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 - 54
Support Vector Machine (Part - I)

Hello guys, today we will start a very interesting and important classifier and it is something like if you this classifier I am talking about support vector machine. If you do not know support vector machines means like you do not know anything it is sort of this support vector machine is like it has become sort of, I can say it has become sort of status symbol. You do not know status back as SVM so that is it so SVM is such an important classifier.

So, and but the problem with this is just SVM you will hardly find any good material in the net even in a book also it is not written many of the textbook it is not written very properly. So, I some try to get information's from all this and try to put it in as simple words as possible because it is a bit difficult to understand. And initially it is very easy but as I proceed well it is a bit difficult to understand.

So, please pay attention and that is all it is a gun let us learn this. Now let us start this support vector machine.
(refer time: 01:35)

So, first we will in this class first we will introduce this support vector machine and then in support vector machine there is a concept called maximum margin hyperplane. So, what is this maximum margin hyperplane that is also we will see here.
(refer time: 01:51)

So, this support vector machine it is fondly called by people as SVM that is not mostly see people call it as SVM. So, it has that has received considerable attention this classification as it is considerable attentions and this technique as one of the co-inventor of this technique is Vapnik, Vapnik is not the only invented is one more person which at present I cannot remember his name and but the Vapnik why he is important he is the person who has used this technique in some in statistical learning theory.

As the first person to have used this technique in the statistical learning theory and made this popular. So, he is one of the co-inventor of this. So, in SVM what it does is that it is as a tax of classification it searches for the optimal hyperplane now also called decision boundary. So, when I talk of hyperplane what is this called also called the distance boundary, separating the tuples of one class from another. Here you can see the figure.

So, suppose this is one class of data this is one class of data these are two different classes of

data. So, to separate how it searches for a hyperplane this is the hyperplane, it searches for some sort of boundary line or something certain which differentiate between these two classes. Suppose a wall it is a wall the data behind and wall is behind this wall is belongs to some other category data in this side of wall belongs to some other category.

So, that is a hyper plane basically and hyperplane has this is a hyperplane and it has a boundary, each hyper level has a boundary. Now we will see what is this boundary. This hyperplane is also called decision boundary means that is give making us that hyperplane is helping us in taking the distance given and it is data whether the data falls in which category, data falls in this and it falls in this category or it falls in this category.
(refer time: 04:02)

So, SVM what are the advantage of SVM, why people when we talk of classification techniques mainly SVM people think of SVM, why it has many advantages. First is SVM works well with higher dimensional data and thus about its dimensionality problem. With higher dimensional data it can work very well like kNN and we have seen kNN and it is a curse of dimensionality it is very difficult to work with higher dimensional problem when if we use kNN algorithm.

And again, the if for different data scales we cannot use kNN algorithm. So, if the different data skills of the attributes. Likewise, this SVM does not have all those disadvantages SVM, it works well with higher dimensional data and thus it does not have the dimensionality problem cards of dimensionality does not have. Although the SVM is based on classification that is the training time is extremely high.

SVM the training time is extremely high, the result is however highly accurate, further testing and unknown data is very fast. Here we can say SVM is the eager learner remember what is in kNN is a lazy learner it does not think during training. So, SVM does the whole work during the training time only it does the whole work during the training time, the whole workman is basically finding out the hyper plane which differentiates between different classes.

So, and then it takes a long time for that of course but then it is very accurate and once that is done for and testing time is very fast, SVM is less prone to overfitting than other methods. So, what is overheating? I will see it in the next slide. It also facilitates compact model for classifications. See here you see and here we have four classes. So, these are the different hyperplane, this is one hyperplane, this is one hyperplane, this is one hyperplane, how it differences between the classes.
(refer time: 06:07)

So, this is the over under fitting, over fitting, optimum fitting what is that? You see here and this the first line, suppose these are the data points this blue colour dots are the data points. The blue colour dots are the data points and suppose I found a function my function what I found is just a straight line. Can you see the straight line? The straight line is my function and classification in all this classification and all what we preservation, (06:39) whatever it is we just try to find that function, is not it?

So, here I found that the function is my straight line. But if this is the straight line and what happens many data speed does not fall into straight line or this figure is not very accurate. Anyway the some lines some points maybe it falls into straight line some points it falls in a straight line there are but many data does not falls in a straight line. In such case we call as under fitting. When in such cases what happens?

For such under fitting curve then what happens our training error is also quite high and our testing error will definitely high. When our training errors are high definitely our testing error will also be quite high in case of under fitting. So, this is an very good example of an under fitting. So, actually this; my true function would have been this but I got this as my function. So, this is there are lots of training error.

Training error means when I the data which I use for training and there is data also when the function what is given even that function did this data is not satisfying the function itself, the training data also. This training data means whose class level is known. When I am using this function so when I am using this function to classify it into two classes suppose I have two classes classified into two classes from my training data only I am you know suppose for some data I need to get it in B class but I am getting it in A class.

For some data I need to get it in A class but I am getting in B class so that is called training error. So, in under fitting case there is lots of training error when there is training error definitely there will be testing error also. So, in case boat and training and testing errors are high in case of under fitting. Now you see over fitting, over fitting means you try to you have found out the hyperplane in such a way that it caters to all the points.

Model means we try to based on the observation based on different trial and error. We are trying to find out the best model is not it A base function that fits the data. So, if we draw if we find such an model which fitted all the points almost all definitely not all the points if it is almost all the points then what happens? In this case our training error will be very low and negligible when we get such a curve.

Training error will be very negligible but when we find this sort of curve because there may be some data's which are outliers. There may be some data which are very rare occurrences we are trying to what to say or trying to address those data also. When we do that then what happens? Our model is not the true it model does not mean the true model. In that case our testing error increases and see in over fitting our training error is very less because we have tried to fit almost all the points almost all the training points.

So, but then training error is very less but our testing error is high. This is the case of overfitting and this is the example of optimum fitting. Optimum fit is that where my training error is also low and testing error is also low that is we call it optimum fit. So, here but SVM is less prone to overfitting than other methods. As we missed definitely it definitely does not do under fitting but it is less prone to over fitting as well.

When I am talking of overfitting let me also tell you one thing. This kNN algorithm what we have seen again and if we take a very less value of k then chances are there that we do overfitting. For less value of k where the success of overfitting is there overfitting means training error is very less but testing error increases.
(refer time: 10:35)

So, now support vector machines the types of SVM linear SVM and non-linear SVM. Now we will be discussing linear SVM definitely will go to non-linear SVM also later. So, linear SVM a classification technique when training data is a linearly separable. This is an example of linear SVM by a straight line by a plane by a hyper plane we can separate the different datas. And this is a very good example of linearly non separable.

Here also we have two classes, one is this red circle another is the magenta blue squares and yellow curves yellow circles. But these are not linearly separable so we call that as non-linear SVM. So, now let us see what is maximum margin hyperplane.
(refer time: 11:27)

See we have talked about the distance boundary. When a data are described in terms of two attributes so two attributes means in distance ax + by the in when we describe a data in terms of two attributes then our distance boundary may will be a straight line. What will be the equation of the straight line? That is ax + by + c = 0 where c is the intercept a and b are the slope. So, ax + by + c = 0 and it has two attributes.

So, this is where distance decision boundaries may be a straight line. If by a straight line I can differentiate between the two classes. Now when there is data's are in three dimensional x 1, x 2 and x 3 or x y z when data is under three dimensional then we will use a plane to differentiate between the two categories. Then my differential boundary will be a plane entry dimensional similarly.

If my data's are in 4d, 5d, 6d so it is very difficult to just what to say picture that it is it then decision boundary will be hyper plane. So, for two attributes it will be a simple straight line, for three attribute it will be a plane, for more than three attribute will be a hyper plane. So, in general I will tell it is a hyper plane only hyperplane which is differentiating between the classes.
(refer time: 13:00)

So, now formulation of the classification problem. So, formally let us formulate it. So, in our subsequent differential assume a simplistic situation that given a training data with a set of n tuples, which belongs to two classes either plus or minus and each tuple is described by two attributes say A 1 and A 2. So, what is given to us some given a training data this is my training data t 1 to t n these are a different observation, t 1 is one observation, t 2 is another observation.

That means total I have an observation our n tuples which belongs to two classes this

observation. Suppose let us first simply stay for simple of discussion I am just considering that it belongs to two classes. The classes name I am giving is plus and minus and each tuple is described by two attributes A 1 and A 2. So, each has two attributes so two attributes we will be using a straight line to find out the two different classes to differentiate between the two classes.

So, for this current example we are considering that data is linearly separable so the current model is linear SVM model.
(refer time: 14:17)

So, now so here you see the figure, the figure three shows a plot of data in 2D. So, it is that we have considered the data has two attributes plus and each data has two attribute and there are two classes, classes are positive and negative. So, suppose these are the classes. Now these data are linearly separable so what happens? So, if this data is linearly separable so how linear is separable means we can separate it using a straight line.

So, now see if these are the data set actually then we can have different type of straight line to different to mark as a boundary decision boundary between these two point. This also can be one of my line, this line can also be one of the distance boundary, this can also be one of the distance boundaries, this can also be one of the distance boundary this can also be any number I cannot find indefinite in finite lines which can act as a decision boundary between these two categories.
(refer time: 15:20)

We can there are an infinite number of separating lines that can be drawn. So, distance boundary there can be infinite number of distance boundaries in the input and here are my distant boundaries and straight line. So, a straight line I can draw it just because to show that the; what is a set of data. Suppose this is my straight line, the set of data falls in different category and this set of data falls in different classes; that is all.

So, now since this is the case in this figure what you see then this figure there can be infinite lines. Now there are two questions whether all hyper planes are equivalent so, for the classification of data is concerned so there are infinite hyperplanes. Let me call it instead of line we are using a general term because data may be 2D, 3D, 4D, 5D anything, is not it? So, we are using a general term hyper plane.

So, if it is an considered and using this term hyperplane so now all this hyper planes are equivalent that means I can consider any of the hyperplane to the but to say to classify it into two classes if not which hyperplane is the best. If all the hyper planes are equivalent if not then which one is best which one should we take that is the two questions that arises.
(refer time: 16:38)

So, we may note that so for the classification error is concerned with training data all of them are with zero error. See if you take any of this line in this figure if you consider this line also this line this straight line if you consider this straight line also training error if I consider this error all

my pluses in this side all my minus in this, this side of this there is no training error. There is no minus here no plus this side so there is no training error.

So, if I consider this line also there is still no training error. If I consider this line also still not any of the line if I consider there is no training error. So, that means one from the training perspective any airline is equivalent. But let me see what happens when I talk of the test data. Suppose this is my test data. However, there is no guarantee that all hyper planes perform equally well on unseen data.

All these hyperplanes perform good in case of training data, I have seen here from this figure but all this hyperplane does it performs well in terms of the test data that we need to find out. Our requirement is since the training data also should be training error also should be low, testing error also should be low. Note that only training errors should be low testing error can be high or training error can be high testing or can below no both we need both low optimum.

We say I have seen optimum free it is that training error is also low testing error is also low. So, for this hyperplane training error is slow for all the cases. Now what about the testing error? You understand by meaning by what I mean by error, error is that training error means a data, a observation which is labelled as plus. When I use the function when I use this prediction model then I have given set of attributes are there and its class level is also given.

I know this class level is plus but I did not specify the class level. What I have done? I have just given the attributes and I have used this model to predict to which class it belong. When I have use this model to predict the switch classes belong for my data suppose I got class minus then that is a training error because I know this should be plus but it has gone to minus, this is in training error. So, similarly testing error is what?

Testing error is also same thing, testing error how about the class levels are not given there. So, based on the model you will have to find out the class to which class it belongs. Thus, for a good classifier it must choose one of the infinite number of hyperplane so that it performs better not only on training data but as well as on test data. So, we have to choose that hyperplane which performs well both on training data as well as test data.
(refer time: 19:42)

To illustrate how the different size of hyperplane influence the classification error consider any two hyperplane. In this figure you see to we will consider any two hyper plane, one there were many hyperplane in the last figure we have seen. Now we will just consider two hyperplane just one this is hyperplane one and another this is hyperplane two. Just two we have considered just two hyperplane to just to illustrate which hyper plane is better.

This to hyperplane H 1 and H 2 have their own boundaries say this hyperplane have their own boundaries. Now how do we find the boundaries? Now this H 1 which is the boundaries of H 1? H 1 boundary is this one b 11 and b 12. This dotted line you can see this dotted red dotted line; this is the boundary of H 1. Similarly, what is the boundary of H 2? Boundary of H 2 is this

dashed line black colour dashed line b 21 and b 22. This is the boundary of H 2.
(refer time: 20:48)

Decision boundary, how do you find out? It is a decision boundary is a boundary which is parallel to hyperplane and touches the closest class in one side of the hyperplane. So, I told for the hyperplane one H 1 my decision boundary is b 11 and b 12. For the hyperplane two my distance boundary is b 21 and b 22. Now how do I find the decision boundary? So, here it is. A decision boundary is a boundary which is parallel to the hyper plane and touches the closest class in one side of the hyper plane.

It is parallel to the hyperplane. So, H 1, this is H 1 so these are the two parallel lines and it touches the closest class. So, my decision boundary suppose this is my hyperplane. So, I will go on increasing my distance boundary till it touches one of the classes. Suppose here I got this I am growing in equal thing both the side. Suppose this side is starches class one plus that is class A. When it touches the class the distance between the hyperplane suppose in the middle it is the hyperplane distance between the hyperplane.

And my decision boundary suppose it is x then he decides also other side also it will be x it is the same distance. So, a descent boundary is a boundary which is parallel to hyperplane and touches the closest class in one side of the hyper plane. The distance between the two distant decision boundaries of a hyper plane is called a margin, this distance between two. So, you can see this m 1, m 1 is the margin of the hyperplane H 1 similarly m 2 is the margin of the hyperplane H 2, see in the figure.

So, if the data is classified using hyperplane H 1 then it is with larger margin than using hyperplane H 2. If I classify data with hyperplane H 1 and I have a larger margin and however if I classify with H 2, I have a lower margin m 2. The margin of the hyperplane implies the error in classifier. In other words, the larger the margin lower is the classification error because I have a great margin here.

So, senses of becoming an error is very less, this is my decision boundary, this is my hyperplane my class B starting from this point, you see here. This is my distance boundary my class B is starting from this point, class A starting from this point. So, I have a lot of gap in between so senses of getting an error given a test data senses of getting an error is very less. On the other hand, suppose this is my hyperplane.

My B started from here my A started from here my chances of becoming an error for a very small variation my B will become A, A will become B, senses of a classification error is very high. So, bigger the margin lesser is the classification error. So, that means they have a different hyper planes we have seen their infinite hyperplane I will look for that I will go for that hyperplane which has the highest margin. So, that my classification error is less.
(refer time: 24:16)

Intuitively the classifier that contains hyper planes with a small margin is more susceptible to

model overfitting and tend to classify with weak confidence on unseen data. Thus, during the training or learning phase the approach should be to search for the hyperplane with maximum margin that is and we will try to select that hyperplane which has the maximum margin. Then what happens? It is maximum margin it is very less susceptible to error.

Margin is lesser more susceptible to error. Our confidence in classification reduces if the margin is less. So, such a hyperplane with maximum margin, such a hyperplane is called maximum margin hyperplane. So, that is the concept of maximum margin hyperplane. Support vector machine means you have to you need to know what is maximum margin hyperplane. We may note that the shortest distance from a hyperplane to one of its decision boundary is equal to the shortest distance from the hyperplane to the decision boundary at its other side.

I told you it is equal both the sides. Alternatively, hyperplane is at the middle of its decision boundaries this is obvious I told you. This whatever is the distance from the hyperplane to the decision boundary whatever is the distance this distance is suppose this distance is X similarly hyperplane to decide also this distance will be x, hyperplane is in the middle of the decision boundaries. So, now linear SVM model. We will not be completing linear as we will model in today's class we will just start.
(refer time: 25:56)

So, SVM which is used to classify data which are linearly separable is called linear SVM data which we are linearly separable. Linearly separable means we can separate the data by using an hyperplane. But if it is too attribute by using a straight line, three attributes by using plane, more than three attribute by using a hyperplane then we call it a linearly separable. In other words, a linear SVM searches for a hyperplane with maximum margin obviously.

SVM means will always try to look for a hyperplane with a maximum margin. This is why linear SVM if often terms as maximal margin classifier. Linear SVM is also called maximum maximal margin classifier because we are looking for the maximum margin hyperplane. So, this is also called maximal margin classifier.
(refer time: 26:53)

So, now finding the maximum margin hyperplane MMH maximum margin hyperplane for linear SVM for a given trading data how do we do that. So, basically formal notation for putting it. Consider a binary classification problem consists of n training data now we are considering binary classification only for case of discussion. Consists of n training data notation of the training the how do we note it each tuple is denoted by X i Y i where x i's are different attribute corresponding to the attribute set of the ith tuple.

It is considering data is in a m dimensional and space that is why we have taken x i 1 x i 2 x i m there are total what to say how many observations we have taken. Suppose we are taken t observations where i = 1 to t and each observation has total m attributes. So, it is an m distance base and Y i is the class so Y i belongs to the; suppose there are two class plus and minus. It denotes its class level.

The choice of which class should be level as plus or minus is arbitrary which one is class plus which one is two classes minus is arbitrary. Given X i Y i for i = 1 to n we are to obtain a hyper plane which separates all the two sides of it which separates all into two sides of it. So, our goal is to find the hyperplane with maximum margin such that it separates all this data all this we have total n data, all this n data into both the sides of the hyper plane.

Some of the data will fall into one side, some of the data will fall into the other side of the hyper plane and this hyperplane should be with maximum margin that is our goal.
(refer time: 28:47)

So, before going to the general equation of a plane in n dimensional let us consider first a hyperplane in 2D plane. In 2D plane what will be a hyper plane? In 2D plane our hyper plane will be nothing just the equation of a straight line ax + by + c = 0 I am writing as w 0 + w 1 x 1 + w 2 x 2 where my attributes are x 1 and x 2, w 0 is the intercept where it is w 0 w 1, w 2 are nothing but the coefficients defining the slope and the intercept of the line.

So, my that means I have to find this line, my task in SVM is to finding this line what is this line w 0 + w 1 x 1 + w 2 x 2 = 0, I have to find this line. Finding this line means essentially, I have to find out the values of w 0, w 1, w 2. This line such that it has the maximum margin.
(refer time: 29:49)

So, this is my; the equation of the line. Now any point lying above such a hyper plane satisfies any point which is lying above this hyper plane point lying in this hyperplane is equals to zero. Any point lying above this hyperplane will be greater than zero any point which is lying below this hyper plane will be less than zero.
(refer time: 30:18)

So, an SVM hyperplane is an n dimensional generation of a straight line in 2D. It can be visualized as a plane surface in 3D but it is not easy to visualize when the dimensionality is greater than three of course. In fact, Euclidean equation of a hyperplane, we can write it in this way where i's are real numbers and b is a real constant called the intercept. Instead of w 0 I have written b you can write it in this way also which can be positive or negative.
(refer time: 30:44)

So, why I bought to this form basically. This particular equation I can represent it in matrix form. How can given a system of equation? How can we solve it is in matrix form? Same nothing so this is my equations w 1 x 1 + w 2 x 2 = b. So, if I write it in matrix form this will be my matrix from W X + b = 0. If you are getting confused you can also write W X + w 0 = 0. So, this is my matrix form where W is a vector having always value w 1 to w m different coefficients.

And axis again different attribute values and b is a real constant. So, here W and b are the parameters of the classifier model to be learned given a training data. Now does W and b we have to find out the parameter value of this W and b given as setup data given a set of

observations.
(refer time: 31:41)

So, consider 2D training data consists of two classes here the plus and minus. So, this is the line W X + b = 0, any point which is above the line is W X + b = 1. Since we have considered as plus one and minus one see here. Suppose b 1 and b 2 are two distant boundaries above and below the hyper plane respectively. So, it should be basically any points above this is W X + b greater than zero below this is W X + b less than zero.

So, consider any point x plus and x minus shown, suppose I have considered this point x plus and x minus here. Suppose this is my decision boundary, this dotted line. This dotted line is decision boundary suppose this dotted line the equation of this line is suppose W X + b = 1. And this distance boundary suppose W X + b = - 1 because this side it will be less than zero. Since W X + b = 0 and below this it will be less than zero. So, W X + b is to - 1 top side is W X + b = + 1, I am just assuming it is.
(refer time: 32:52)

So, now for x plus for any point x plus located above the decision boundary the equation can be written as W X + b = K where K is greater than zero. Any point which is above the hyperplane it will be greater than zero, is not it? So, W X + b I am writing a square where K is greater than zero. Any point which is below this point that is X bar I can write is W dot X bar + b it is K dash where K dash is less than zero.
(refer time: 33:21)

So, now basically thus if we label all plus as class label plus and all minus is class label minus then we can predict the class label y for any test data. How we can predict? Y is equal to plus if W X + b is greater than zero W will be minus if W X + b is less than zero. So, given any value of X what will I do? I will try to find out what is the value of W X + b. If the W X p+ b if I get greater than 0 then it will fall in plus class. If W X + b i find it is less than 0 then it will fall in the minus class.

So, that is how we will do the classification. Now but that will happen once I know what is W what is b once I know what is W, what is b that is the equation for my hyper plane with the maximum margin. Once I know the value of W and b then I can easily classify my data whether it falls in plus class or it falls in minus class. If it if W X + b is greater than 0 it falls in above the highway plane if W X + b is less than 0 it falls below the hyper plane.

Now how to find the margin of W X + b? How to find the margin of W X + b how to find out the value of W and b that will be discussing in the next lecture.
(refer time: 34:40)

So, in this lecture what we learned? We learned the basic concept of SVM and its advantage linear and non-linear SVM, the concept of maximum merge in hyperplane then the margin of maximum margin happened for linearly separable data in 2D. In the next lecture we will learn

more about MMH of a linear SVM.
(refer time: 35:04)

Thank you, guys.