Chapter 10: Support Vector Machines

We're going to find a hyperplane that separates biopsy samples as either "benign" or "malignant" based on the expression levels of proteins. Our training data are

status	Ras	Mek	Erk	p53
benign	9.1	0.2	0.9	6.5
benign	3.8	2.8	1.0	7.7
benign	5.2	3.5	0.7	5.9
malignant	0.9	5.6	1.5	4.2
malignant	2.1	9.2	1.4	2.6

1. **Assign codes of $+1$ or -1 to the samples.**

Let's write a quadratic SVM program to find the hyperplane. We need constraints such that

$$\begin{aligned} \mathbf{a}\cdot\mathbf{x} &\geq b+1 \quad \text{for the } +1 \text{ points} \\ \mathbf{a}\cdot\mathbf{x} &\leq b-1 \quad \text{for the } -1 \text{ points} \end{aligned}$$

2. **What are the dimensions of a and b for this problem?**

3. **Write out the constraints for all five training samples.**

4. **Write the objective for the quadratic program.**

5. **Which of these two hyperplanes is a solution to the SVM problem?**

$$\mathbf{a} = \begin{pmatrix} -1.21 \\ -0.27 \\ 1.21 \\ 0.35 \end{pmatrix}, \quad b = -0.86$$

$$\mathbf{a} = \begin{pmatrix} -0.30 \\ 0.18 \\ 0.06 \\ -0.17 \end{pmatrix}, \quad b = -0.92$$

Solutions

1. **Assign codes of $+1$ or -1 to the samples.**

We assign -1 to the benign samples and +1 to the malignant samples. The coding is arbitrary — we chose -1 for benign samples only because it would be easy to remember that benign samples are a “negative" test result.

The coded training data appear below.

status	code	Ras	Mek	Erk	p53
benign	-1	9.1	0.2	0.9	6.5
benign	-1	3.8	2.8	1.0	7.7
benign	-1	5.2	3.5	0.7	5.9
malignant	+1	0.9	5.6	1.5	4.2
malignant	+1	2.1	9.2	1.4	2.6

2. **What are the dimensions of a and b for this problem?**

There are four features (Ras, Mek, Erk, and p53), for **a** is vector of length four:

$$\mathbf{a} = \begin{pmatrix} a_{\text{Ras}} \\ a_{\text{Mek}} \\ a_{\text{Erk}} \\ a_{\text{p53}} \end{pmatrix}$$

The unknown parameter b is always a scalar regardless of the number of features.

3. **Write out the constraints for all five training samples.**

Sample 1 (benign, −1):

$$9.1a_{\text{Ras}} + 0.2a_{\text{Mek}} + 0.9a_{\text{Erk}} + 6.5a_{\text{p53}} \leq b - 1$$

Sample 2 (benign, −1):

$$3.8a_{\text{Ras}} + 2.8a_{\text{Mek}} + 1.0a_{\text{Erk}} + 7.7a_{\text{p53}} \leq b - 1$$

Sample 3 (benign, −1):

$$5.2a_{\text{Ras}} + 3.5a_{\text{Mek}} + 0.7a_{\text{Erk}} + 5.9a_{\text{p53}} \leq b - 1$$

Sample 4 (malignant, +1):

$$0.9a_{\text{Ras}} + 5.6a_{\text{Mek}} + 1.5a_{\text{Erk}} + 4.2a_{\text{p53}} \geq b + 1$$

Sample 5 (malignant, +1):

$$2.1a_{\text{Ras}} + 9.2a_{\text{Mek}} + 1.4a_{\text{Erk}} + 2.6a_{\text{p53}} \geq b + 1$$

4. Write the objective for the quadratic program.

The objective is

$$\underset{a_{\text{Ras}}, a_{\text{Mek}}, a_{\text{Erk}}, a_{\text{p53}}, b}{\text{minimize}} \quad a_{\text{Ras}}^2 + a_{\text{Mek}}^2 + a_{\text{Erk}}^2 + a_{\text{p53}}^2$$

5. Which of these two hyperplanes is a solution to the SVM problem?

$$\mathbf{a} = \begin{pmatrix} -1.21 \\ -0.27 \\ 1.21 \\ 0.35 \end{pmatrix}, \quad b = -0.86$$

$$\mathbf{a} = \begin{pmatrix} -0.30 \\ 0.18 \\ 0.06 \\ -0.17 \end{pmatrix}, \quad b = -0.92$$

The separating hyperplane sits at $\mathbf{a} \cdot \mathbf{x} = b$, which is midway between the $\mathbf{a} \cdot \mathbf{x} = b + 1$ and $\mathbf{a} \cdot \mathbf{x} = b - 1$ plates we separated during classification. All the negative samples should be below $\mathbf{a} \cdot \mathbf{x} = b$, and the positive samples should be above.

Using

$$\mathbf{a} = \begin{pmatrix} -1.21 \\ -0.27 \\ 1.21 \\ 0.35 \end{pmatrix}, \quad b = -0.86$$

Sample 1 (benign, −1):

$$\mathbf{a} \cdot \mathbf{x} = -7.7 \leq -0.86 \quad \text{correct}$$

Sample 2 (benign, −1):

$$\mathbf{a} \cdot \mathbf{x} = -1.4 \leq -0.86 \quad \text{correct}$$

Sample 3 (benign, −1):

$$\mathbf{a} \cdot \mathbf{x} = -4.3 \leq -0.86 \quad \text{correct}$$

Sample 4 (malignant, +1):

$$\mathbf{a} \cdot \mathbf{x} = 0.7 \geq -0.86 \quad \text{correct}$$

Sample 5 (malignant, +1):

$$\mathbf{a} \cdot \mathbf{x} = -2.4 \leq -0.86 \quad \textbf{incorrect}$$

These values for $\mathbf{a}$ and b do not classify the samples. Let's try the second set:

$$\mathbf{a} = \begin{pmatrix} -0.30 \\ 0.18 \\ 0.06 \\ -0.17 \end{pmatrix}, \quad b = -0.92$$

Sample 1 (benign, −1):

$$\mathbf{a} \cdot \mathbf{x} = -3.7 \leq -0.92 \quad \text{correct}$$

Sample 2 (benign, −1):

$$\mathbf{a} \cdot \mathbf{x} = -1.9 \leq -0.92 \quad \text{correct}$$

Sample 3 (benign, −1):

$$\mathbf{a} \cdot \mathbf{x} = -1.9 \leq -0.92 \quad \text{correct}$$

Sample 4 (malignant, +1):

$$\mathbf{a} \cdot \mathbf{x} = 0.1 \geq -0.92 \quad \text{correct}$$

Sample 5 (malignant, +1):

$$\mathbf{a} \cdot \mathbf{x} = 0.7 \geq -0.92 \quad \text{correct}$$

These values for **a** and b correctly classify all the samples.