6.5 EXAMPLE: XOR PROBLEM (REVISITED)

To illustrate the procedure for the design of a support vector machine, we revisit the XOR (Exclusive OR) problem discussed in Chapters 4 and 5. Table 6.2 presents a summary of the input vectors and desired responses for the four possible states.

To proceed, let (Cherkassky and Mulier, 1998)

\[ K(\mathbf{x}, \mathbf{x}_i) = (1 + \mathbf{x}^T \mathbf{x}_i)^2 \] \hspace{1cm} (6.43)

With \( \mathbf{x} = [x_1, x_2]^T \) and \( \mathbf{x}_i = [x_{i1}, x_{i2}]^T \), we may thus express the inner-product kernel \( K(\mathbf{x}, \mathbf{x}_i) \) in terms of monomials of various orders as follows:

\[ K(\mathbf{x}, \mathbf{x}_i) = 1 + x_1^2 x_{i1}^2 + 2x_1x_2x_{i1}x_{i2} + x_2^2 x_{i2}^2 + 2x_1 x_{i1} + 2x_2 x_{i2} \]

The image of the input vector \( \mathbf{x} \) induced in the feature space is therefore deduced to be

\[ \varphi(\mathbf{x}) = [1, x_1^2, \sqrt{2}x_1 x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2]^T \]

Similarly,

\[ \varphi(\mathbf{x}_i) = [1, x_{i1}^2, \sqrt{2}x_{i1} x_{i2}, x_{i2}^2, \sqrt{2}x_{i1}, \sqrt{2}x_{i2}]^T, \hspace{1cm} i = 1, 2, 3, 4 \]

From Eq. (6.41), we also find that

\[
K = \begin{bmatrix}
9 & 1 & 1 & 1 \\
1 & 9 & 1 & 1 \\
1 & 1 & 9 & 1 \\
1 & 1 & 1 & 9 \\
\end{bmatrix}
\]

The objective function for the dual form is therefore (see Eq. (6.40))

\[
Q(\alpha) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2} (9\alpha_1^2 - 2\alpha_1\alpha_2 - 2\alpha_1\alpha_3 + 2\alpha_1\alpha_4 \\
+ 9\alpha_2^2 + 2\alpha_2\alpha_3 - 2\alpha_2\alpha_4 + 9\alpha_3^2 - 2\alpha_3\alpha_4 + 9\alpha_4^2)
\]

Optimizing \( Q(\alpha) \) with respect to the Lagrange multipliers yields the following set of simultaneous equations:

\[
\begin{align*}
9\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 &= 1 \\
-\alpha_1 + 9\alpha_2 + \alpha_3 - \alpha_4 &= 1 \\
-\alpha_1 + \alpha_2 + 9\alpha_3 - \alpha_4 &= 1 \\
\alpha_1 - \alpha_2 - \alpha_3 + 9\alpha_4 &= 1
\end{align*}
\]

<table>
<thead>
<tr>
<th>TABLE 6.2 XOR Problem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input vector, ( \mathbf{x} )</td>
</tr>
<tr>
<td>((-1, -1))</td>
</tr>
<tr>
<td>((-1, +1))</td>
</tr>
<tr>
<td>((+1, -1))</td>
</tr>
<tr>
<td>((+1, +1))</td>
</tr>
</tbody>
</table>
Hence, the optimum values of the Lagrange multipliers are

\[ \alpha_{o,1} = \alpha_{o,2} = \alpha_{o,3} = \alpha_{o,4} = \frac{1}{8} \]

This result indicates that in this example all four input vectors \( \{x_i\}_{i=1}^4 \) are support vectors. The optimum value of \( Q(\alpha) \) is

\[ Q_o(\alpha) = \frac{1}{4} \]

Correspondingly, we may write

\[ \frac{1}{2} \|w_o\|^2 = \frac{1}{4} \]

or

\[ \|w_o\| = \frac{1}{\sqrt{2}} \]

From Eq. (6.42), we find that the optimum weight vector is

\[ w_o = \frac{1}{8} \left[ -\varphi(x_1) + \varphi(x_2) + \varphi(x_3) - \varphi(x_4) \right] \]

\[
\begin{bmatrix}
1 \\
1 \\
\sqrt{2} \\
1 \\
-\sqrt{2} \\
-\sqrt{2}
\end{bmatrix}
+ 
\begin{bmatrix}
1 \\
-\sqrt{2} \\
-\sqrt{2} \\
\sqrt{2}
\end{bmatrix}
+ 
\begin{bmatrix}
1 \\
1 \\
\sqrt{2} \\
\sqrt{2}
\end{bmatrix}
- 
\begin{bmatrix}
1 \\
1 \\
\sqrt{2} \\
\sqrt{2}
\end{bmatrix}
\]

\[
\begin{bmatrix}
0 \\
0 \\
-1/\sqrt{2} \\
0 \\
0 \\
0
\end{bmatrix}
\]

The first element of \( w_o \) indicates that the bias \( b \) is zero.

The optimal hyperplane is defined by (see Eq. (6.33))

\[ w_o^T \varphi(x) = 0 \]
That is,

\[
\begin{bmatrix}
0, 0, \frac{-1}{\sqrt{2}}, 0, 0, 0
\end{bmatrix}
\begin{bmatrix}
1 \\
x_1^2 \\
\sqrt{2}x_1x_2 \\
x_2^2 \\
\sqrt{2}x_1 \\
\sqrt{2}x_2
\end{bmatrix} = 0
\]

which reduces to

\[-x_1x_2 = 0\]

The polynomial form of support vector machine for the XOR problem is as shown in Fig. 6.6a. For both \(x_1 = x_2 = -1\) and \(x_1 = x_2 = +1\), the output \(y = -1\); and for both \(x_1 = -1, x_2 = +1\) and \(x_1 = +1 \text{ and } x_2 = -1\), we have \(y = +1\). Thus the XOR problem is solved as indicated in Fig. 6.6b.

4.6 COMPUTER EXPERIMENT

In this computer experiment we revisit the pattern-classification problem that we studied in Chapters 4 and 5. The experiment involved the classification of two overlapping two-dimensional Gaussian distributions labeled 1 (Class \(\mathcal{C}_1\)) and 2 (Class \(\mathcal{C}_2\)). The scatter plots for these two sets of data are shown in Fig. 4.14. The probability of correct classification produced by the Bayesian (optimum) classifier is calculated to be

\[p_c = 81.51\%\]