

## 18 Support Vector Machine: Finding the Perfect Margin

### 18.1 Introduction: The Quest for the Optimal Hyperplane

#### 18.1.1 SVM Origins and Motivation

The core concept behind Support Vector Machine (SVM) is to identify the optimal hyperplane that separates data points from different classes by the widest possible margin. This hyperplane serves as a decision boundary, allowing the model to classify new, unseen data points.

To address cases where data are not linearly separable, SVMs employ kernel functions to transform input data into a higher-dimensional feature space where a linear separation is possible. Commonly used kernel functions include the linear kernel, polynomial kernel, and radial basis function (RBF) kernel, which will be discussed in the next chapter.

SVMs are considered highly effective for complex datasets because they perform well in high-dimensional spaces and can model nonlinear relationships. They also incorporate regularization to reduce the risk of overfitting. However, SVMs can be computationally intensive when applied to large datasets, and choosing and tuning the appropriate kernel function requires expertise and experience.

Figure 1 illustrates the motivation behind Support Vector Machine. As shown, imagine a lake with groups of red and blue markers representing different categories on either side of a narrow channel, analogous to a kayak navigating between reefs. The objective is to chart a straight path through the water that allows the kayak to pass with the greatest possible distance from the closest reefs on both sides.

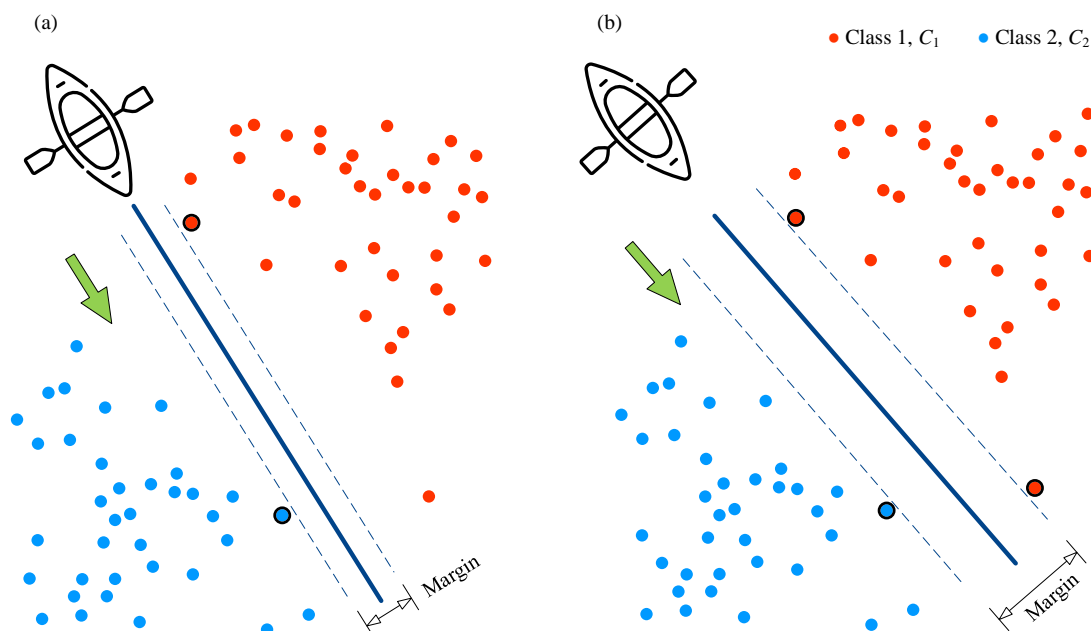


Figure 1. The blue and red points represent two labeled classes. SVM finds the optimal separating hyperplane that maximizes the margin—the distance to the nearest points (support vectors)

In this scenario, the optimal path provides the widest clearance, which is precisely what SVM aims to achieve: identifying the decision boundary that maximizes the margin between categories. Comparing Figure 1 (a) and Figure 1 (b), it is evident that the boundary in Figure 1 (b) offers a larger margin and is therefore preferable.

In Figure 1 (b), the black-circled points highlight the so-called support vectors. These are the data points closest to the decision boundary and play a critical role in determining its precise position and orientation.

Other data points do not influence the location of the boundary. This property allows SVMs to remain effective even when working with data sets where the number of features is much greater than the number of samples.

The distance between the two dashed lines parallel to the decision boundary is called the margin. The SVM's optimization objective is explicitly to maximize this margin. This principle makes SVM robust and effective for many types of classification problems.

### 18.1.2 From Linear to Nonlinear: Challenges in Real Data

In practice, many datasets are not linearly separable, meaning no straight line (or hyperplane in higher dimensions) can perfectly separate the two classes, such as the blue and red points in Figure 2.

To handle such cases, SVM use two key approaches: the soft margin and the kernel trick.

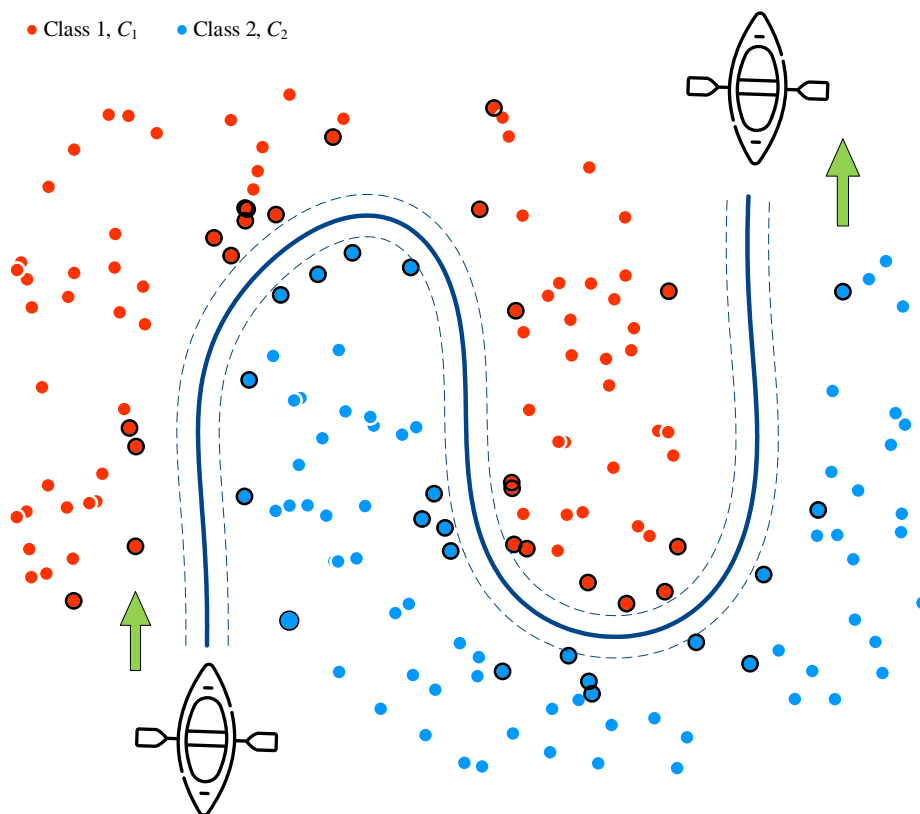


Figure 2. Non-linearly separable data refers to data that cannot be separated into distinct classes using a straight line (or hyperplane)

### 18.1.3 Core Components: Margin, Support Vectors, and Decision Boundary

The soft margin, as shown in Figure 3, allows some misclassification by introducing slack variables, enabling the model to tolerate overlapping data while still maximizing the margin. We will explore this in depth later in this chapter.

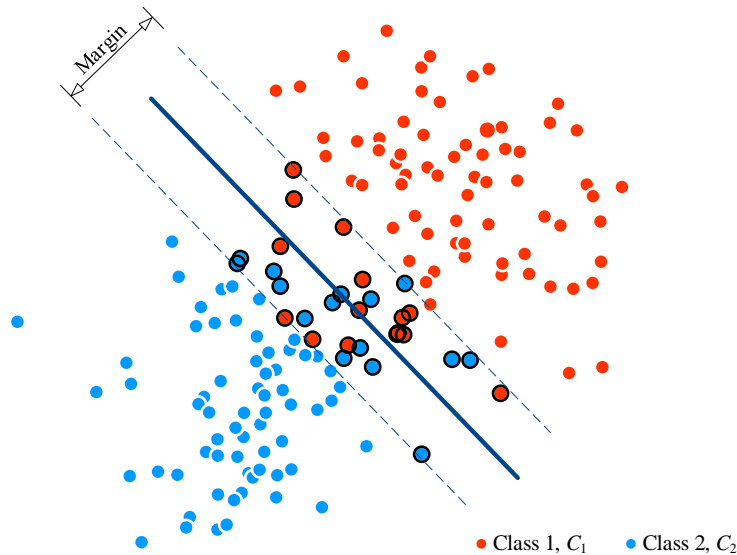


Figure 3. Soft margin in SVM introduces slack variables to allow some misclassification.

The kernel trick implicitly maps data into a higher-dimensional feature space, where it becomes linearly separable, allowing SVM to find a linear decision boundary in that transformed space.

As illustrated in Figure 4, the raw sample data have two features visualized in a two-dimensional plane, where the blue and red points cannot be separated by a straight line. By applying a kernel function, the data points are implicitly mapped from this original 2D space into a higher-dimensional feature space, where the points become linearly separable.

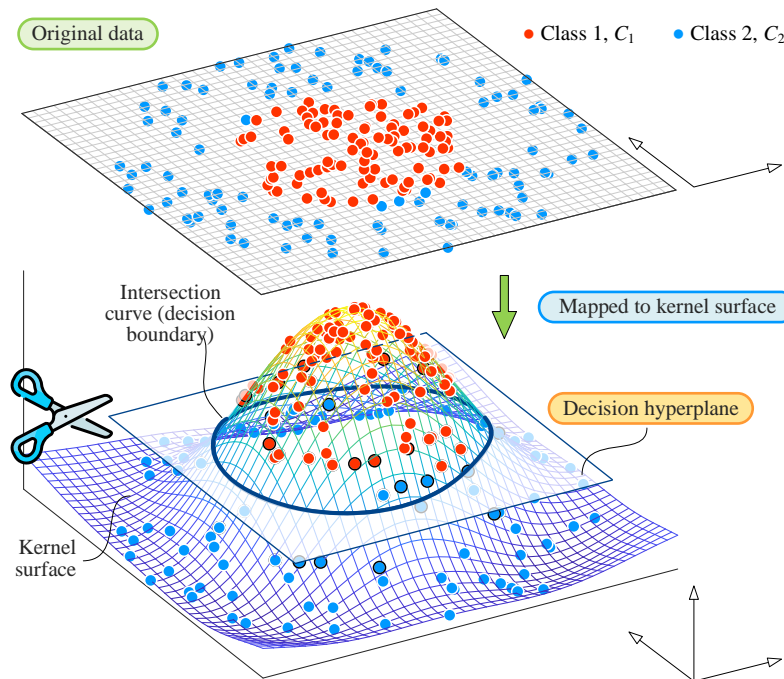


Figure 4. Mapping to a higher dimension allows SVM to find a linear separating hyperplane in the transformed space, effectively separating the blue and red classes that are inseparable in lower dimensions.

To solve linear inseparability problems, SVMs typically combine the use of kernels with the soft margin approach, which allows some data points to be misclassified to improve generalization. Figure 5 demonstrates how SVM uses kernel techniques alongside soft margins to classify complex data distributions, such as ring-shaped data. We will talk more about kernel trick in the next chapter.

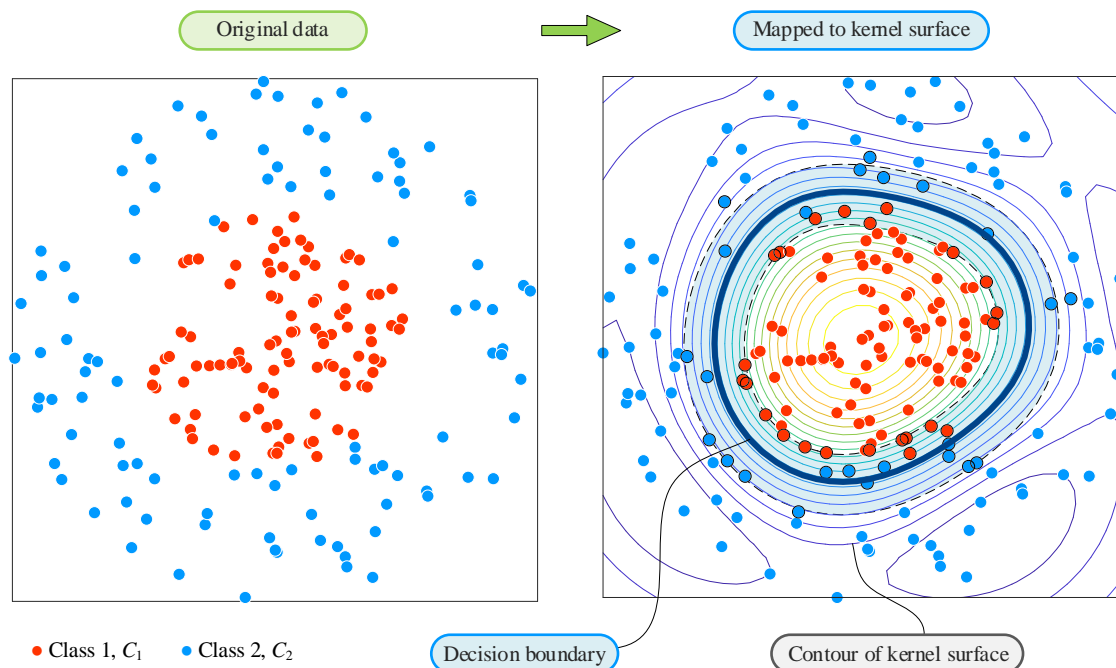


Figure 5. SVMs combine kernel functions with the soft margin approach, allowing some misclassification to maximize the margin and improve the model's ability to generalize on non-linearly separable data.

## 18.2 Linearly Separable Data: Hard Margin SVM

### 18.2.1 Decision Boundary and Labels

The hard spacing method in the support vector machine is used to process linear separable data. Using the vector geometry knowledge explained in the volume "Matrix Power", this section will construct the mathematical relationships between elements such as support vectors, decision boundaries, classification labels, and intervals in SVM.

As shown in Figure 6, the blue and red points represent two different classes. The solid dark blue line is the decision boundary (optimal hyperplane) that separates the two classes.

The dashed lines show the margin, the maximum distance from the decision boundary to the nearest data points of each class. The width of margin shown in Figure 6 is  $2h$  ( $h > 0$ ).

The black circled points are the support vectors, which define the margin and the position of the decision boundary. The green arrow indicates the gradient vector  $\mathbf{w}$  perpendicular to the decision boundary.

The decision boundary is defined as

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = 0 \quad (1)$$

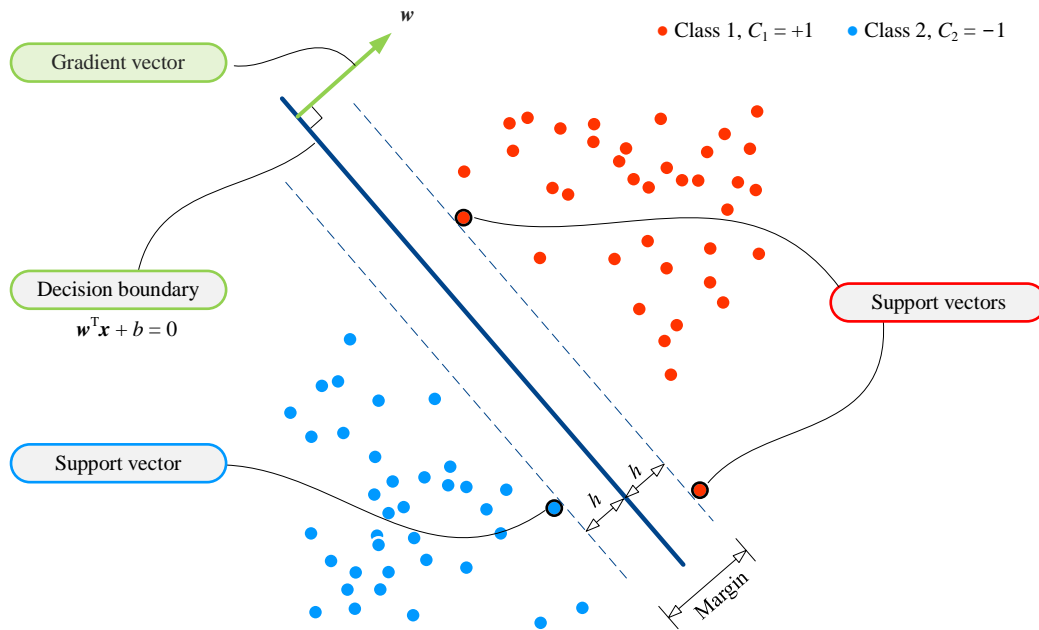


Figure 6. Hard spaced SVM deals with binary classification problems

For binary classification, the data points "above" the decision boundary satisfy

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b > 0 \quad (2)$$

The data points "below" the decision boundary satisfy

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b < 0 \quad (3)$$

For any query point  $\mathbf{q}$ , the binary decision function  $p(\mathbf{q})$  can be expressed as

$$p(\mathbf{q}) = \text{sign}(\mathbf{w}^T \mathbf{q} + b) \quad (4)$$

where  $\text{sign}()$  is the sign function that outputs  $+1$  or  $-1$ .

The data points that lie on one side of the decision boundary are classified as  $+1$ , while those on the opposite side are classified as  $-1$ . In other words, points "above" the decision boundary are predicted as class  $+1$ , and points "below" it as class  $-1$ .

### 18.2.2 Measuring Distance and Parallel Margins

Consider a support vector whose coordinates are represented by the column vector  $\mathbf{q}$ . The distance from this support vector  $\mathbf{q}$  to the decision boundary is given by

$$d = \frac{|\mathbf{w}^T \mathbf{q} + b|}{\|\mathbf{w}\|} = \frac{|\mathbf{w} \cdot \mathbf{q} + b|}{\|\mathbf{w}\|} \quad (5)$$

Generally, when measuring the distance between points and lines (or hyperplanes), the absolute value is used to ensure a non-negative scalar. However, in classification problems, retaining the sign of the distance is useful, as it indicates on which side of the hyperplane the point lies.

By removing the absolute value in the numerator, the signed distance is

$$d = \frac{\mathbf{w}^T \mathbf{q} + b}{\|\mathbf{w}\|} = \frac{\mathbf{w} \cdot \mathbf{q} + b}{\|\mathbf{w}\|} \quad (6)$$

If  $d > 0$ , the point  $\mathbf{q}$  is above the hyperplane; If  $d < 0$ , the point  $\mathbf{q}$  is below the hyperplane.

For instance, as shown in Figure 7,  $\mathbf{q}_1$  is above the decision boundary (positive  $d$ ), while  $\mathbf{q}_2$  is below it (negative  $d$ ).

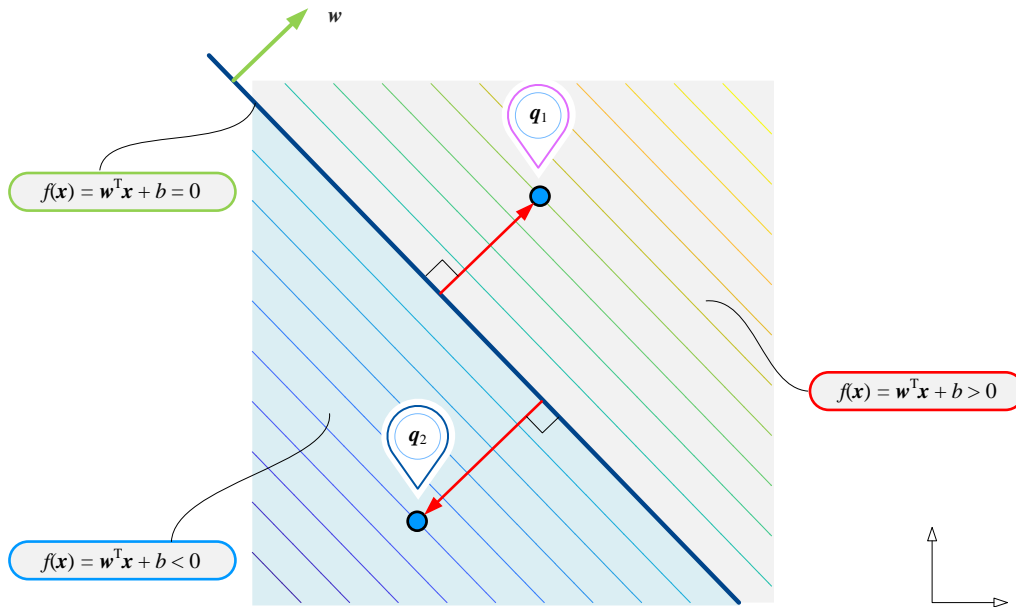


Figure 7. The solid dark blue line represents the decision boundary (hyperplane). Other contour lines show the values of function  $f(\mathbf{x})$ .

### 18.2.3 Mathematical Formulation of Margin

As shown in Figure 8, the hard margin defines two parallel boundaries, called the margin boundaries.

The lower boundary is denoted as  $l_1$ , which is located at a distance of  $-h$  from the decision boundary (the optimal separating hyperplane).

The support vector  $C$  lies exactly on this boundary, so it satisfies the equation

$$\frac{\mathbf{w}^T \mathbf{x} + b}{\|\mathbf{w}\|} = -h \quad (7)$$

Similarly, the upper boundary is denoted as  $l_2$ , located at a distance of  $+h$  from the decision boundary. Support vectors  $A$  and  $B$  lie on this upper boundary and satisfy

$$\frac{\mathbf{w}^T \mathbf{x} + b}{\|\mathbf{w}\|} = +h \quad (8)$$

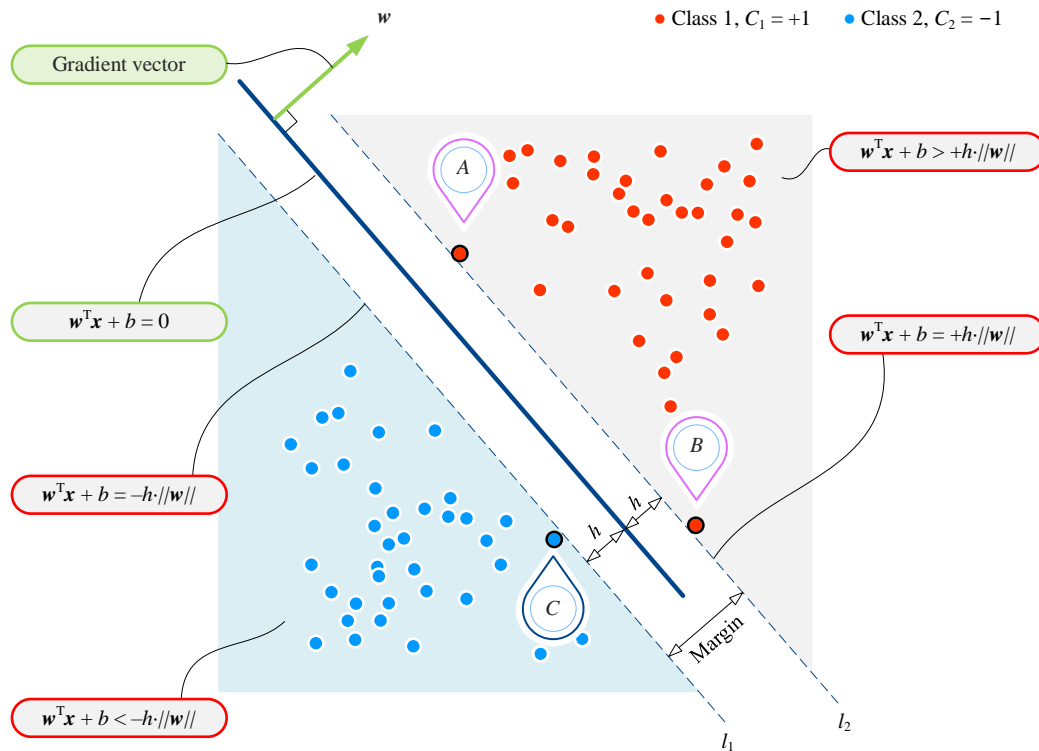


Figure 8. The hard margin creates two parallel boundaries—one lower boundary at a negative distance and one upper boundary at a positive distance from the decision boundary—with support vector  $C$  lying exactly on the lower boundary and support vectors  $A$  and  $B$  lying on the upper boundary, each satisfying their respective boundary conditions.

The decision boundary, shown as the dark blue line in Figure 8, effectively separates the two classes of sample data.

The data points located on or beyond the upper boundary which satisfy

$$\frac{\mathbf{w}^T \mathbf{x} + b}{\|\mathbf{w}\|} \geq +h, \quad y = +1 \quad (9)$$

are labeled as  $y = +1$ .

On the contrary, those on or beyond the lower boundary which satisfy

$$\frac{\mathbf{w}^T \mathbf{x} + b}{\|\mathbf{w}\|} \leq -h, \quad y = -1 \quad (10)$$

are labeled as  $y = -1$ .

Formally, this can be summarized as

$$\frac{(\mathbf{w}^T \mathbf{x} + b)y}{\|\mathbf{w}\|h} \geq 1 \quad (11)$$

To simplify the equation above, let

$$\|\mathbf{w}\|h = 1 \quad (12)$$

Then we can get

$$(\mathbf{w}^T \mathbf{x} + b)y \geq 1 \quad (13)$$

The margin width  $2h$  can be expressed as a function of  $\mathbf{w}$  by

$$2h = \frac{2}{\|\mathbf{w}\|} \quad (14)$$

## 18.3 Optimization Problem: Maximizing the Margin

### 18.3.1 Objective: Margin Maximization

The core idea of SVM is to maximize the margin — the interval separating classes — to achieve robust classification. This section addresses the formulation and solution of the SVM optimization problem using the method of Lagrange multipliers.

We use  $\mathbf{w}$  and  $b$  as optimization variables to maximize the margin under the constraints that ensure correct classification of all  $n$  training samples.

The optimization problem is formulated as

$$\begin{aligned} \arg \max_{\mathbf{w}, b} \quad & \frac{2}{\|\mathbf{w}\|} \\ \text{subject to} \quad & (\mathbf{x}^{(i)} \mathbf{w} + b)y^{(i)} \geq 1, \quad i = 1, 2, 3, \dots, n \end{aligned} \quad (15)$$

Maximizing the margin is equivalent to minimizing the squared norm of  $\mathbf{w}$

$$\begin{aligned} \arg \min_{\mathbf{w}, b} \quad & \frac{\|\mathbf{w}\|^2}{2} = \frac{\mathbf{w}^T \mathbf{w}}{2} = \frac{\mathbf{w} \cdot \mathbf{w}}{2} \\ \text{subject to} \quad & (\mathbf{x}^{(i)} \mathbf{w} + b)y^{(i)} \geq 1, \quad i = 1, 2, 3, \dots, n \end{aligned} \quad (16)$$

### 18.3.2 Lagrange Formulation

To handle the constraints, the Lagrangian function  $L(\mathbf{w}, b, \boldsymbol{\lambda})$  is constructed by introducing Lagrange multipliers for each constraint

$$L(\mathbf{w}, b, \boldsymbol{\lambda}) = \frac{\mathbf{w} \cdot \mathbf{w}}{2} + \sum_{i=1}^n \lambda_i \left( 1 - y^{(i)} (\mathbf{x}^{(i)} \mathbf{w} + b) \right) \quad (17)$$

This reformulates the constrained optimization problem into an unconstrained one.

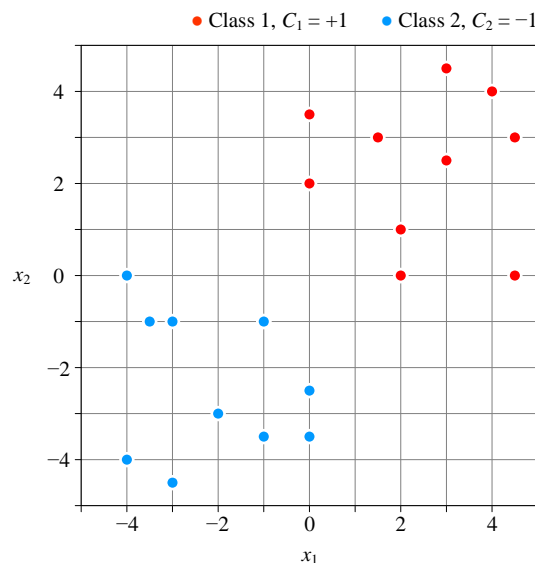
To avoid overwhelming readers, we will skip the detailed derivation of the Lagrange dual problem here. Interested readers are encouraged to explore this important topic on their own to gain a deeper understanding of the underlying optimization theory behind Support Vector Machine.

## 18.4 Hard Margin in Action: A Worked Example

### 18.4.1 Visualizing Linearly Separable Data

This section presents a practical example to demonstrate the implementation of a hard margin SVM algorithm.

Figure 9 shows 20 sample data points that are clearly linearly separable. The task is to use an SVM to classify these samples correctly.



The hard margin approach defines two boundary lines parallel to the decision boundary.

The upper boundary  $l_1$  satisfies

$$w_1x_1 + w_2x_2 + b = 1 \Rightarrow x_2 = -\frac{w_1}{w_2}x_1 - \frac{b-1}{w_2} \quad (20)$$

The lower boundary  $l_2$  satisfies

$$w_1x_1 + w_2x_2 + b = -1 \Rightarrow x_2 = -\frac{w_1}{w_2}x_1 - \frac{b+1}{w_2} \quad (21)$$

Note that because  $w_2 = 0$ , these expressions hold and help visualize the margin, but represent a simplification for understanding.

Figure 10 displays the classification outcomes, highlighting three support vectors located at points **Figure 10**, A (0, 2), B (2, 0) and C (-1, -1). These points lie exactly on the margin boundaries and determine the position of the decision boundary. The remaining 17 samples do not influence the decision boundary.

The decision boundary corresponds analytically to

$$\frac{x_1}{2} + \frac{x_2}{2} = 0 \Rightarrow x_1 + x_2 = 0 \Rightarrow x_2 = -x_1 \quad (22)$$

The classification decision function is

$$p(x_1, x_2) = \text{sign}(x_1 + x_2) \quad (23)$$

The upper boundary  $l_1$  is

$$\frac{x_1}{2} + \frac{x_2}{2} = 1 \Rightarrow x_2 = -x_1 + 2 \quad (24)$$

The lower boundary  $l_2$  is

$$\frac{x_1}{2} + \frac{x_2}{2} = -1 \Rightarrow x_2 = -x_1 - 2 \quad (25)$$

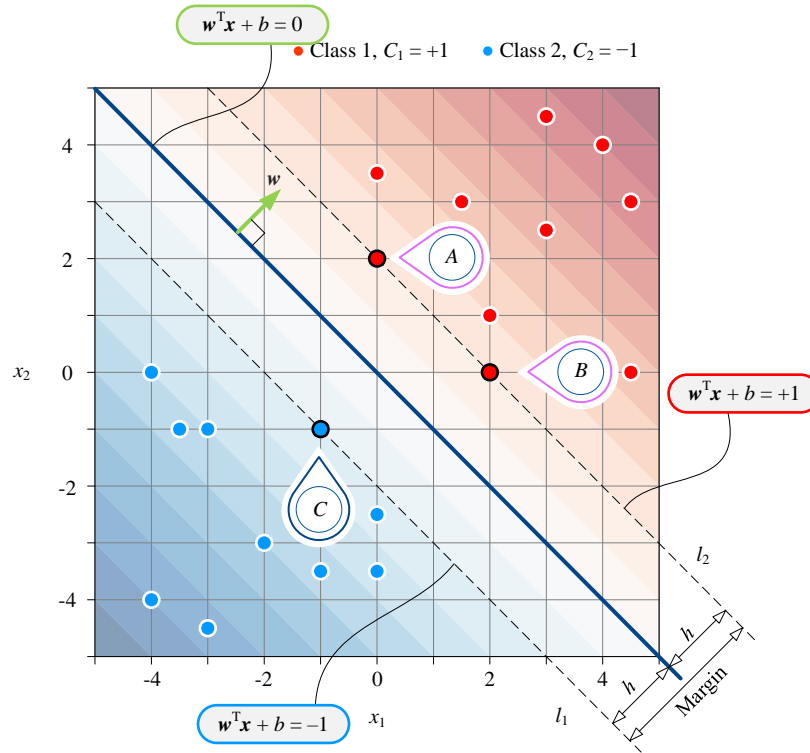


Figure 10. The three support vectors, A, B, and C, uniquely determine the position of the decision boundary and the width of the margin.

At (4, 4), classification is

$$p(4, 4) = \text{sign}(4 + 4) = +1 \quad (26)$$

At (-2, -3), classification is

$$p(-2, -3) = \text{sign}(-2 - 3) = -1 \quad (27)$$

#### 18.4.2 Effect of Support Vectors on the Decision Boundary

Figure 11 illustrates the effect of removing a support vector (point A) on the decision boundary and margin position. This demonstrates a key aspect of SVM: only support vectors influence the decision boundary; all other samples do not affect its placement.

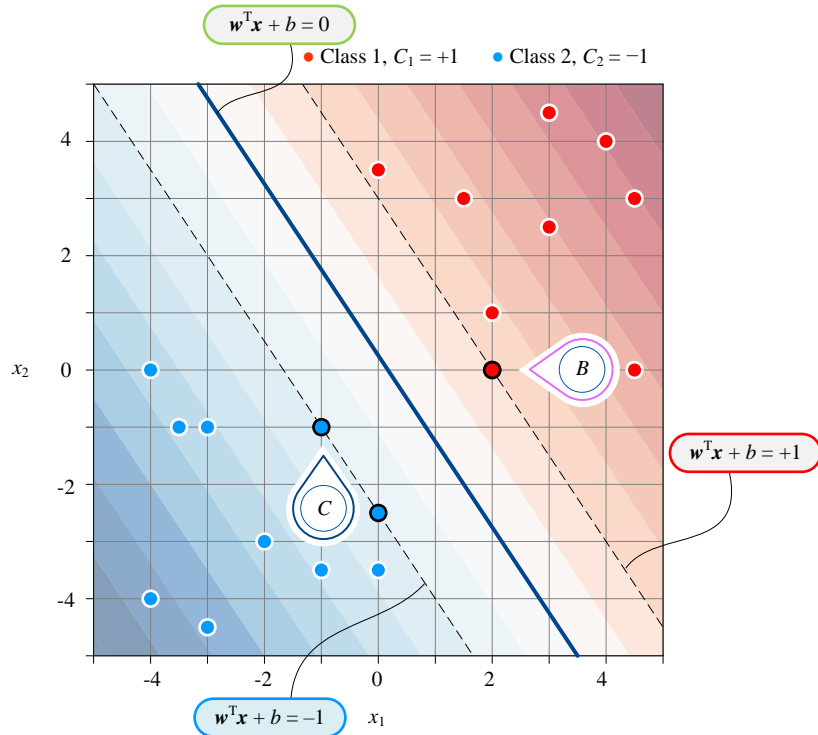


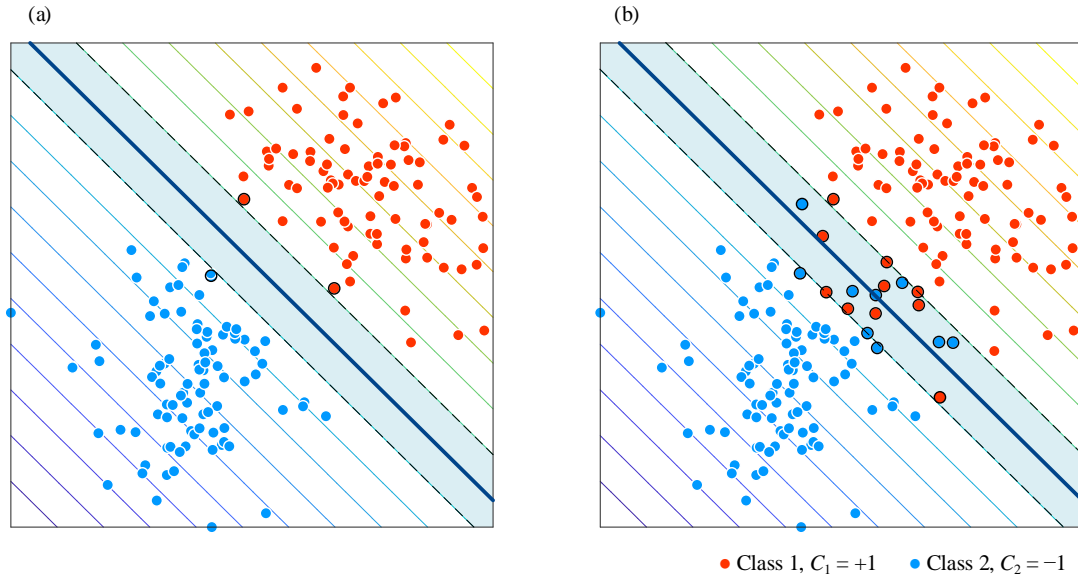
Figure 11. Removing the support vector at point A causes the decision boundary and margin to shift positions. Figures generated by Ch18\_01\_Support\_Vector\_Machine.ipynb.

## 18.5 Soft Margin: Handling Overlapping and Noisy Data

### 18.5.1 From Hard to Soft: Motivation for Relaxation

The first section of this chapter introduced SVM and explained how the soft margin approach is used to handle data that are not linearly separable—i.e., data that cannot be perfectly divided by a straight line or hyperplane.

In contrast, the hard margin method applies to linearly separable data, where a clear, distinct boundary can separate classes without error, as illustrated in Figure 12 (a). When data are more complex or overlapping, as shown in Figure 12 (b), the soft margin method is adopted to allow some classification errors for better generalization.



**Figure 12.** Hard margin SVM is suitable for perfectly separable data, while soft margin SVM handles overlapping data by allowing classification errors to improve generalization.

### 18.5.2 Slack Variables and Penalty Parameter

The core idea behind the soft margin SVM is to sacrifice perfect classification accuracy for a wider margin, enabling the model to tolerate some misclassified points while maintaining robustness. Two critical parameters govern the soft margin formulation:

- ◀ The slack variable (denoted by  $\xi$ ) quantifies the degree to which individual data points violate the margin constraints.
- ◀ The penalty parameter  $C$  controls the trade-off between maximizing the margin width and minimizing classification errors.

The slack variables introduce a "relaxation" to the strict margin boundaries, as depicted schematically in Figure 13.

For samples with label  $+1$ , the constraint becomes

$$(\mathbf{w} \cdot \mathbf{x} + b) \geq 1 - \xi \quad (28)$$

For samples with label  $-1$ , the constraint becomes

$$(\mathbf{w} \cdot \mathbf{x} + b) \leq -1 + \xi \quad (29)$$

As shown in Figure 13, when  $\xi = 0$ , data points lie on or beyond the correct side of the margin, indicating correct classification within the margin boundaries. When  $\xi > 0$ , points fall within the margin or are misclassified.

In Figure 13, the red band illustrates areas where large slack variables appear (more margin violation), while the blue band shows smaller slack values, meaning fewer violations.

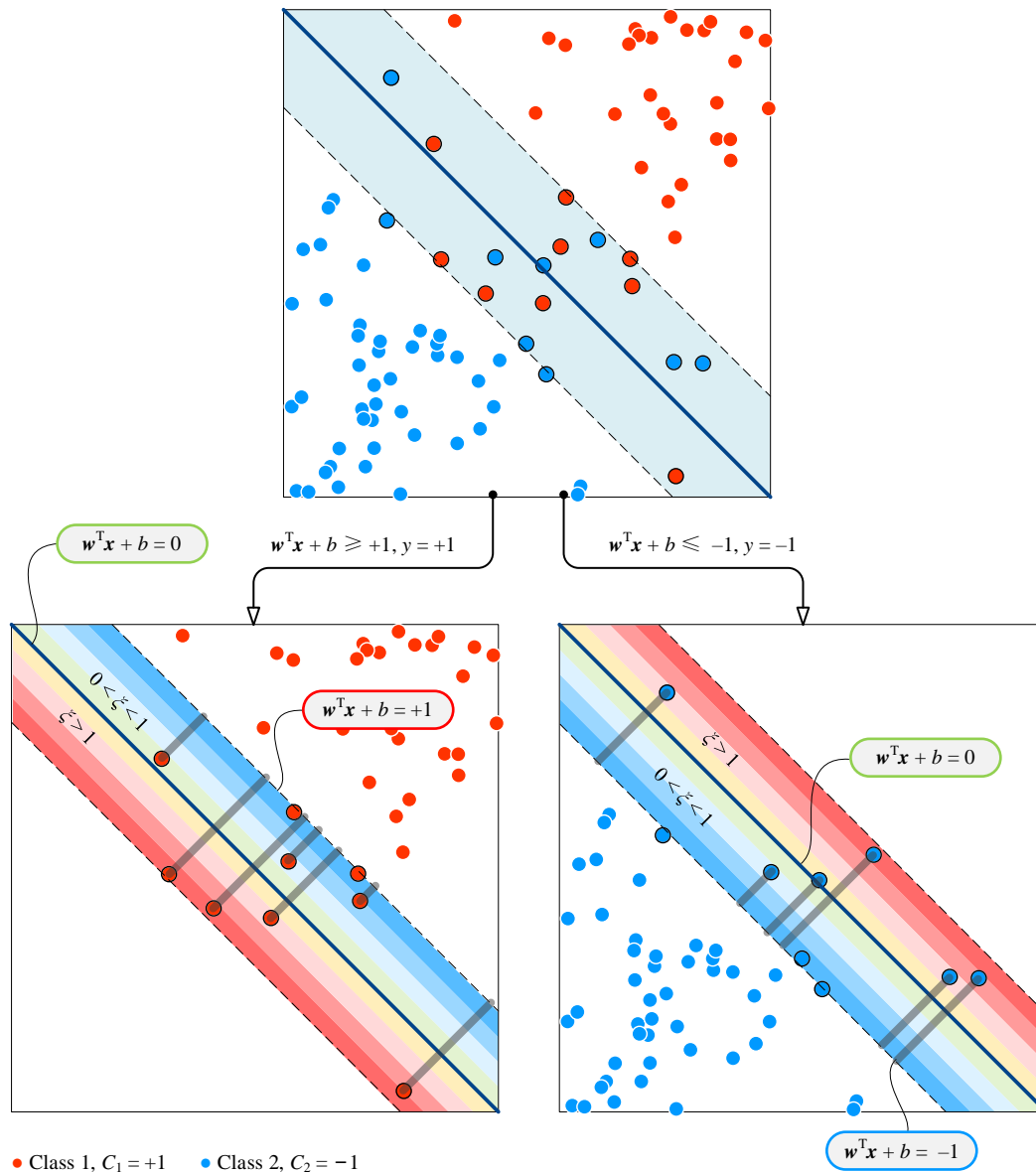


Figure 13. Colored bands indicate different levels of margin violation, demonstrating how the model handles non-linearly separable data using slack variables.

### 18.5.3 Soft-Margin SVM in Practice

Figure 14 illustrates how the support vectors, decision boundary, and margin width change as the penalty parameter  $C$  varies in a soft-margin SVM.

Larger values of  $C$  (e.g., 1) enforce a stricter penalty on margin violations, leading to a narrower margin and fewer misclassified training points, but potentially lower generalization.

Smaller values of  $C$  allow greater flexibility, tolerating more violations of the margin to achieve a wider margin and better generalization, though possibly at the cost of increased training error.

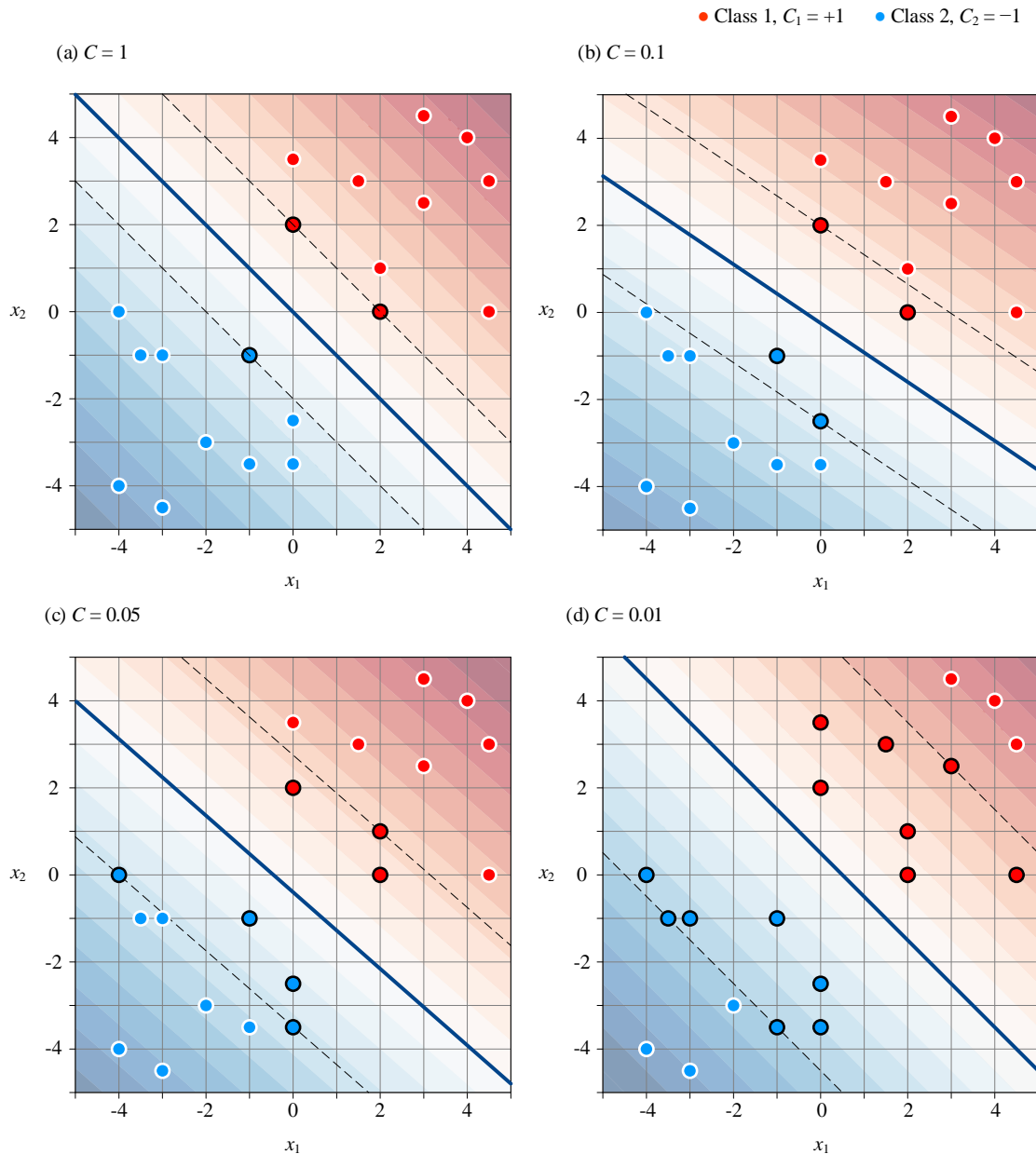


Figure 14. Effect of penalty parameter  $C$  on the SVM decision boundary and margins. Figures generated by Ch18\_01\_Support\_Vector\_Machine.ipynb.

## 18.6 Conclusion

The goal of a Support Vector Machine (SVM) is to identify the optimal hyperplane that best separates two classes of data in a feature space. This hyperplane is determined by solving an optimization problem that seeks to maximize the margin between the two classes while satisfying certain constraints.

There are two types of SVM formulations: hard-margin and soft-margin. The hard-margin SVM assumes that the data is linearly separable with no classification errors, whereas the soft-margin SVM allows for some misclassifications by introducing slack variables, making it suitable for non-separable or noisy data.

In the next chapter, we will explore the kernel trick used in Support Vector Machines. The kernel trick enables SVM to handle nonlinear data by implicitly mapping inputs into high-dimensional spaces. This allows the model to find linear decision boundaries in complex problems without explicitly computing the transformation, improving flexibility and performance.