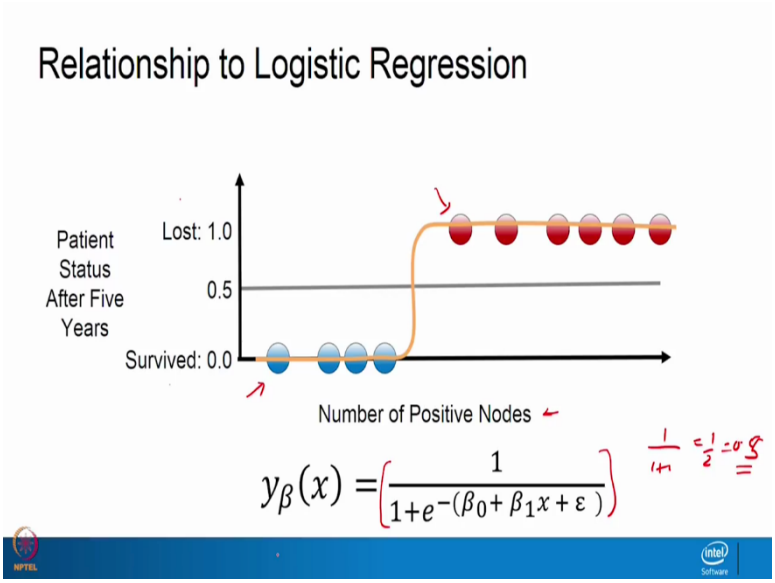
Machine Learning for Engineering and Science Applications
Professor Dr. Ganapathy Krishnamurthy
Department of Engineering Design
Indian Institute of Technology, Madras
Support Vector Machines

Hello and welcome back, in this video we will look at support vector machines. An introduction to support vector machines all the slides are provided by the Intel software.

(Refer Slide Time: 0:25)



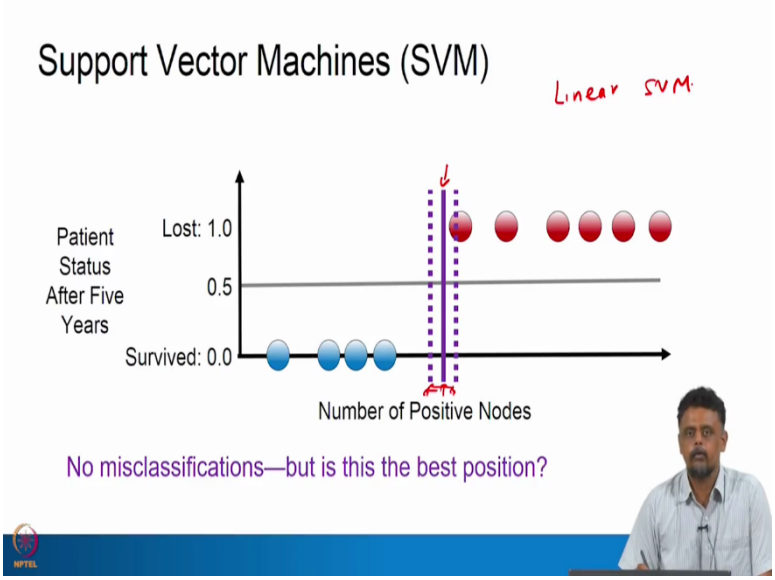
Let's first look at the relationship to logistic regression. So let's consider this example where we are trying to determine whether a patient survived or he was lost based on the number of cancerous modules from the patient, okay. Now if you use logistic regression then the idea is if the output of the logistic function is greater than 0.5 then we classify it as class I.

And if the output of the logistic function is less than 0.5 then we classified as class 0. Now if you recall the logistic function if that happens when the argument to the logistic function is greater than zero then we get 0.5 or above and if the argument to the logistic function is less than zero then we get a value which is less than 0.5, right?

Remember take this function then if the argument becomes zero then it is 1 over 1 plus 1 which is half that is like 0.5, right? If this evaluates let's say, if the argument becomes then this is point at which we draw the threshold, correct? So far all arguments which are greater than zero the logistic function outputs point greater than 0.5 and all arguments less than zero the point is that the logistic function outputs less than 0.5 though we classify that as a class I or plus 0.

So the idea behind fitting to a logistic Sigma is such that, for all class I's the argument is much much greater than zero and for all class zeroes the argument is much much less than zero, okay. So that then this function evaluates to some value much greater than 0.5 and we can with confidence say that it is class 1 or class 2, okay.

(Refer Slide Time: 2:15)



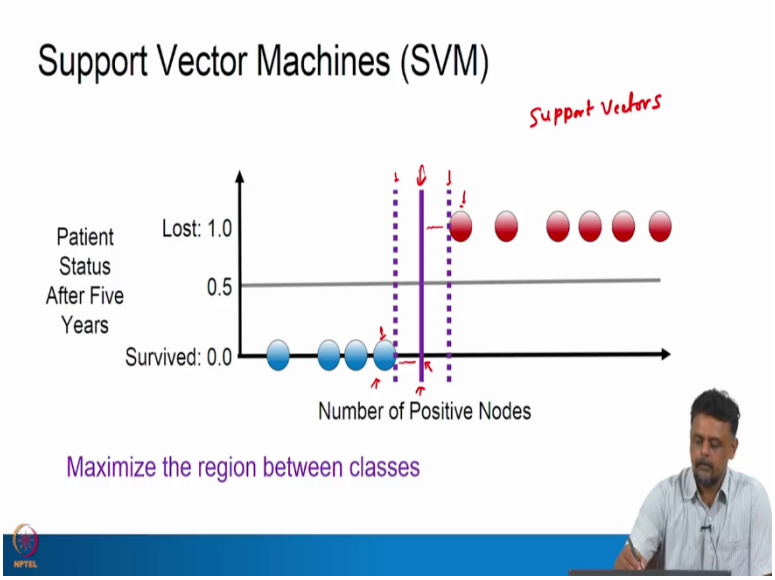
In other way of looking at this problem is to look at the, in the case of SVM it looks that terms of decision boundaries, right? So the idea is, okay let's say we draw a decision boundary here (())(2:24) only one feature we can just think of it like a threshold we draw the decision boundary given by this particular line, okay. Then what happens we have 3 misclassifications corresponding to this blue points are misclassified, so then we can just try another addition boundary which is over here, right?

Again we have 2 misclassifications corresponds to those 2 points, right there. However if we draw this line is more to some point in between , this one green, one blue and one red then we have no misclassifications. However there are multiple charges for this, right? We have an entire range for this and it is very difficult to determine where you can draw this line.

So the idea behind SVM is to figure out where to draw this line. So this is called the in 1D, so this is one straight line and of course in 2-D also we have a line more like a threshold in 1D and in 2-D we have a line. In multiple dimensions it corresponds to a separating hyper plate, okay. We will look at what this 2 dotted lines mean later on but this width typically is known as the margin that's what the and SVM is referred to as a maximum margin where by maximizing that margin. So we will see that later on.

However, the point behind support for vector machines is to figure out this boundary this separating plane line or boundary between 2 classes. And one of the criteria here is that the 2 classes should not have any overlap, so they should be linearly separable, so that is a problem we will be considering, we will be looking at linear SVM, okay. So that's the first intro introductory topic to SVM. So we will look at classes that can be linearly separated, okay.

(Refer Slide Time: 4:07)

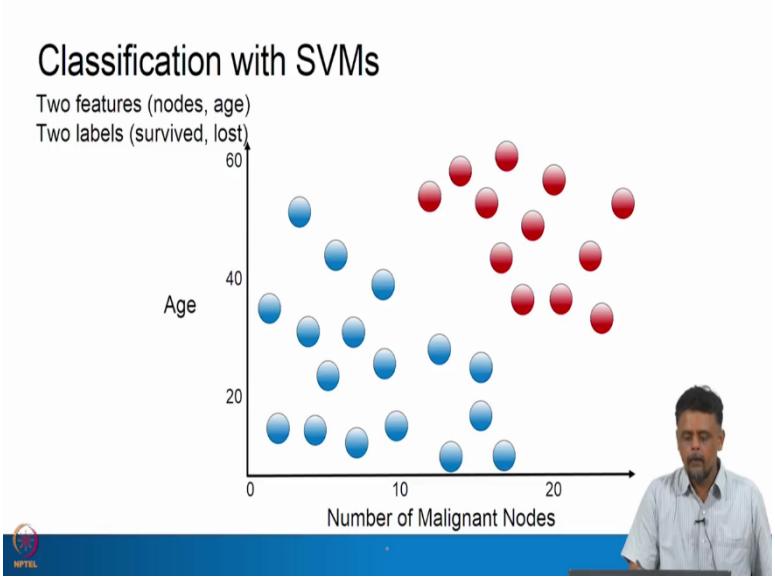


So ideally if we are coming back to a 1D example, ideally intuitively you can see that this is the ideal separating part, right? Because then we have some leeway here and here to and we also see that these 2 dotted lines they correspond to the points on either classes, so there is one in the red class, there is one in the Blue class and these 2 lines passed through points which are closest to these boundary, okay.

So you can see that these dotted lines pass are very nearby the red and this dotted line passes very nearby the blue dot and these 2 points are the closest to the optimal boundary that we have drawn here. This is the optimal boundary, okay. So then these points are referred to as support vectors, okay. I will show why they are called support vectors because we can just think of them as vectors in N dimensions, that's all.

It's a point in N dimension is, if you are dataset is N dimensional, okay. So that is the purpose of the objective of a support vector machine to find out this optimum boundary with respect to the support vectors which are basically the closest points of either class to separating boundary that we are looking at, okay or the separating plane we are looking at.

(Refer Slide Time: 5:33)



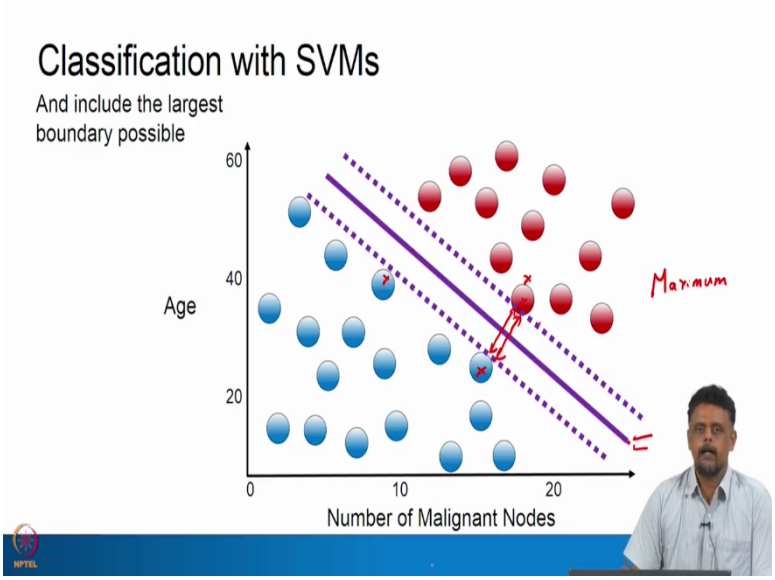
So we will consider slightly more complicated in the sense 2D that way we can actually illustrate it much better, we have 2 features number of malignant nodes cancerous nodes and the age of the patient. And let's say we are just trying to predict survival, so there are 2 types, the patient is lost or the patient survives which can be denoted by these 2 red and the blue dots.

And in the case of SVMs typically it as we will shift 0 to 1, 01 classification to plus minus 1, okay. So in logistic regression we saw it is class 0 or 1, in SVM is it is typically referred to as the class labels are minus 1 and plus 1, okay. So then how do we draw this line? Like you actually saw before for the 1D case. We can just draw many such separating lines, so as we go through we will see that each of them has a misclassification associated with it.

But the way to interpret this again one side of the line is class one the other side of the line is class minus 1, okay. So that's how we separate determine the classes, okay. So this is a very nice separating boundary much closer but you see it is also very sensitive. So suppose we have a Red point here. it's kind of dubious in the sense which it is very difficult to figure out which last it belongs to.

So again the similar way this line is also not the most best line because again it is very sensitive to small points very near that boundary.

(Refer Slide Time: 6:56)

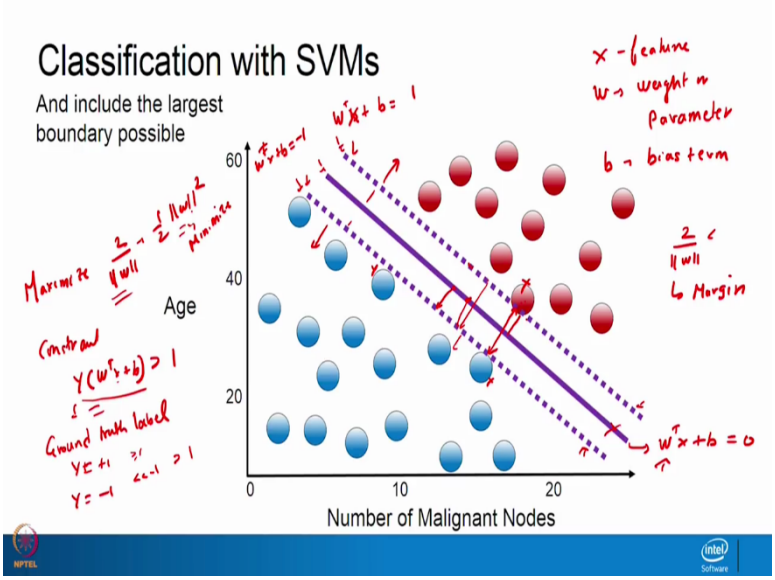


So ideally, so this is an excellent boundary, right? This is very nice because when it actually separates the 2 classes very clearly and then we have this margin, right? Once again if you look the margin that we talk about is the distance between these classes, okay. As determined by points (())(7:14) which are closest to the separating line in this case for separating plane, okay.

This is the optimal separating plane or separating line and these 2 lines are determined by the points that are closest to it, okay. So once again to reiterate the point behind SVM is to determine this line or plane in more in one direction it is a plane and more than 2 dimensions it's a plane and the idea is to determine this separating hyper plane based on the support vectors.

The support vectors are points on either classes plus and minus 1 classes that are closest to the, in terms of geometric distance they are closest to the plane, to the separating hyper plane, okay. And we use these 2 lines to, this is what we actually optimized for in the SVM algorithm that is we want to make sure that this distance is maximum, we. and this distance refers to as margin. So maximizing the margin, so as to determine the optimal separating hyper plane, so that the idea behind the algorithm for support vector machine.

(Refer Slide Time: 8:30)



So how do we go about doing this, so the idea is to what we did for logistic regression in terms of the model it is similar but we will exquisitely state intercept the biased, okay. So this you can think of it as an equation of a line or a plane in 3-D or hyper plane in multiple dimension more than 3 dimension. So we will call this as w transports x then we have the highest term w transports b equal to 0, okay. That is the equation of this line where the x is your feature and w is your weight vector or parameter vector, b is your term, right?

Okay. Now what we do is, we fix the distance between this line and the support vectors. So based on the support vectors we figure out these 2 equation of these 2 lines, we call them Yes transports x plus b equal to 1 and this line is w transports x plus b equal to minus 1, okay. So let's put them at unit distance from separating hyper plane. The idea is then what we talked about was that, we want to maximize is margin, okay.

So the maximizing this margin is then to figure out the distance from the optimal plane to let's say the closest support vector or you can think of some point on the plane to this line and some point on this again this plane or line to this line here. The sum of these 2 distances is called the margin and that's what we want to maximize. So how do you compute that distance? Then you please look up maybe from your high school.

Please look up how to determine distance of a point to a line, okay. So if you consider point from here to that line you can look it up, okay. And you will see that the total distance will come up to be 2 over w, okay. So this is the margin, okay. So you consider a point in this line

and you calculate the distance to this line and then consider a point on this line and calculate the distance to this line.

We know the equations of those lines and you take any arbitrary point you will see that it can be done, okay or the other way round also. Either way you should be able to calculate this distance to be true over (())(11:12) of w, okay. So this is the margin and we want to maximize this margin. Remember in the process when we maximize the margin with respect to w then you will get this value here, okay. Plus there is a constant, right?

So what have we done here? We have drawn these 2 supporting planes here such that those lines are drawn by considering the points that are closest to the optimal separating line or separating plane, okay. So then we are considering these points that's what we have used maybe on this one to draw these lines, these dotted lines, okay. So there should be no points in between these 2 lines, right?

So there should be no data points lying in between these 2. All class one points should be on the other side of this line all class minus 1 point should be on the other side of this line. So that constraint is written as Y times is greater than one, okay where Y is your ground tooth label. So the idea is, when Y is plus 1, you can think about it when Y is plus 1, Y equal to plus 1 then w transports X plus B should evaluate the sum number much greater than one hopefully and that means that if you multiply this to you will get a positive number greater than one.

Similarly when Y is minus 1 and w transports x plus B should evaluate to a number much less than minus 1 and this again a part of these 2 will give you a number much greater than one, okay. So that is the point behind having this constraint. So maximizing this quantity subject to this constraint or you can minimize the same, okay. So this is the loss function for support vector machines, okay.

So again to recap what we're trying to do is to, we are considering this classification problem where the classes are linearly separable. So that's an important constraint, okay. So the non-linear case is handled by something called a kernel trick, okay. But if time permits we will go there but otherwise we won't consider it but I will just stop it linear SVM.

So we are only considering linearly separable classes, okay that means that we should be able to draw line have a threshold or separating hyper plane between the classes if you are looking at more than 3 dimensions, okay. And one of the things to consider is, let's say we have this

line, we just considered 2D for us for our conversation here. Let's say we have this line we figure out this line.

What it means is that, if we consider that distance from this line to the nearest points on either class, okay. The geometric distance between this lines to the nearest points in either class that distance should be maximized, okay. That is the objective of the SVM function, okay. So when we do that then we get this, not only do you find out this hyper plane we also get the support, I call them support planes because these are the points that are closest to the optimization boundary are called the support vectors, okay.

So then if we have this trillions then it means that everything on the other side of these support lines are belonging to a particular class and actually there is no point in between them, okay. In the margin there are no training data points falling in the margin, okay. So this is typically used for binary classification plus or minus 1, if you have multiple classes then you do one against the rest, okay.

But typically SVM's are usually for binary classification and the way to formulate this, again like I said is to maximize imagine and to do that is that you define these 2 supporting planes which passed through the support that is by these equations w transports, I'm sorry have made a mistake here it is w transport x plus b equal to 1 and w transport x plus b equal to minus 1.

Basically you're putting them at kind of a unit distance if you say and we find out the distance between these 2 lines in terms of w which turns out to be nothing but 2 over models of w this is coordinate geometry you should try it out. And of course maximizing there would be same as minimizing, so this is minimizing half Omega square w square. Subject to the constraint that there are no points in between the supporting planes, okay.

And that constraint is met by this equation Y times, Y is the ground truth, Y times w transports x plus b should be greater than one, okay. So this is the objective function for SVM again this is called that quadratic programming problem and it is solved using landline multiply techniques we don't have time to go in there because of that again we have to go through the derivations to figure out how in the end the form of that of cross function will be such that we should be able to handle even non-linear stats boundaries.

Of course if time permits we will just look at it otherwise we will refer it for later class or maybe I will just post you some resources to read from, okay.