Support Vector Machines - Math Summary

1. Model (Hyperplane)

For d input features and binary labels $y_i \in \{-1, +1\}$:

$$f(x_i) = \mathbf{w}^\top \mathbf{x}_i + b$$

Prediction rule:

$$\hat{y}_i = \text{sign}(f(x_i))$$

Decision boundary:

$$\mathbf{w}^\top \mathbf{x} + b = 0$$

2. Geometric Margin

Distance from a point to the hyperplane:

$$\text{distance} = \frac{|\mathbf{w}^\top \mathbf{x}_i + b|}{\|\mathbf{w}\|}$$

Margin size (hard-margin intuition):

$$\text{Margin} = \frac{2}{\|\mathbf{w}\|}$$

3. Hard-Margin Optimization

If data is linearly separable:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

Subject to:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \quad \forall i$$

4. Soft-Margin Optimization

To allow misclassification, introduce slack variables $\xi_i \geq 0$:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

Subject to:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i$$

5. Hinge Loss Formulation

Equivalent unconstrained formulation:

$$J(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b))$$

6. Subgradient Updates (Primal SGD)

If:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$$

Then:

$$\nabla_{\mathbf{w}} = \mathbf{w}, \qquad \nabla_b = 0$$

Else:

$$\nabla_{\mathbf{w}} = \mathbf{w} - Cy_i\mathbf{x}_i$$

$$\nabla_b = -Cy_i$$

Update rule:

$$\mathbf{w} \leftarrow \mathbf{w} - \eta\nabla_{\mathbf{w}}$$

$$b \leftarrow b - \eta\nabla_b$$

7. Dual Formulation

The dual optimization problem:

$$\max_{\alpha} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j$$

Subject to:

$$0 \leq \alpha_i \leq C$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

8. Kernel Trick

Replace inner product with kernel function:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$

Example: Radial Basis Function (RBF)

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2)$$

9. Decision Function (Kernel SVM)

After optimization:

$$f(\mathbf{x}) = \sum_{i \in SV} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$$

Only support vectors (SV) have $\alpha_i > 0$.

10. Feature Scaling (Standardization)

For each feature:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Scaling ensures geometric consistency in distance-based optimization.

11. Classification Metrics

Accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score:

$$F_1 = 2\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$