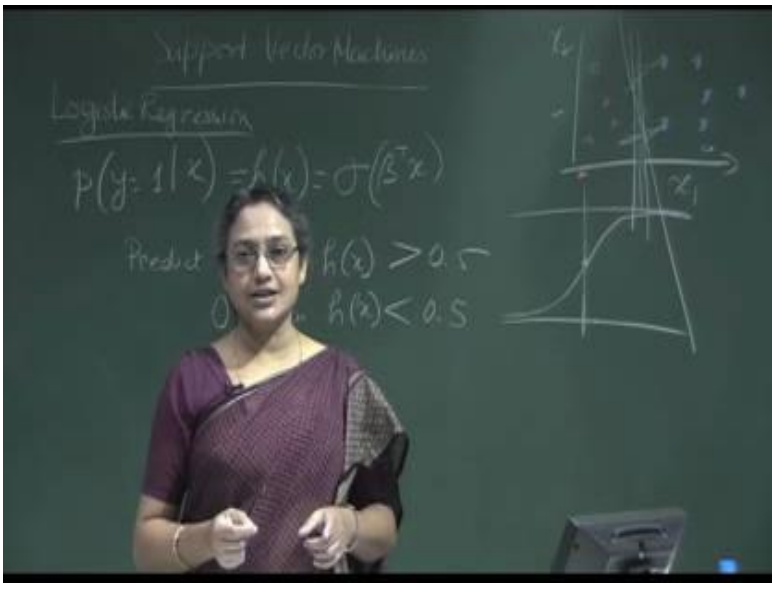
Introduction to Machine Learning
Prof. Sudeshna Sarkar
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Module - 5
Lecture - 21
Introduction Support Vector Machine

Good morning, so we start the second part of the lecture of this week. Today, we will give an introduction to support vector machine. We have studied several classifiers, and they have different properties. Support vector machine is one of the most effective classifiers among those, which has sort of linear. It has a very you know good mathematical intuition behind the support vector machine and we are able to handle certain cases where there is non-linearity by using non-linear basis functions or in particular we will see these are called kernel functions.

Now why support vector machine is so popular, we will see that support vector machines have a clever way to prevent over fitting. And we can work with relatively larger number of features without requiring too much computation.

(Refer Slide Time: 01:31)



So, in order to motivate support vector machine, let us look back at logistic regression, which I talked about in the last class. In logistic regression, we have try to find out the probability of y equal to 1 given the input x. And we have model this h x as the logistic function sigma or g of beta T x. Now we predict one when this h x is greater than 0.5, so we predict one when h x is greater than 0.5 and 0 when h x is less than 0.5. And the probability value is higher probability of confidence in the output is higher when h x is much larger than 0.5, because we are using this sigmoid function which is s shaped you know the function has a shape like this, so at this point h x is 0.5.

But as we go towards the side h x has value closer to one our probability or confidence in the output would be higher. So, the larger the value of h x or this smaller the value of h x higher is our confidence in the output. So, we have more confident on the predictions which have further from the decision surface. So, near the decision surface, so logistic regression actually gives you linear classifier.

Let us say just look at, suppose these are your features x 1, x 2. And you how different point so suppose these are you are positive points, these are you are positive points, and let us say these are you are negative points. Now suppose this is you are decision surface now if this is you are decision surface, then you have less confidence if these is a point here, you have less confidence in you are output of a point, which is closer to the decision surface. You have a higher confidence for those points, which you are further from the decision surface. If you have a large number of features, you could use Bayesian learning.

In Bayesian learning, we will look at all possible classifiers, wait them by their aposteriori probability and combined them to get the answer. And as we have discussed Bayesian learning is computationally intractable. We want to a good alternative to Bayesian learning, which is computationally attractive. So, this brings us to support vector machines.

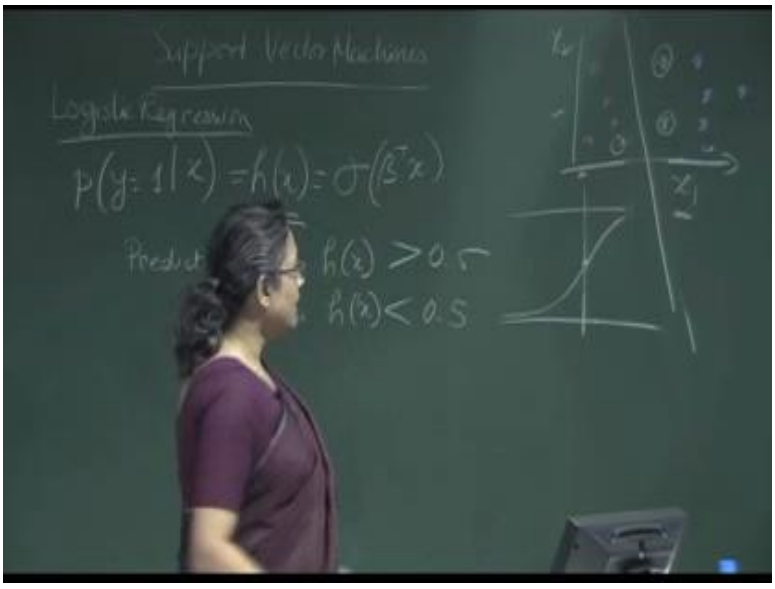
In support vector machine, we want a classifier, so that it maximizes the separation of between the points and the decision surface. So, for that we can define, what we mean by the margin so or you know if you look at this point, we can find the distance of this point

from the decision surface. We can find the distance of this point from the decision surface, this point, this point etcetera for each point; we can find the distance from the decision surface.

Now, if you look at the negative points, this is the closest negative point to the decision surface. In fact, this is the closest point over all to the decision surface and this is the distance of the closest point to the decision surface can be defined as the margin. And these two points these two positive points are closes from the decision surface.

Now we could have shifted this line here so that the distance from the closes positive points and the closes negative point is the same. In that case the margin width will be slightly large, so the width of this band will be twice this. But there are other decision surfaces for which there can be other margins, and among all possible decision surfaces we want to choose that one for which the margin width is highest. So my margin is denoted by the minimum distance of a training instance from the decision surface, and we want to choose that decision surface for which the margin width is highest. All this is under the assumption that the positive and the negative points are linearly separable by a linear decision surface.

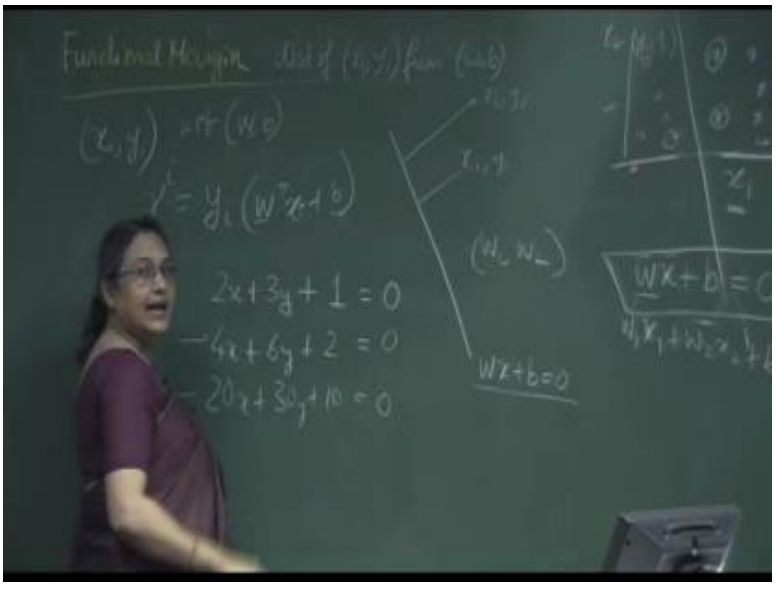
(Refer Slide Time: 07:14)



Now, you know if you consider a particular decision surface, and then you considered the points, which are closes to the decision surface. Now these three points are closes to the decision surface and then it create distance from the decision surface these points are called the support vectors. So, in the case where you are maximizing the margin width, the distance of the closest negative point to this line and the closest positive point to this line will be same.

And there are minimum of two support vectors, but there can be more than two support vectors, but typically the number of support vectors is extremely small. And if you notice this support vectors are the one which actually determine by equation of the line, and typically the support vector number will be very small. And this particular decision surface decides which point given a test point, you can do the test which side of the decision surface, it is based on that you can classify the points has positive or negative. Now, with this introduction let us define what we mean by formally by margin.

(Refer Slide Time: 08:49)



So, first will talk about what we call functional margin. Now functional margin of a point you take a point x i, y i is an arbitrary point; suppose this is a point x i, y i. So, this point with respect to a decision surface, now suppose the equation of this decision surface is given by w x plus b equal to 0. So W is a vector, x is a vector.

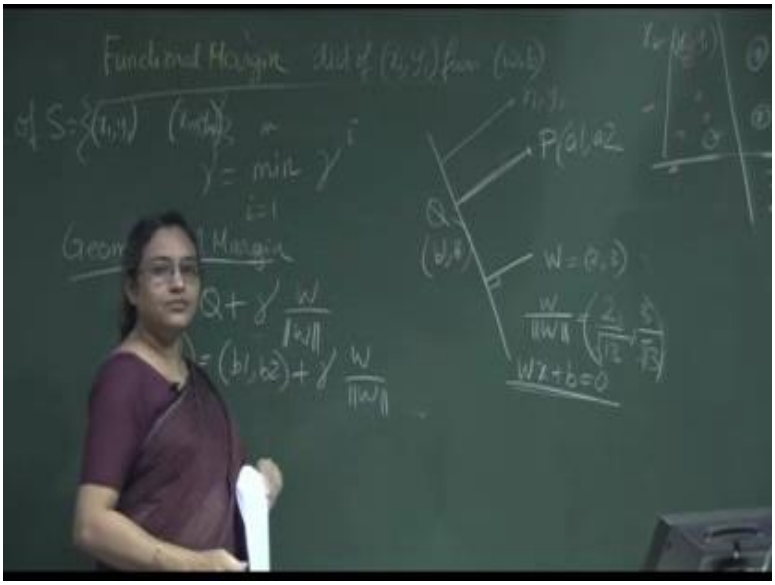
Suppose x 1 and x 2 are the two features, so this is W 1 x 1 plus W 2 x 2 plus b equal to 0, this is the equation of the line, and it can be written as W x plus b equal to 0. So, with respect to this line W x plus b equal to 0 which is parameterized by W b the functional margin is measure by the distance of the point x i, y i from the decision boundary. So, if the distance of x i, y i from the decision boundary, it is given by W b.

So, now let us see how to find this distance. Suppose, this is a straight line, whose equation is given by W x plus b equal to 0. Now the vector which is normal to this line has the normal to this line has, so this vector W 1, W 2 is normal to this straight-line. So, we can define the distance of x i, y i from W b as let us say gamma i and gamma i equal to y i times W 2 x i plus b. So, gamma I will be given by y i into W t x i plus b which is the normal distance from this point x i, y i to the decision surface, so this is given by this.

Now, this functional margin you know among another if I take another point, let us say x j y j, it will have another function margin. So, the functional margin of this is more than the functional margin of this. So, if the point is here, further away from the decision surface, we have higher confidence in the classification of this point. So, larger functional margin means more confidence in predicting the class of that point.

The problem of defining the margin as this distance is as follows. This W 1 b can be arbitrarily scaled. Suppose, the equation of this line is 2 x plus 3 y plus 1 equal to 0, so w is 2 3, D is 1. Now I can write the same equation as 4 x plus 6 y plus 2 equal to 0; and for that matter as 20 x plus 30 y plus 10 equal to 0. So, if we scale this equation the equation of the line remain same, but the functional margin becomes larger. So, the functional margin does not depend on the actual line, it also depends on the coefficients that we are using. Therefore, we need to use some normalization, so that we can look at normalized distance, we will come to that presently. But before that, we have seen what is the functional margin of a point at one point from the line.

(Refer Slide Time: 13:37)



But let us define the functional margin of set of points so we have a set of points x 1 y one x 2 y 2 x n y n. So, we have a set of points x 1 y 1 x 2 y 2 x m or rather let us write x m, y m these are my training points. And with respect to them, the functional margin is defined to be gamma equal to minimum i equal to 1 to m gamma i. So, among all the functional margins, I want to choose the smallest functional margin that is a smallest functional margin of the different points is the find to be the functional margin of the set of points.

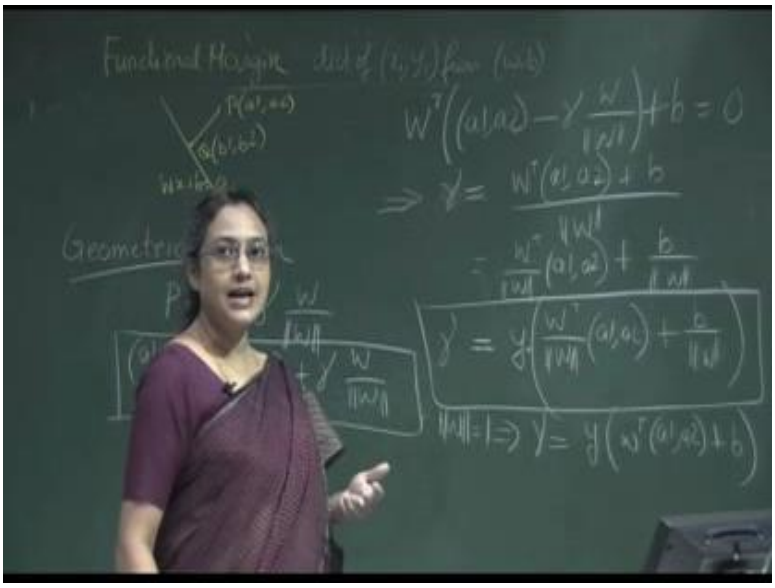
Now, let us come to as I said definition of a functional margin suffers from these drawbacks. So, we want to use form which does not have this drawback and we come to the concept of geometrical margin, which is invariant to the scaling of the equation. Now again let us consider the decision surface given by W x plus b equal to 0. And let us look at a normal to this decision surface. So, the normal to this decision surface is given by W - the vector w is normal to this decision surface.

And the unit vector normal to the decision surface, so this is the vector w and the unit vector normal to the decision surface is given by W divided by W norm. So, if the equation of the line is 2 x plus 3 y plus 1 equal to 0, so W equal to 2 3. And W by W hat

is 2 by root over 2 square 3 square that is root over 13 3 by root over 13, so this is the unit vector, which is normal to the equation normal to this decision surface.

Now, let us look at this point p. So, suppose p has a value a, b. And if you draw normal to the decision surface suppose it needs a decision surface at q. And let us say that this has this is a 1, a 2; and q has the coordinates b 1, b 2. Now if you want to find the distance of p from q, the distance is in the direction of the normal vector. So, we can write p equal to q plus gamma times W by W hat, W by W hat is the vector which is normal to this. So, the coordinates of Q plus if gamma is the distance of p from this decision surface the coordinates of P is given by coordinate of Q plus gamma times W by W hat that is we can write other way round a 1 a 2 equal to b 1 b 2 plus gamma times w by w hat. And given w we can write this equation and from that we can find gamma. Let me just rub this; please keep this basic diagram in mind.

(Refer Slide Time: 18:44)

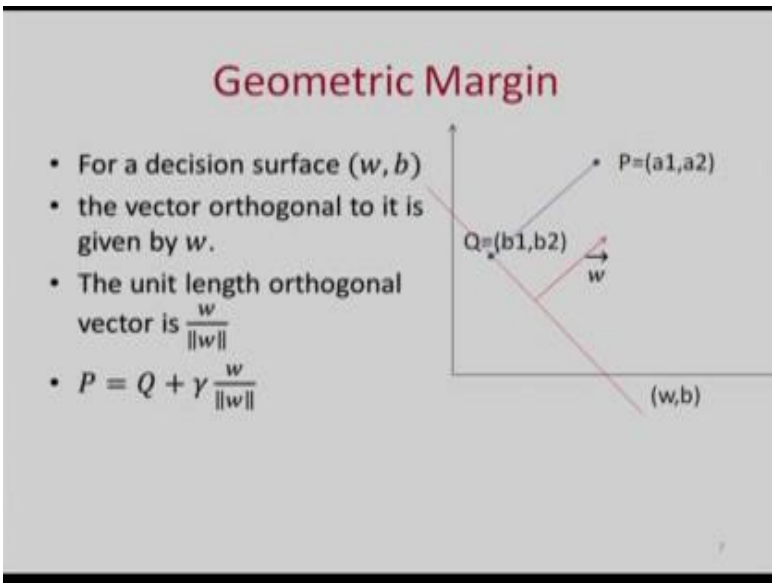


So, from this, we can find out that w transports a 1 a 2 minus gamma w by w hat plus b equal to c, because this point b 1 b 2. b 1 b 2, if you just recall, let me just draw the diagram again. So, this is the decision surface this is p - a 1, a 2; and this is q given b1, b 2. So, q given b 1 b 2 lies on the line W x plus b equal to 0, so this is b 1 b 2, so it lies on this line, so W into this plus b equal to 0. So, from this, we get W equal to we can solve

for gamma and we can find gamma equal to W t a 1 a 2 plus b by W hat, which is equal to W t by W hat a 1 a 2 plus b by W norm. So, from this, we find gamma equal to y times W by W hat a 1 a 2 plus b by W y is plus 1 or minus 1.

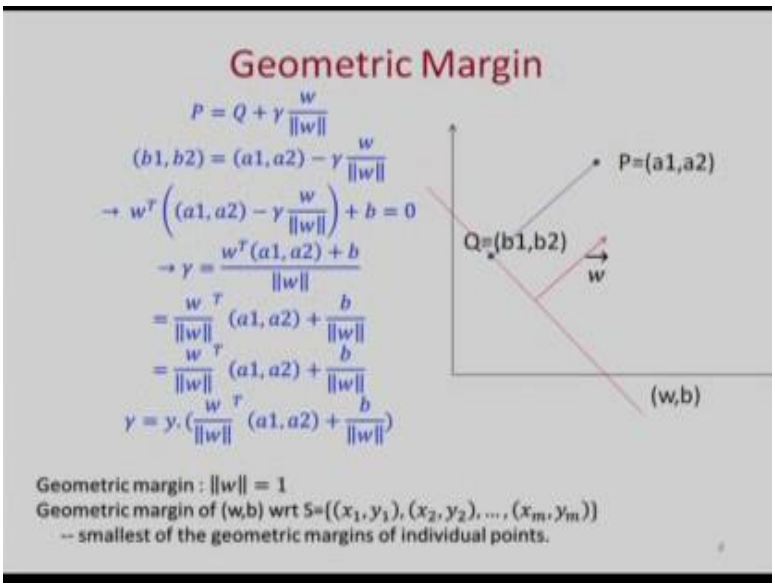
So, for geometric margin, what we will do is that we will scale W, you know w is the weights of the line W b, we can scale w b arbitrarily by dividing all of them by some same number or multiplying them by the same number, we will scale W so that w norm equal to 1. We will scale w so that w norm equal to 1, and then we will find the geometric margin. So, geometric margin, so will scale W, so that w hat equal to 1 and then the geometric margin will be given by gamma equal to y by y times W t a 1 a 2 plus b. So, this is the geometric margin, which we get after normalization.

(Refer Slide Time: 22:22)

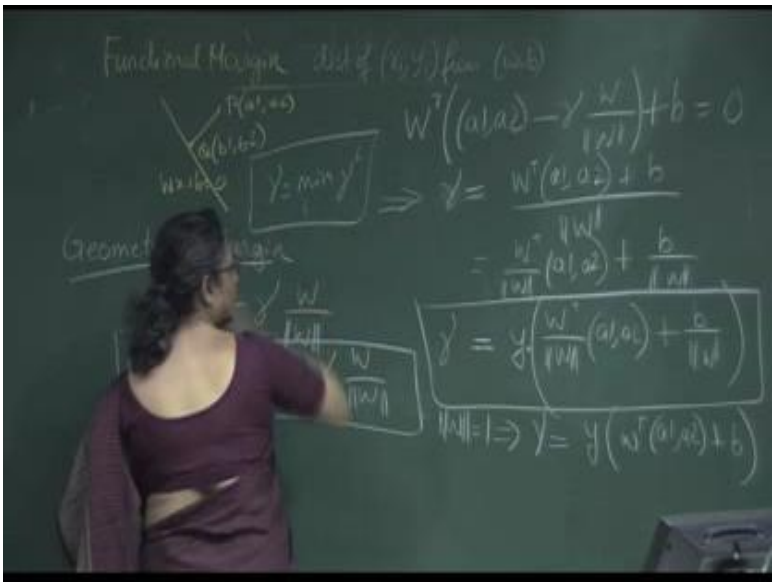


And as in the previous case if you have a set of points x 1, y 1, x 2, y 2, x m, y m, we can find out the geometric margin as the one which is smallest.

(Refer Slide Time: 22:25)

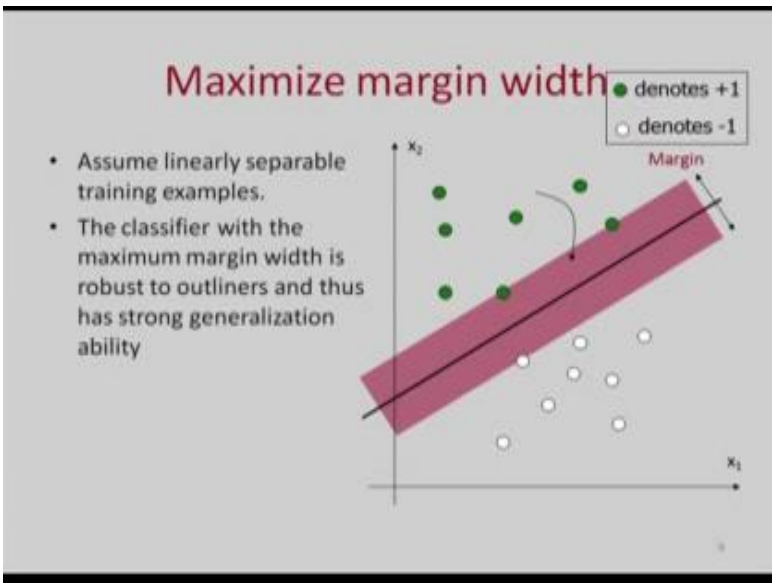


(Refer Slide Time: 22:31)



So, geometric margin will be as before the minimum of i of gamma i that will be the geometric margin of set of points.

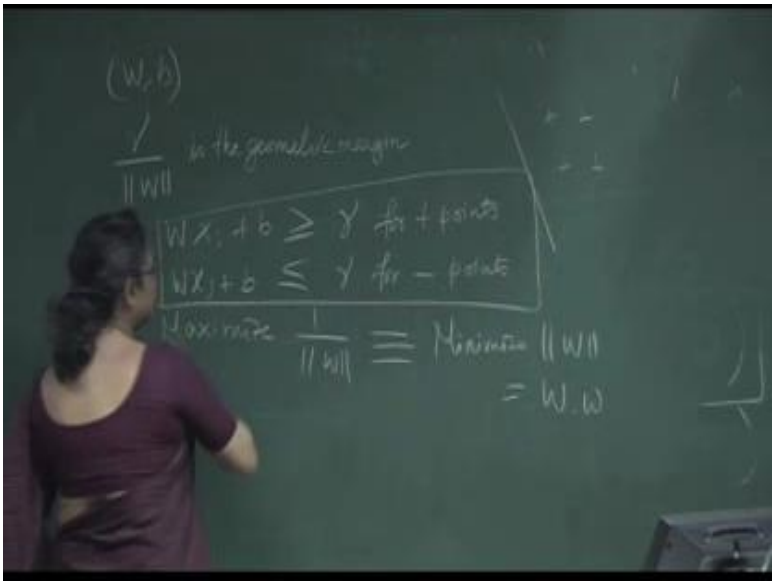
(Refer Slide Time: 22:51)



Now, we will look at how what we really want to do. If we look at this diagram, we assume that the training examples are linearly separable, and let us say this line is our decision surface, and this red band is the margin. And the green points here these two green points lie on the margin this white point also lie on the margin, these are the support vectors. So, classify with the maximum margin width is what we want it is robust to outliers and it has strong generalization ability.

Now once we have defined this we have seen that gamma is our geometric margin, and we want to maximize this margin. You know if we without normalization, we get if you have a gamma by norm of W is the geometric margin. And we need to maximize gamma by W subject to constrain. So, we pose the our optimization problem as follows given a set of training examples labeled as positive and negative.

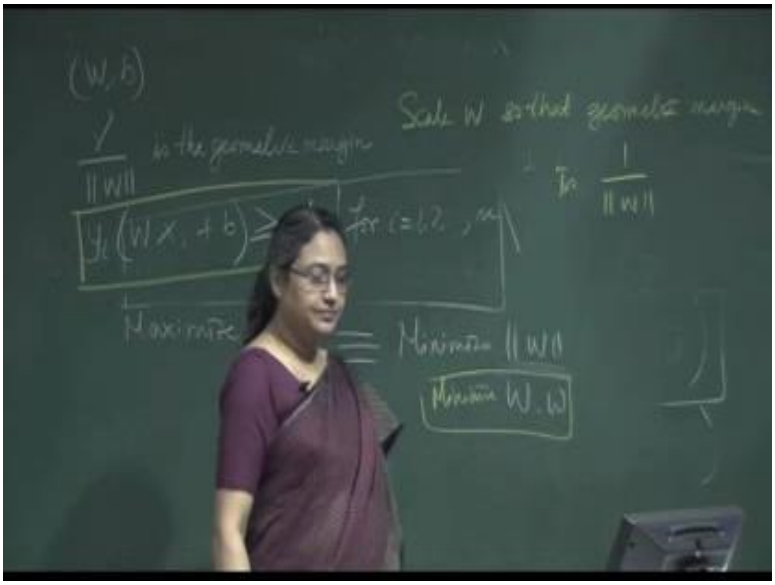
(Refer Slide Time: 24:18)



If W b characterizes the decision surface then gamma by W is the geometric margin. And we want to learn the values of W b, so that this geometric margin is largest subject to constrain. What are the constrains, we have the positive points on this side, for each positive point, W x plus b will be greater than equal to so all positive points will be at a distance of greater than equal to gamma from the margin. So, W x i plus b will be greater than equal to gamma for plus points; and W x j plus b will be less than equal to gamma for minus points, so and gamma by W is what we want to maximize.

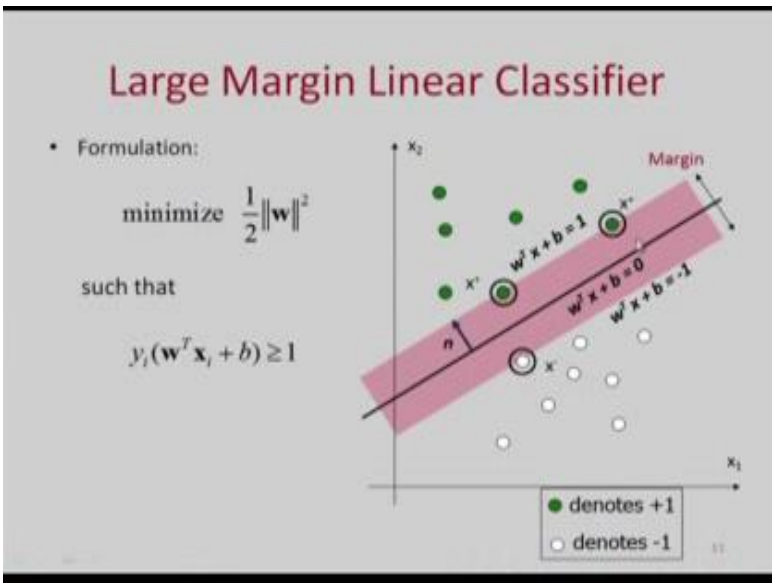
Now, if we so we can say we can maximize so we want to scales, so that so that this gamma is 1. We can scale so that this gamma is 1, and we can say maximize 1 by W, which is the same as equivalent to minimize W, and W is w dot w or w transpose w, so this is the w dot w dot product. So, we want to minimize W dot W, subject to this constrains. And now these two constrains can be combined and written in the same form as y i times W x i plus b greater than equal to gamma for all points.

(Refer Slide Time: 26:45)



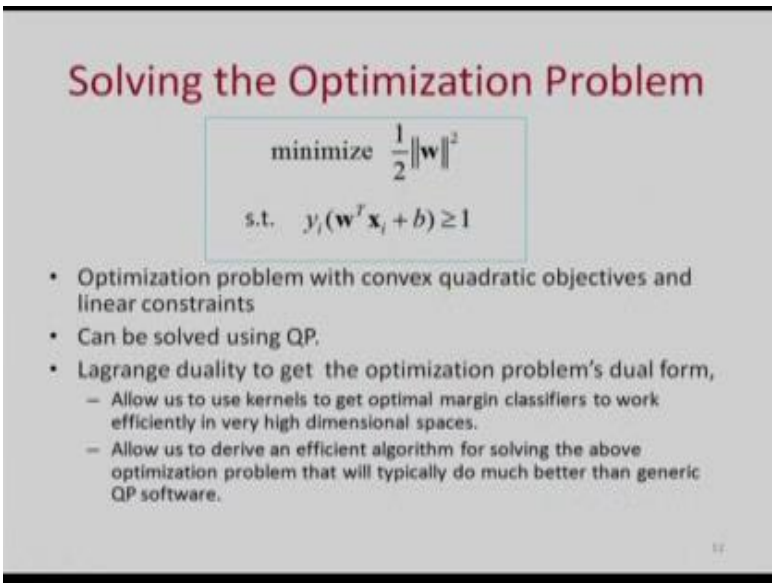
And we will normalize the values of W, so that actually I was wrong we do not normalize w to be 1, we keep w we normalize so that gamma equal to 1. So, the geometric margin is gamma by W, we will change W, so that so we will scale W for normalization this is what we will do, please pay careful attention. We will scale W, so that geometric margin is 1 by W, so we scale it so that geometric margin is 1 by W. And we want to maximize this margin subject to this constrains y i, W x i plus b greater than equal to gamma which will become greater than equal to 1, for all training instances right and equivalently we can say minimize W or minimize W dot W, so minimize W dot W.

(Refer Slide Time: 28:35)



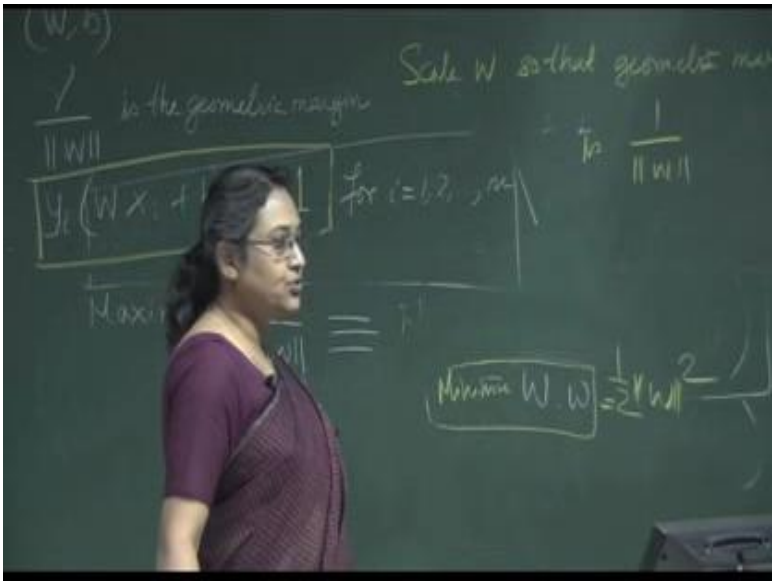
So, if you look at the slide now so this is the formulation of our optimization problem we want to minimize W dot W. Or we can say minimize W square that is W 1 square plus W 2 square minimize W square, so that these constrains y i times W x i plus b is greater than equal to 1. What is translates to is that suppose this is you are decision surface which has the equation W x plus b equal to 0. And the line parallel to it and this side is w x plus b equal to minus 1, the line parallel to this, and this side is W x plus b equal to 1. And the margin has been scaled, so that the geometric margin has width 1, and the positive points on the margin have the satisfy the equation W x plus b equal to 1, the negative points on the margin satisfy the equation W x plus b equal to minus 1.

(Refer Slide Time: 29:40)



So, how to solve this optimization problem this is a quadratic programming problem. We have a quadratic objective function, and we have convex quadratic objectives. We have convex quadratic objective function, and we have a set of m constrains m inequality constrains which are linear. So, this optimization problem can be solved using a commercial quadratic programming solver. So, by solving this, we get our classifier and that should be the end of the story. However, there are few other things that we can do by this very powerful formulation and this is what we will discuss now.

(Refer Slide Time: 30:48)



Now, we can straight away solve this problem using, so we are we are minimizing W dot W or W square. And we are we can also say minimize half W square this half is a so that we have a nice form, when we get the solution we could have solve this problem straight away, but we are going to covert this problem into dual formulation. And we will see that the dual formulation have certain properties.

Now, if you want to solve this optimization problem the method to solve this is by using Lagrange multiplier, you can use Lagrangian, and you can get a formulation of this problem, which you can solve and you can get the Lagrange duality to get the dual of this optimization problem. And what we gain by using this is that we can you know this formulation has some nice properties which we will are we able to see later.

With this, we come to the end of this introductory part of the support vector machine. In the next class, we will talk about the dual formulation.

Thank you.