Neural Networks for Signal Processing-I

Prof. Shayan Srinivasa Garani

Department of Electronic System Engineering

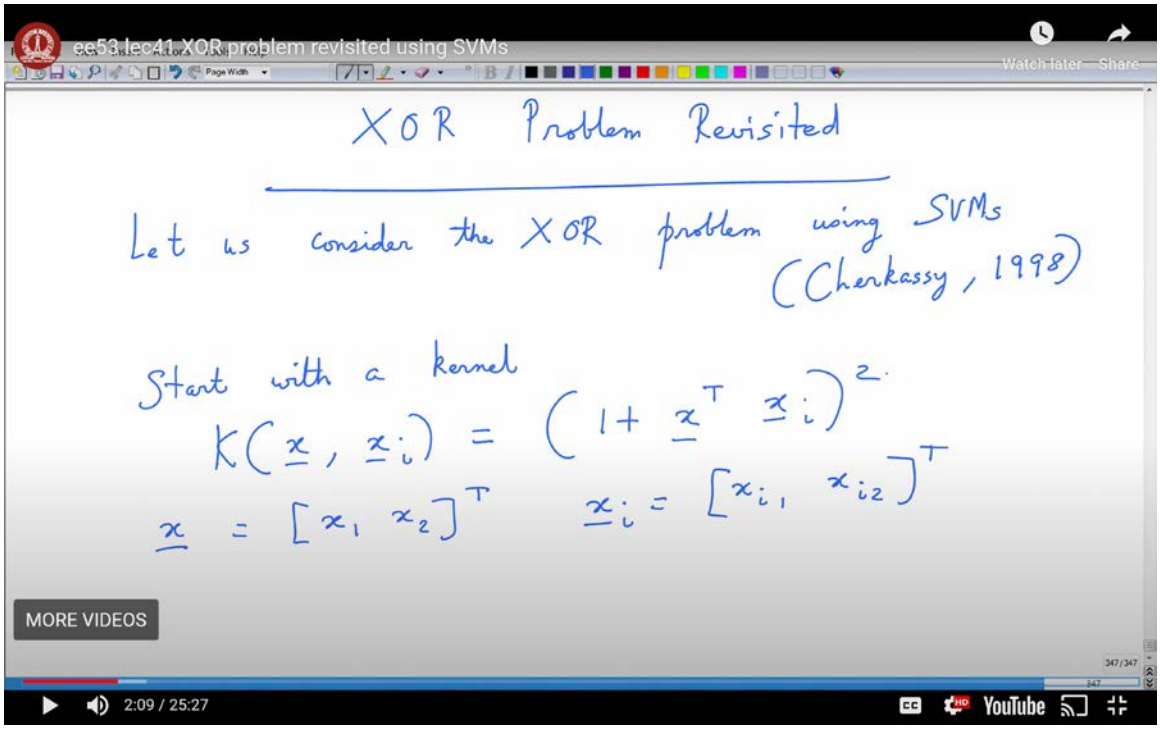
Indian Institute of Science, Bengaluru

Lecture – 41

XOR Problem Revisited using SVMs

Let us delve into the XOR problem, which we've previously explored through multilayer perceptrons and radial basis function networks. Now, we will examine this problem from the perspective of Support Vector Machines (SVMs), as studied by Cherkassy in 1998.

(Refer Slide Time: 02:09)



To start, we use a polynomial kernel, defined as:

$$k(x, x_i) = (1 + x^T x_i)^2$$

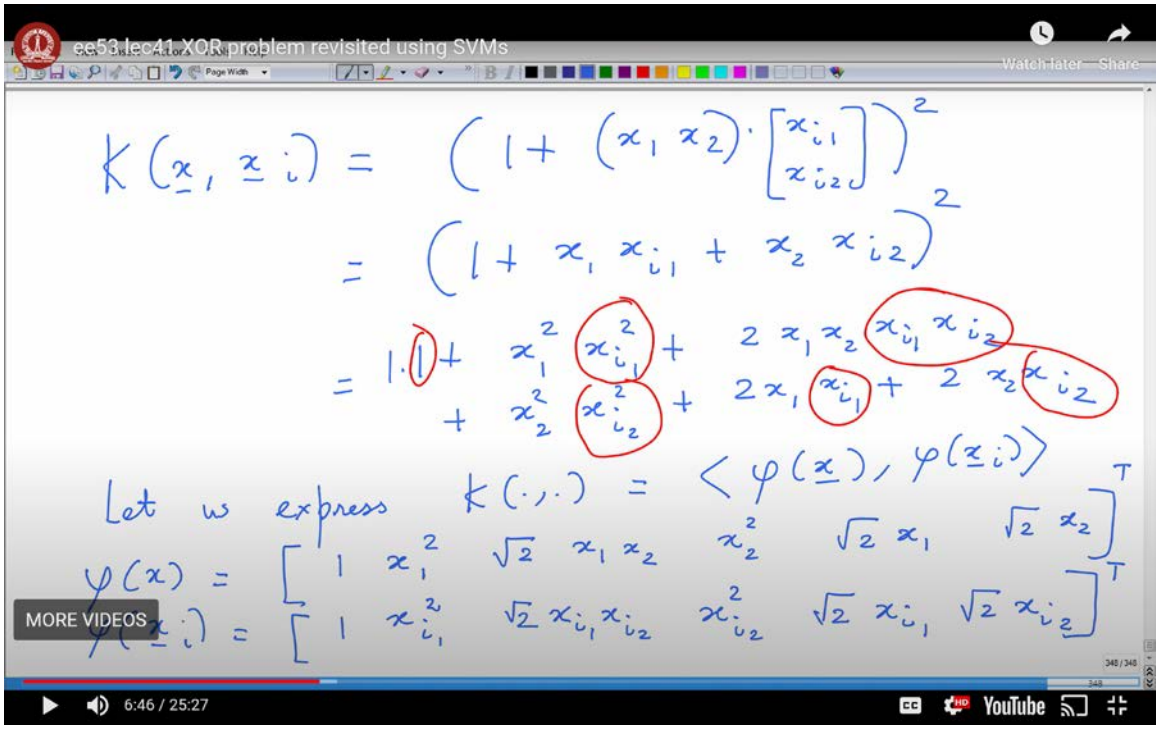
Assuming x and x_i are vectors in a two-dimensional space, where $x = (x_1, x_2)^T$ and $x_i = (x_{i1}, x_{i2})^T$, we can compute this kernel as follows:

$$k(x, x_i) = [1 + (x_1 x_{i1} + x_2 x_{i2})]^2$$

Expanding this expression, we get:

$$k(x, x_i) = 1 + 2(x_1 x_{i1} + x_2 x_{i2}) + (x_1 x_{i1} + x_2 x_{i2})^2$$

(Refer Slide Time: 06:46)



Breaking down the squared term:

$$(x_1 x_{i1} + x_2 x_{i2})^2 = x_1^2 x_{i1}^2 + 2x_1 x_{i1} x_2 x_{i2} + x_2^2 x_{i2}^2$$

Thus, combining all terms, we have:

$$k(x, x_i) = 1 + x_1^2 x_{i1}^2 + x_2^2 x_{i2}^2 + 2x_1 x_{i1} x_2 x_{i2} + 2x_1 x_{i1} + 2x_2 x_{i2}$$

In total, this expansion results in six terms. This can be interpreted as the inner product of two vectors, φ(x) and φ(x_i), where:

$$\phi(x) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2\right)^T$$

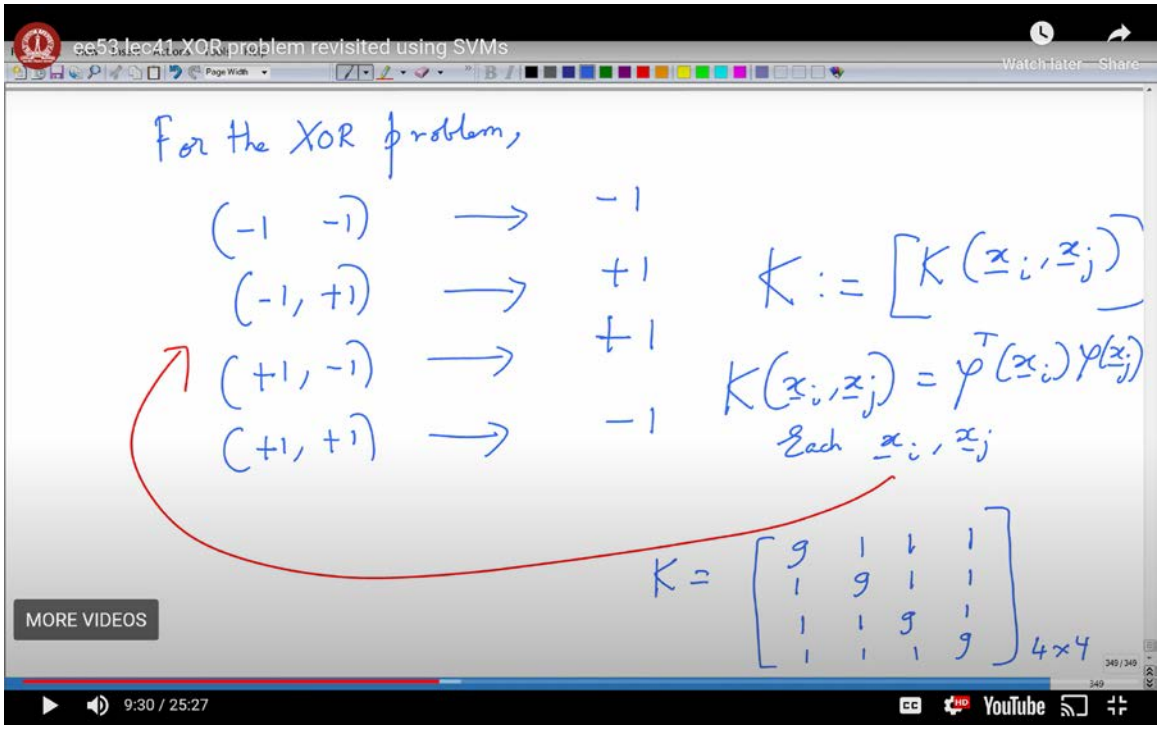
Similarly,

$$\phi(x_i) = \left(1, \sqrt{2}x_{i1}, \sqrt{2}x_{i2}, x_{i1}^2, x_{i2}^2, \sqrt{2}x_{i1}x_{i2}\right)^T$$

The inner product between φ(x) and φ(x_i) gives us the kernel function:

$$\phi(x)^T\phi(x_i) = k(x, x_i)$$

(Refer Slide Time: 09:30)



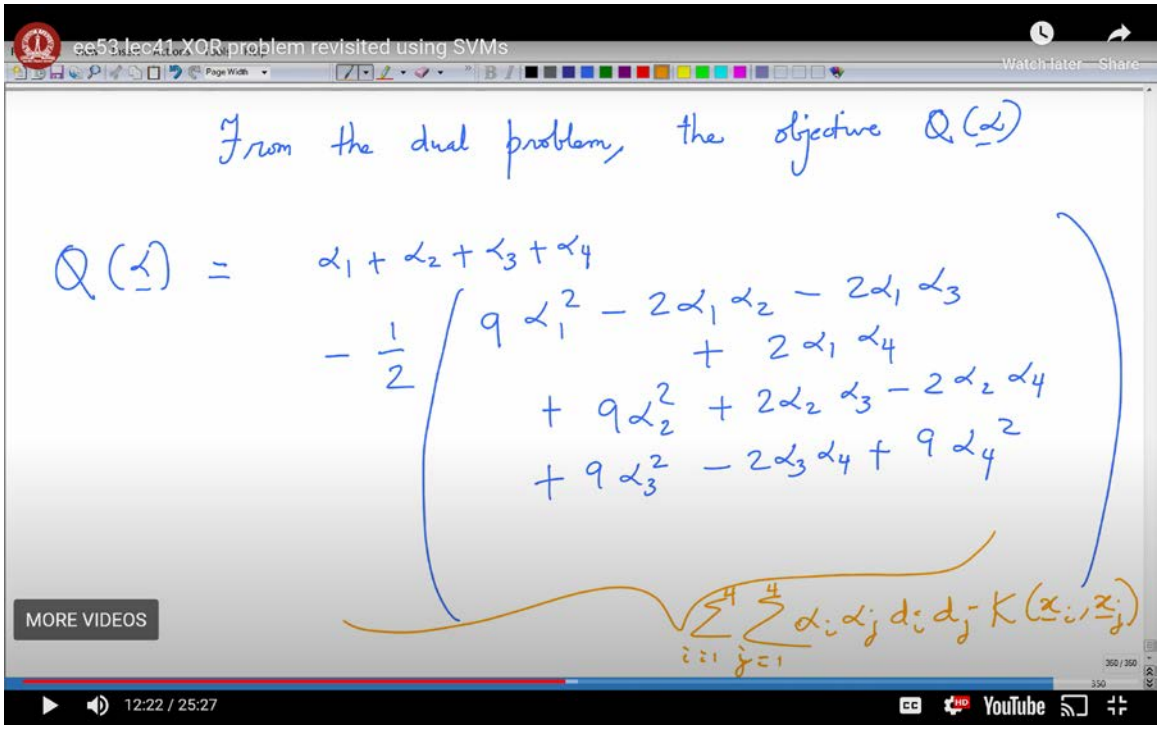
For the XOR problem, the data points and their corresponding responses are:

- (-1, -1) corresponds to a response of -1
- (-1, +1) corresponds to a response of +1
- (+1, -1) corresponds to a response of +1
- (+1, +1) corresponds to a response of -1

This kernel transformation allows us to map the XOR problem into a higher-dimensional space, where it becomes linearly separable, facilitating the use of SVMs to solve it.

Let us now formulate the Gram matrix K, where each element is given by the kernel function $k(x_i, x_j)$, which equals $\phi(x_i)^T \phi(x_j)$. For the XOR problem, the vectors x_i and x_j can be the pairs (-1, -1), (-1, +1), (+1, -1), and (+1, +1). This will result in a 4×4 matrix when we compute the Gram matrix.

(Refer Slide Time: 12:22)



Evaluating $\varphi(x_i)$ and $\varphi(x_j)$ for each pair of points, we obtain the following Gram matrix:

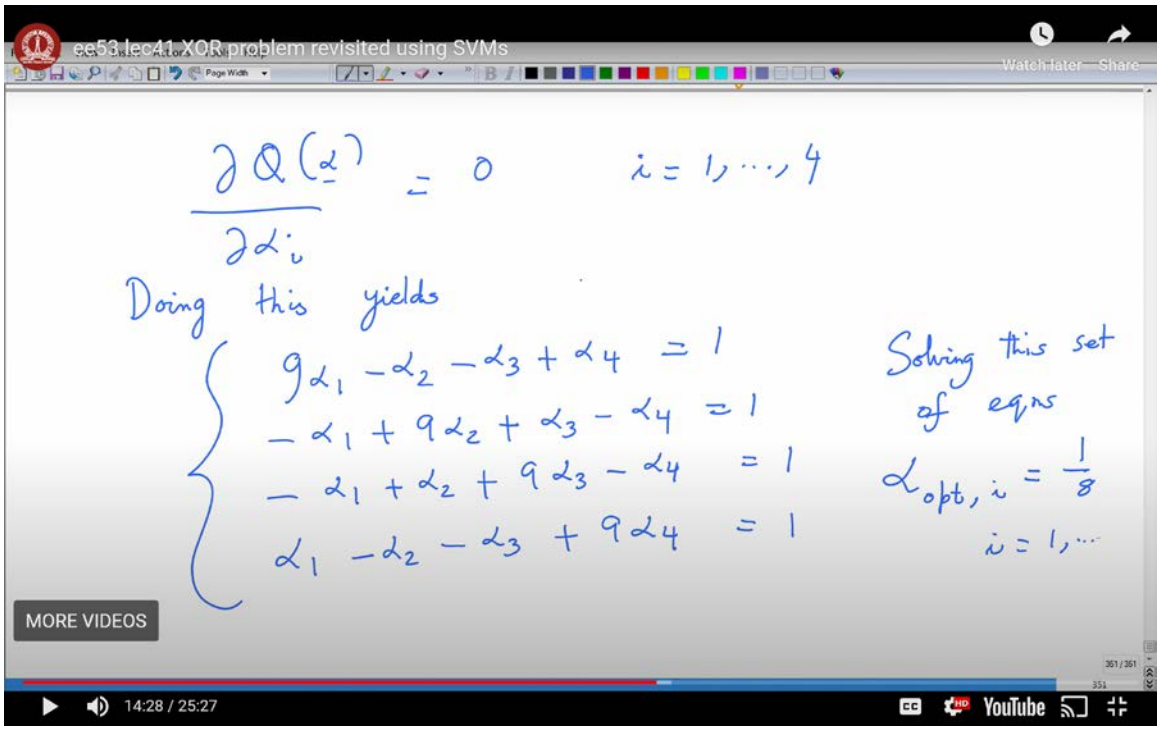
$$K = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 9 & 1 \\ 1 & 9 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}$$

In this matrix, we have 9 along the diagonal and 1 in the off-diagonal positions. This matrix is derived from the evaluation of $\varphi(x_i)$ and $\varphi(x_j)$ over all pairs of points.

Moving on to the dual problem, we set up the objective function Q as a function of the Lagrange multipliers α_i. The objective function is computed as follows:

$$Q = \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_i \alpha_j d_i d_j k(x_i, x_j)$$

(Refer Slide Time: 14:28)



Simplifying this, we have:

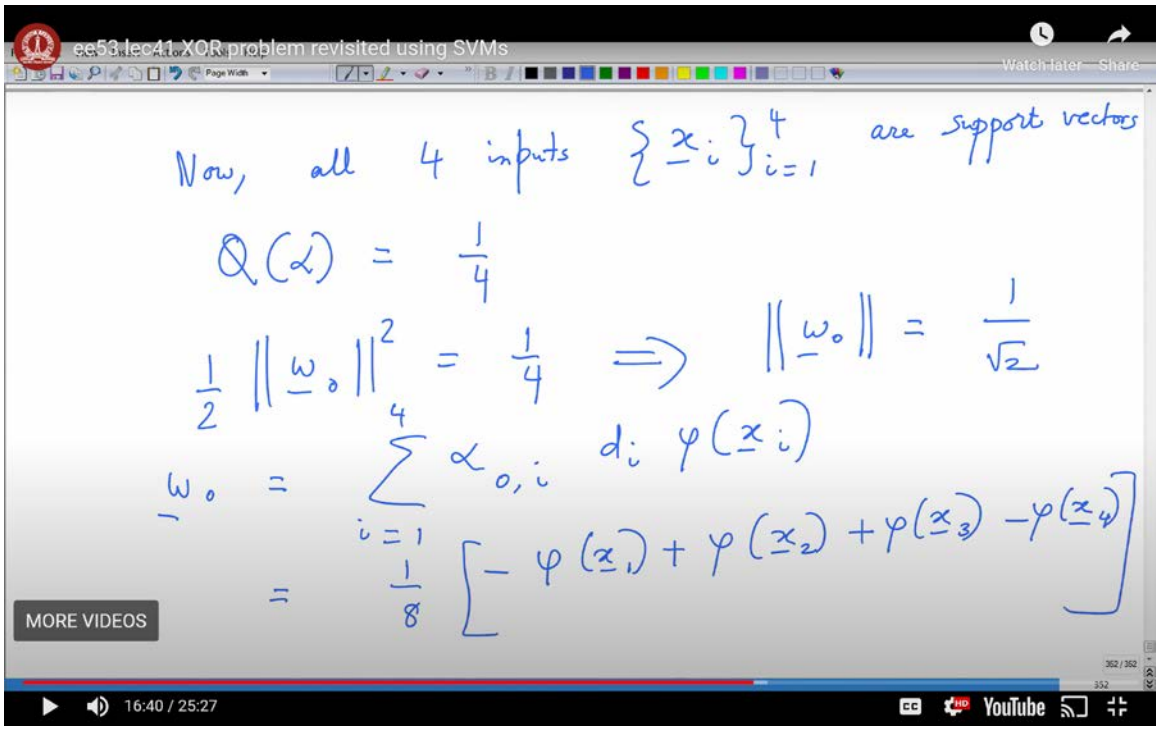
$$Q = \sum_{i=1}^{4} \alpha_i - \frac{1}{2} (\text{the quantity involving all the terms})$$

To find the optimal α, we take the partial derivatives of Q with respect to each α_i (for i = 1, 2, 3, 4) and set them to zero. This yields four different equations. For instance:

$$9\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 = 1$$

We repeat this process for each α_i, resulting in a system of equations. Solving this system, we find that the optimal α_i for i = 1, 2, 3, 4 is $\frac{1}{8}$. Thus, all four input points are support vectors, as expected.

(Refer Slide Time: 16:40)



When I substitute the values into Q(α), I find that the result is $\frac{1}{4}$, while the cost in the second part is zero. This means that half of the norm of the weight vector w_0 squared is $\frac{1}{4}$, which implies that the norm of the optimal weight vector is $\frac{1}{\sqrt{2}}$. With this information, I can determine the optimal weight vector w_0 using the formula:

$$w_0 = \sum_{i=1}^{4} \alpha_i^{\text{opt}}\, d_i\, \phi(x_i)$$

Here, α_i^{opt} for i = 1 to 4 is $\frac{1}{8}$. Substituting these values, I calculate:

$$w_0 = \frac{1}{8}\left(-\phi(x_1) + \phi(x_2) + \phi(x_3) - \phi(x_4)\right)$$

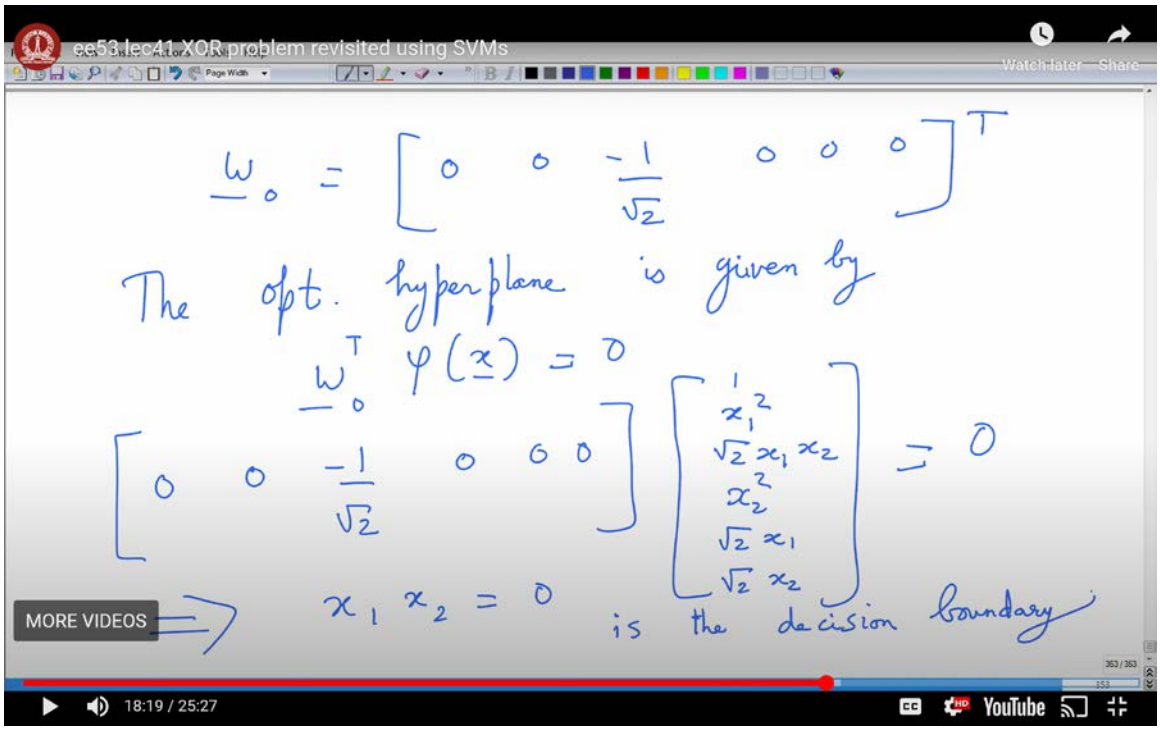
This results in:

$$w_0 = \left(0,0,-\frac{1}{\sqrt{2}}\right)$$

This weight vector is verified as correct. From the optimal weight vector, I can compute the optimal hyperplane, which is given by:

$$w_0^T \phi(x) = 0$$

(Refer Slide Time: 18:19)



Substituting the values, we have:

$$\left(0,0,-\frac{1}{\sqrt{2}}\right) \cdot \left(1, x_1, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2\right) = 0$$

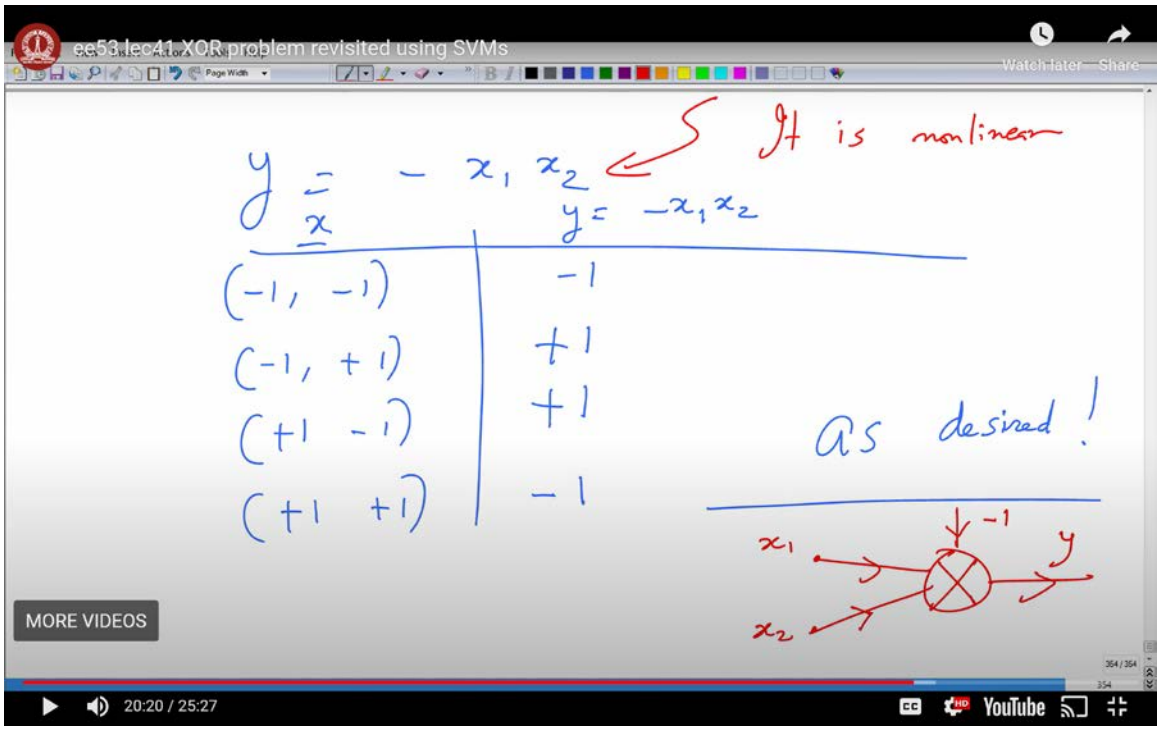
This simplifies to:

$$-\frac{1}{\sqrt{2}} \cdot \sqrt{2}x_1x_2 = 0$$

Thus, the decision boundary is:

$$x_1 x_2 = 0$$

(Refer Slide Time: 20:20)



To ensure a consistent sign, I reformulate it as:

$$y = -x_1 x_2$$

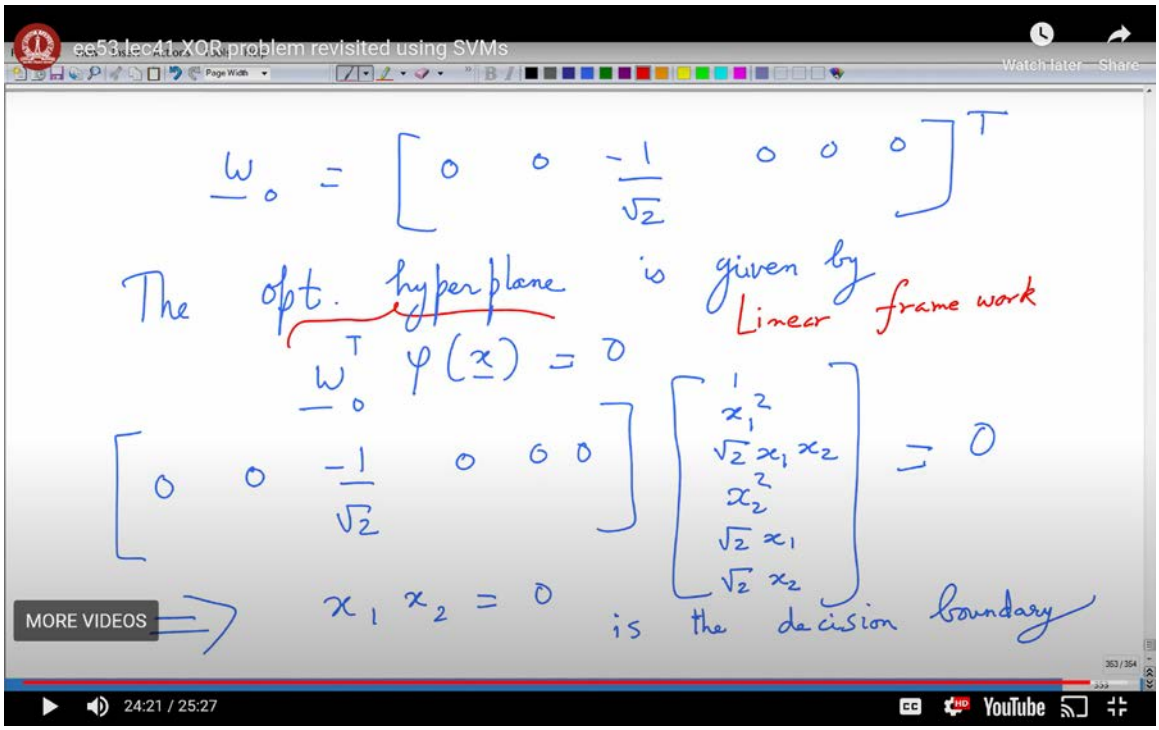
Let's see how this works with the data points:

- For (-1, -1), the response is -1.
- For (-1, +1), the response is +1.
- For (+1, -1), the response is +1.
- For (+1, +1), the response is -1.

These results are consistent with the desired responses for the XOR problem. Starting with the XOR response in this manner, we conclude that the equation $y = -x_1 x_2$ represents the decision boundary for the XOR problem.

The key takeaway is that while the function $y = -x_1 x_2$ is nonlinear, we solved it using a linear optimization framework. The hyperplane is essentially a linear equation in the transformed feature space, which allows us to handle nonlinear classification problems through a linear framework. This example demonstrates how SVMs can be employed to solve classification problems effectively.

(Refer Slide Time: 24:21)



Given the dataset, which includes the set of points and their corresponding responses, I then formulate the kernel. You might wonder how to choose the right kernel. The choice of kernel is indeed subjective and depends on the specific application. Various inner product kernels are available, and the selection of the appropriate one typically involves using polynomial kernels. By adjusting the power p of the polynomial kernel, I can map the input vector to an arbitrary higher dimension as needed.

In this example, I transformed a two-dimensional vector into a six-dimensional vector, allowing me to solve the problem in this higher-dimensional space. This approach enables

us to derive a linear classifier in six dimensions, which wouldn't be possible in just two dimensions.

The choice of dimension is crucial and often requires experimentation. There is a connection to the VC (Vapnik-Chervonenkis) dimension, which relates to how much we can lift the input space. Once we have the feature space in a higher dimension, we construct the Gram matrix for all pairs of data points and derive the objective function.

In this case, solving the dual problem proved easier since it involved only the Lagrange multipliers and directly depended on the data points through the kernel. Thus, I opted for this framework. By solving the dual optimization problem, we obtain the optimal Lagrange multipliers, which can then be substituted into the equation for the optimal weight vector. This is achieved by setting the derivative of the Lagrangian with respect to the weight vector to zero, solving for w, and then evaluating the cost.

The final step involves computing $w^T \phi(x)$, which resolves the problem. This process is crucial, as it integrates the $\varphi(x_i)$ values into the solution for w_0, ensuring that all components from the kernel are accounted for.

This example illustrates the fundamental approach, but in real-world scenarios, problems may involve additional complexities such as noise or more intricate loss functions. In such cases, setting up the problem carefully, defining the constraints, and utilizing an optimization solver can help obtain accurate results. The key takeaway is to grasp the underlying concepts and methodologies, which is central to effectively addressing practical problems using this formulation.