

An Introduction to Support Vector Machines (SVMs)

Finding the Fattest Separator

Core Concept

The goal in classification problems (with linearly separable data) is to find the **best hyperplane** that separates two classes with the **largest possible margin** (distance between the hyperplane and the closest data points). This approach is the foundation of **Support Vector Machines (SVMs)**.

Key Terms

- **Hyperplane:** A decision boundary that separates classes in space.
 - Defined by $\mathbf{w}^\top \mathbf{x} + b = 0$
 - \mathbf{w} : Weight vector (normal to the hyperplane)
 - b : Bias term shifting the hyperplane
 - **Margin:** The distance between the hyperplane and the nearest data point.
 - Larger margin \Rightarrow better generalization and lower test error.
 - **Support Vectors:** The data points closest to the hyperplane. They “support” the boundary.
-

Detailed Explanation

1. Finding the Fattest Separator

- Among all possible separating lines or hyperplanes, the one with the **largest margin** is preferable because it reduces the chance of misclassifying unseen data.

- The margin is given by $\frac{1}{\|\mathbf{w}\|}$, meaning we want to **minimize** $\|\mathbf{w}\|$ to maximize the margin.

2. Why Fattest is Better

- A wider margin ensures the classifier is robust to small variations or noise in data.
- The **out-of-sample error (E_out)** decreases with a larger margin because the model generalizes better.

3. Mathematical Representation

- All points must satisfy:

$$y_n(\mathbf{w}^\top \mathbf{x}_n + b) > 0$$

- The closest point satisfies:

$$y_n(\mathbf{w}^\top \mathbf{x}_n + b) = 1$$

- The margin is:

$$\frac{1}{\|\mathbf{w}\|}$$

4. Geometry of the Problem

- The weight vector \mathbf{w} is perpendicular to the hyperplane.
- The distance from a point \mathbf{x} to the hyperplane is:

$$\frac{|\mathbf{w}^\top \mathbf{x} + b|}{\|\mathbf{w}\|}$$

5. What if Data is Not Linearly Separable?

- We can use **non-linear decision boundaries** or **kernel methods** to transform the data into higher dimensions where it becomes linearly separable.

Important Intuitions

- A **larger margin** helps the model be more robust and generalize better.

- The geometry of the problem is crucial — the hyperplane is always orthogonal to the weight vector.
 - The model's complexity and data noise can affect how tight or wide the margin should be.
 - SVMs are powerful even for complex, non-linear datasets when extended with kernels.
-

Soft margin SVC

1. Problem with Hard Margin SVM

- In hard margin SVM, we require:

$$y_n(w^T x_n + b) \geq 1 \quad \text{for all } n$$

- This works **only if the data is perfectly separable**.
 - But in real life, data often overlaps → **no perfect separator exists**.
-

2. Introducing slack variables ξ_n

- To allow flexibility, we add slack variables $\xi_n \geq 0$.
- Modified condition:

$$y_n(w^T x_n + b) \geq 1 - \xi_n$$

- If $\xi_n = 0$: point is correctly classified outside margin.
- If $0 < \xi_n < 1$: point is inside margin but still correctly classified.
- If $\xi_n > 1$: point is misclassified.

So, **slack = margin violation**.

3. Optimization Objective

We now balance:

1. **Large margin** (small $\|w\|$)
2. **Few violations** (small $\sum \xi_n$)

The new optimization problem:

$$\min_{w,b,\xi} \quad \frac{1}{2} \|w\|^2 + C \sum_{n=1}^N \xi_n$$

subject to:

$$y_n(w^T x_n + b) \geq 1 - \xi_n, \quad \xi_n \geq 0$$

- First term: maximizes margin.
 - Second term: penalizes violations.
 - $C > 0$: regularization parameter (trade-off).
-

4. Role of C

- **Small C:**
 - Margin violations (slack) are allowed.
 - Wider margin, but more misclassifications.
 - Model is more tolerant → better generalization.
 - **Large C:**
 - Violations are heavily penalized → slack must be small.
 - Forces classifier to correctly classify almost all training points.
 - May overfit.
-

5. Intuition

- Soft margin SVM is a compromise:

- **Too strict (hard margin):** can fail when data is noisy.
 - **Too lenient (very small C):** may underfit.
 - The hyperparameter C is tuned (often via cross-validation) to balance bias-variance.
-

Concise Note Version:

- Hard margin fails if data isn't linearly separable.
 - Add slack variables $\xi_n \rightarrow$ allow margin violations.
 - New condition: $y_n(w^T x_n + b) \geq 1 - \xi_n$.
 - Objective: minimize $\frac{1}{2} \|w\|^2 + C \sum \xi_n$.
 - C controls trade-off:
 - Small C \rightarrow wider margin, more violations, better generalization.
 - Large C \rightarrow narrower margin, fewer violations, risk of overfitting.
-