitamin magnesium, zinc whatever it has and accordingly the other fruits all over it has and how it helps the person who consumes it how it improves the humidity of the person who consumes it. So, accordingly you will find out the similarity measure of these 2 fruits that is how intelligent person will do the other dumb person will just based on the colour shape size will give in compression.

Now you see this the intelligent person will do this sort of compare similarity between this to fruit. Now, likewise suppose if consider two other fruits in the different soil this is in Indian soil suppose in UK soil, in UK soil suppose we have two different fruits. So, in the similarity measures so if we try to find out a similarity measure suppose this when you are comparing that you say two different fruits here.

I have compared to fruits X and Y there in the UK soil another two fruits A and B so similarity measure may be same level of vitamins content zinc content magnesium content whatever it is and how it improves the immunity level. So, if that is there is a correlation between those two fruits and these two fruits in Indian soil then we can say these are correlated, these two maybe these are different fruits the loop wise it is may be very different.

So, but at similarity wise this is the simulative measure of this two, what how it affects a human being whoever is consume it and similarly these two fruits in the Indian soil how it is affects how when a person consumes it this way, we can say this to food if the same similar we can say that they are correlated. So, that is what basically is the kernel function, now what is key there are many standard functions.

Which we will which can be used as a kernel function which and how we can use this for us in this SVM that will be seeing it in my next lecture.
(refer time: 32:51)

So, in this lecture we learned about a dual formulation of the optimization problem, from the primal form how we have converted it to the dual form and we have also learned an important concept that kernel trick we have just learned it we how it can be used and how it helps us in solving this transformation problem that we will be seeing in my next lecture.
(refer time: 33:17)

With that thank you guys.