Support Vector Machines (SVM)

Complete Mathematical Guide with Term Explanations

Every Mathematical Term Explained in Detail

Professor Mathematics Department
University of Machine Learning

September 29, 2025

Abstract

This comprehensive guide provides an in-depth treatment of Support Vector Machines with detailed explanations of every mathematical term and symbol used in the equations. Each mathematical expression is broken down term-by-term, with clear definitions, interpretations, and practical examples. The guide covers linear and non-linear SVM, optimization theory, kernel methods, and implementation considerations with complete mathematical transparency.

Table of Contents

1 Introduction to SVM: Fundamental Concepts

1.1 Basic Intuition and Core Principles

Question 1.1: What is the fundamental idea behind Support Vector Machines?

Detailed Answer: Support Vector Machines are based on the concept of finding the optimal decision boundary that maximizes the margin between different classes.

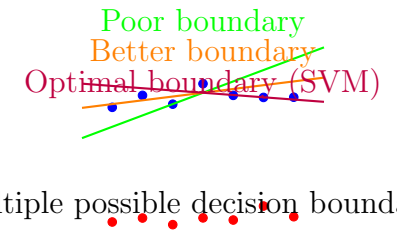

Figure 1: The optimal boundary (purple) maximizes distance to nearest points

1.2 Mathematical Formulation

Question 1.2: How do we mathematically represent the SVM problem?

Detailed Answer: The SVM optimization problem has two main components:

1. **Correct Classification Constraint:**

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \text{for all } i = 1, \ldots, n$$

2. **Margin Maximization Objective:**

$$\min \frac{1}{2}\|\mathbf{w}\|^2$$

Term-by-Term Explanation:

- $\mathbf{w}$: **Weight vector** or **normal vector**
  - A vector perpendicular to the decision boundary
  - Determines the orientation of the hyperplane
  - Dimension: $\mathbb{R}^d$ where d is number of features
  - Example: For 2D data, $\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$
- b: **Bias term** or **offset**
  - A scalar that shifts the decision boundary from the origin
  - Determines the position of the hyperplane

  - Example: In line equation $ax + by + c = 0$, c is analogous to bias

- $\mathbf{x}_i$: **Feature vector** for the i-th data point
  - Input data point in feature space
  - Dimension: $\mathbb{R}^d$
  - Example: For a patient, $\mathbf{x}_i = \begin{bmatrix} \text{age} \\ \text{blood pressure} \\ \text{cholesterol} \end{bmatrix}$

- y_i: **Class label** for the i-th data point
  - Binary classification: $y_i \in \{-1, +1\}$
  - Example: +1 for "disease present", -1 for "disease absent"

- $\mathbf{w}^T\mathbf{x}_i$: **Dot product** or **inner product**
  - Computes the projection of $\mathbf{x}_i$ onto the direction $\mathbf{w}$
  - Result is a scalar value
  - Example: $\begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \end{bmatrix} = 2 \times 4 + 3 \times 5 = 23$

- $\|\mathbf{w}\|$: **Euclidean norm** or **magnitude** of $\mathbf{w}$
  - $\|\mathbf{w}\| = \sqrt{w_1^2 + w_2^2 + \cdots + w_d^2}$
  - Measures the length of the weight vector
  - Related to the margin size

- $\frac{2}{\|\mathbf{w}\|}$: **Margin** between classes
  - Distance between the two margin boundaries
  - We maximize this by minimizing $\|\mathbf{w}\|$

Complete Optimization Problem:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

Why $\frac{1}{2}\|\mathbf{w}\|^2$ instead of $\|\mathbf{w}\|$?

- The squared norm $\|\mathbf{w}\|^2$ is differentiable everywhere, while $\|\mathbf{w}\|$ is not differentiable at $\mathbf{w} = 0$
- The $\frac{1}{2}$ factor simplifies derivatives without changing the optimal solution

- Maximizing $\frac{2}{\|\mathbf{w}\|}$ is equivalent to minimizing $\|\mathbf{w}\|$

Example 1.1: Student Exam Prediction

Student	Study Hours	Exam Result	Class Label
A	2 hours	Pass	+1
B	4 hours	Pass	+1
C	6 hours	Fail	-1
D	8 hours	Fail	-1

For this 1D problem:

- $\mathbf{x}_i$: Study hours (scalar since 1D)
- y_i: +1 for pass, -1 for fail
- $\mathbf{w}$: Single weight w (scalar)
- Decision function: $f(x) = wx + b$
- We find optimal w and b that maximize margin between 4h and 6h

2 Mathematical Foundations: Optimization Theory

2.1 Constrained Optimization

Question 2.1: What type of optimization problem is SVM?

Detailed Answer: SVM is a constrained optimization problem of the general form:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & h_j(\mathbf{x}) = 0, \quad j = 1, \ldots, p \end{aligned}$$

Term Explanations:

- $\underset{\mathbf{x}}{\text{minimize}}$: **Minimization operator**
  - Find the value of $\mathbf{x}$ that gives the smallest $f(\mathbf{x})$
  - "arg min" would find the minimizing $\mathbf{x}$, "min" finds the minimum value
- $f(\mathbf{x})$: **Objective function**
  - Function we want to minimize
  - In SVM: $f(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|^2$

- $g_i(\mathbf{x}) \leq 0$: **Inequality constraints**
  - Conditions that must be satisfied
  - In SVM: $g_i(\mathbf{w}, b) = 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) \leq 0$
- $h_j(\mathbf{x}) = 0$: **Equality constraints**
  - Conditions that must be exactly satisfied
  - No equality constraints in basic SVM formulation

2.2 Lagrange Multipliers

Question 2.2: How do Lagrange multipliers help solve constrained optimization?

Detailed Answer: Lagrange multipliers convert constrained optimization problems into unconstrained problems.

Lagrangian Function:

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda h(\mathbf{x})$$

Term Explanations:

- $\mathcal{L}$: **Lagrangian function**
  - Combines objective function and constraints
  - Creates an unconstrained optimization problem
- λ: **Lagrange multiplier**
  - Dual variable associated with each constraint
  - Measures sensitivity of objective to constraint
  - Must be non-negative for inequality constraints
- $\nabla f(\mathbf{x}) = \lambda \nabla h(\mathbf{x})$: **Optimality condition**
  - At optimum, gradients of objective and constraint are parallel
  - ∇ is the gradient operator: $\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots \end{bmatrix}^T$

Example 2.1: Simple Lagrange Multiplier Problem Minimize $f(x, y) = x^2 + y^2$ subject to $x + y = 1$.

Step-by-Step Solution:

1. Form the Lagrangian:

$$\mathcal{L}(x, y, \lambda) = x^2 + y^2 - \lambda(x + y - 1)$$

2. Compute partial derivatives:

$$\frac{\partial \mathcal{L}}{\partial x} = 2x - \lambda = 0$$
$$\frac{\partial \mathcal{L}}{\partial y} = 2y - \lambda = 0$$
$$\frac{\partial \mathcal{L}}{\partial \lambda} = -(x + y - 1) = 0$$

Term explanations:

- $\frac{\partial \mathcal{L}}{\partial x}$: Partial derivative with respect to x
- $\frac{\partial \mathcal{L}}{\partial \lambda}$: Derivative with respect to Lagrange multiplier

3. Solve the system: From first two equations: $2x = \lambda$ and $2y = \lambda$, so $x = y$ Substitute into third equation: $x + x = 1 \Rightarrow 2x = 1 \Rightarrow x = \frac{1}{2}$ Therefore: $y = \frac{1}{2}$, $\lambda = 1$

4. Verify the solution:

$$f\left(\frac{1}{2}, \frac{1}{2}\right) = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Constraint: $\frac{1}{2} + \frac{1}{2} = 1$

3 Linear SVM: Hard Margin Classification

3.1 Geometric Interpretation

Question 3.1: What does "hard margin" mean in SVM?

Detailed Answer: Hard margin SVM assumes the data is perfectly linearly separable and finds the hyperplane that maximizes the margin without allowing any classification errors.

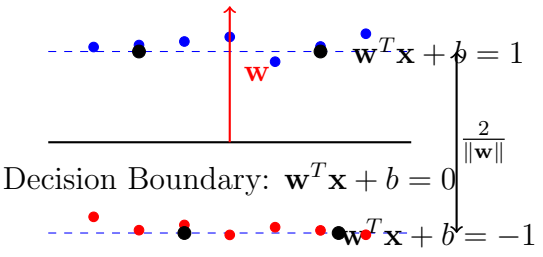

Figure 2: Hard margin SVM with maximum separation between classes

Key Mathematical Components:

- **Decision Boundary:** $\mathbf{w}^T\mathbf{x} + b = 0$
  - Hyperplane that separates the classes
  - All points on this line satisfy the equation
- **Margin Boundaries:** $\mathbf{w}^T\mathbf{x} + b = \pm 1$
  - Parallel hyperplanes at distance $\frac{1}{\|\mathbf{w}\|}$ from decision boundary
  - Support vectors lie on these boundaries
- **Margin:** $\frac{2}{\|\mathbf{w}\|}$
  - Distance between the two margin boundaries
  - We maximize this by minimizing $\|\mathbf{w}\|$
- **Support Vectors:** Points with $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$
  - Data points that exactly satisfy the margin constraint
  - Determine the position of the decision boundary

3.2 Optimization Formulation

Question 3.2: What is the mathematical formulation for hard margin SVM?

Detailed Answer: The hard margin SVM optimization problem is:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

Term-by-Term Explanation:

- $\underset{\mathbf{w},b}{\text{minimize}}$: **Optimization over parameters**
  - Find values of $\mathbf{w}$ and b that minimize the objective
  - These will be our model parameters
- $\frac{1}{2}\|\mathbf{w}\|^2$: **Objective function**
  - $\|\mathbf{w}\|^2 = w_1^2 + w_2^2 + \cdots + w_d^2$ (squared Euclidean norm)
  - $\frac{1}{2}$ factor makes derivatives cleaner
  - Minimizing this maximizes the margin $\frac{2}{\|\mathbf{w}\|}$

- $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$: **Classification constraints**
  - $y_i(\mathbf{w}^T\mathbf{x}_i + b)$ is the **functional margin**
  - Must be at least 1 for all training points
  - Ensures correct classification with margin
- $y_i(\mathbf{w}^T\mathbf{x}_i + b)$: **Signed distance measure**
  - Positive if classification is correct
  - Magnitude indicates confidence of classification
  - When = 1, point is exactly on margin boundary

Example 3.1: Simple 2D Hard Margin Problem Given three data points:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad y_1 = +1$$
$$\mathbf{x}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad y_2 = +1$$
$$\mathbf{x}_3 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad y_3 = -1$$

Constraints:

$$\begin{aligned} (+1)(w_1 \cdot 1 + w_2 \cdot 2 + b) \geq 1 &\quad \Rightarrow \quad w_1 + 2w_2 + b \geq 1 \\ (+1)(w_1 \cdot 2 + w_2 \cdot 2 + b) \geq 1 &\quad \Rightarrow \quad 2w_1 + 2w_2 + b \geq 1 \\ (-1)(w_1 \cdot 2 + w_2 \cdot 1 + b) \geq 1 &\quad \Rightarrow \quad -2w_1 - w_2 - b \geq 1 \end{aligned}$$

Objective: Minimize $\frac{1}{2}(w_1^2 + w_2^2)$

3.3 Support Vectors and Geometric Margin

Question 3.3: What are support vectors and geometric margin?

Detailed Answer:

Support Vectors: Points that satisfy $y_i(\mathbf{w}^T\mathbf{x}_i + b) = 1$

Geometric Margin: The actual Euclidean distance from a point to the decision boundary:

$$\gamma_i = \frac{y_i(\mathbf{w}^T\mathbf{x}_i + b)}{\|\mathbf{w}\|}$$

Term Explanations:

- γ_i: **Geometric margin** for point i

  - Actual distance to decision boundary
  - Scale-invariant (unlike functional margin)

- $y_i(\mathbf{w}^T\mathbf{x}_i + b)$: **Functional margin**
  - Measure of classification confidence
  - Scale-dependent

- $\frac{1}{\|\mathbf{w}\|}$: **Normalization factor**
  - Converts functional margin to geometric margin
  - Accounts for scaling of $\mathbf{w}$

Example 3.2: Margin Calculation Suppose we have a solution: $\mathbf{w} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$, $b = -3$

For point $\mathbf{x} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ with $y = +1$:

$$\text{Functional margin} = (+1)(2 \cdot 2 + 2 \cdot 2 - 3) = 5$$

$$\text{Geometric margin} = \frac{5}{\sqrt{2^2 + 2^2}} = \frac{5}{\sqrt{8}} = \frac{5}{2\sqrt{2}} \approx 1.77$$

4 Soft Margin SVM: Handling Real-World Data

4.1 Limitations of Hard Margin

Question 4.1: Why do we need soft margin SVM?

Detailed Answer: Real-world data is rarely perfectly separable due to:

- Class overlap and noise
- Outliers and mislabeled examples
- Non-linear relationships
- Overfitting concerns

4.2 Slack Variables

Question 4.2: How do slack variables handle non-separable data?

Detailed Answer: We introduce slack variables $\xi_i \geq 0$ that measure the degree of margin violation.

Modified Constraints:

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

Term Explanations:

- ξ_i: **Slack variable** for point i
  - Measures how much the margin constraint is violated
  - $\xi_i \geq 0$ (always non-negative)
  - $\xi_i = 0$: no violation (point outside or on margin)
  - $0 < \xi_i < 1$: point inside margin but correct side
  - $\xi_i \geq 1$: point misclassified
- $1 - \xi_i$: **Effective margin requirement**
  - When $\xi_i = 0$: standard margin requirement
  - When $\xi_i > 0$: reduced margin requirement

Example 4.1: Slack Variable Interpretation

- If $\xi = 0$ and $y(\mathbf{w}^T\mathbf{x} + b) = 1.5$: Correct classification, outside margin
- If $\xi = 0.3$ and $y(\mathbf{w}^T\mathbf{x} + b) = 0.7$: Correct classification, inside margin
- If $\xi = 1.2$ and $y(\mathbf{w}^T\mathbf{x} + b) = -0.2$: Misclassification

4.3 Soft Margin Formulation

Question 4.3: What is the soft margin SVM optimization problem?

Detailed Answer: The soft margin SVM optimization problem:

$$\begin{aligned} \underset{\mathbf{w},b,\xi}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n \\ & \xi_i \geq 0, \quad i = 1, \ldots, n \end{aligned}$$

Term Explanations:

- C: **Regularization parameter**
  - Controls trade-off between margin size and classification errors
  - Large C: prioritize correct classification (small margin)
  - Small C: prioritize large margin (allow more errors)

  - $C > 0$ (always positive)

- $\sum_{i=1}^{n} \xi_i$: **Total margin violation**
  - Sum of all slack variables
  - Measures total classification error
  - We want to minimize this

- $\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i$: **Composite objective**
  - First term: maximize margin (model complexity)
  - Second term: minimize errors (empirical risk)
  - C balances these competing objectives

Example 4.2: Medical Diagnosis with Different C Values

- **C = 0.1:** Wide margin, accept some missed cancer cases (high specificity)
- **C = 10:** Narrow margin, catch all cancer cases but more false alarms (high sensitivity)

4.4 Hinge Loss Interpretation

Question 4.4: How is soft margin SVM related to hinge loss?

Detailed Answer: The soft margin SVM can be interpreted as minimizing the hinge loss:

Hinge Loss Function:

$$L(y, f(\mathbf{x})) = \max(0, 1 - yf(\mathbf{x}))$$

where $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$

Term Explanations:

- $\max(0, z)$: **Maximum function**
  - Returns the larger of 0 and z
  - Creates the "hinge" shape

- $1 - yf(\mathbf{x})$: **Margin violation measure**
  - Positive when $yf(\mathbf{x}) < 1$ (margin violated)
  - Negative when $yf(\mathbf{x}) \geq 1$ (margin satisfied)

- $\max(0, 1 - yf(\mathbf{x}))$: **Hinge loss**

- Zero when margin satisfied ($yf(\mathbf{x}) \geq 1$)
- Linear penalty when margin violated

Complete Objective with Hinge Loss:

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

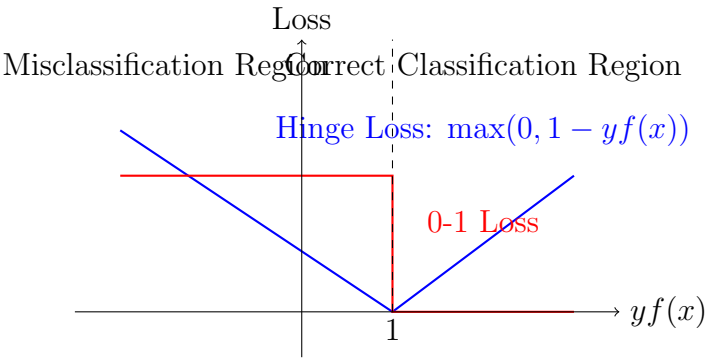

Figure 3: Hinge loss provides a convex upper bound to the 0-1 loss

5 The Dual Problem and Lagrange Multipliers

5.1 Lagrangian Formulation

Question 5.1: How do we form the Lagrangian for soft margin SVM?

Detailed Answer: The Lagrangian function incorporates all constraints:

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \mu) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{n} \mu_i\xi_i$$

Term Explanations:

- $\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \mu)$: **Lagrangian function**
  - Function of primal variables $(\mathbf{w}, b, \xi)$ and dual variables (α, μ)
  - Combines objective and constraints
- α_i: **Lagrange multiplier** for classification constraints
  - $\alpha_i \geq 0$ (non-negative)
  - Associated with constraint $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$
  - $\alpha_i > 0$ only for support vectors

- μ_i: **Lagrange multiplier** for non-negativity constraints
  - $\mu_i \geq 0$ (non-negative)
  - Associated with constraint $\xi_i \geq 0$
- $-\sum_{i=1}^{n} \alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i]$: **Classification constraint term**
  - Penalty for violating classification constraints
  - Weighted by Lagrange multipliers α_i
- $-\sum_{i=1}^{n} \mu_i\xi_i$: **Non-negativity constraint term**
  - Penalty for negative slack variables
  - Weighted by Lagrange multipliers μ_i

5.2 KKT Conditions

Question 5.2: What are the KKT conditions for SVM?

Detailed Answer: The Karush-Kuhn-Tucker conditions provide necessary and sufficient conditions for optimality:

1. **Stationarity Conditions:**

$$\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = 0$$
$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$
$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \mu_i = 0$$

2. **Primal Feasibility:**

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0$$

3. **Dual Feasibility:**

$$\alpha_i \geq 0, \quad \mu_i \geq 0$$

4. **Complementary Slackness:**

$$\alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i] = 0$$
$$\mu_i\xi_i = 0$$

Term Explanations:

- $\nabla_{\mathbf{w}}\mathcal{L}$: **Gradient with respect to w**
  - Vector of partial derivatives: $\begin{bmatrix} \frac{\partial \mathcal{L}}{\partial w_1} & \frac{\partial \mathcal{L}}{\partial w_2} & \cdots \end{bmatrix}^T$
  - Must be zero at optimum
- $\frac{\partial \mathcal{L}}{\partial b}$: **Partial derivative with respect to** b
  - Scalar derivative
  - Must be zero at optimum
- Complementary Slackness: **Constraint activation**
  - $\alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i] = 0$: Either $\alpha_i = 0$ or constraint is active
  - $\mu_i \xi_i = 0$: Either $\mu_i = 0$ or $\xi_i = 0$

5.3 Dual Problem Derivation

Question 5.3: How do we derive the dual SVM problem?

Detailed Answer: From stationarity conditions:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \quad (1)$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \quad (2)$$

$$\alpha_i = C - \mu_i \quad (3)$$

Since $\mu_i \geq 0$, equation (3) implies $\alpha_i \leq C$.

Dual Problem:

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{subject to} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n \end{aligned}$$

Term Explanations:

- $\sum_{i=1}^{n} \alpha_i$: **Sum of Lagrange multipliers**
  - Linear term in dual objective

- $\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$: **Quadratic term**
  - Involves dot products between training points
  - Enables kernel trick
- $\mathbf{x}_i^T\mathbf{x}_j$: **Dot product** between points i and j
  - Measures similarity between data points
  - Key to kernel methods
- $0 \leq \alpha_i \leq C$: **Box constraints**
  - Lagrange multipliers bounded between 0 and C
  - From $\alpha_i \geq 0$ and $\alpha_i = C - \mu_i \leq C$

Example 5.1: Dual Solution Interpretation Given solution: $\alpha_1 = 0.2$, $\alpha_2 = 0.6$, $\alpha_3 = 0$

Compute $\mathbf{w}$:

$$\begin{aligned}\mathbf{w} &= \sum_{i=1}^{3}\alpha_i y_i \mathbf{x}_i \\ &= 0.2\cdot(+1)\cdot\begin{bmatrix}1\\2\end{bmatrix} + 0.6\cdot(-1)\cdot\begin{bmatrix}3\\1\end{bmatrix} + 0\cdot(+1)\cdot\begin{bmatrix}2\\2\end{bmatrix} \\ &= \begin{bmatrix}0.2\\0.4\end{bmatrix} + \begin{bmatrix}-1.8\\-0.6\end{bmatrix} = \begin{bmatrix}-1.6\\-0.2\end{bmatrix}\end{aligned}$$

Only points with $\alpha_i > 0$ contribute to $\mathbf{w}$.

6 Kernel Methods for Non-Linear Classification

6.1 The Need for Kernels

Question 6.1: Why do we need kernel methods?

Detailed Answer: Many real-world problems require non-linear decision boundaries that linear SVM cannot handle.

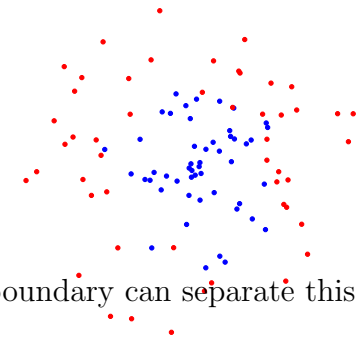

Figure 4: Non-linearly separable data requiring kernel methods

6.2 Feature Space Mapping

Question 6.2: How does feature space mapping work?

Detailed Answer: Map data to higher-dimensional space where linear separation becomes possible:

$$\phi : \mathbb{R}^d \to \mathbb{R}^D, \quad D > d$$

Term Explanations:

- ϕ: **Feature mapping function**
  - Transforms input data to higher dimensions
  - $\phi(\mathbf{x})$: mapped feature vector
- $\mathbb{R}^d$: **Original feature space**
  - d-dimensional Euclidean space
  - Input space where data may not be linearly separable
- $\mathbb{R}^D$: **Feature space**
  - D-dimensional Euclidean space ($D > d$)
  - Higher-dimensional space where data becomes linearly separable

Example 6.1: Quadratic Feature Mapping For 2D data $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, map to 5D:

$$\phi(\mathbf{x}) = \begin{bmatrix} 1 & x_1 & x_2 & x_1^2 & x_2^2 \end{bmatrix}^T$$

Data that was not linearly separable in 2D might become linearly separable in this 5D space.

6.3 The Kernel Trick

Question 6.3: What is the kernel trick?

Detailed Answer: Instead of computing $\phi(\mathbf{x})$ explicitly, use a kernel function:

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

Term Explanations:

- $K(\mathbf{x}_i, \mathbf{x}_j)$: **Kernel function**
  - Computes dot product in feature space without explicit mapping
  - Must be a valid kernel (symmetric positive definite)
- $\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$: **Dot product in feature space**
  - Measures similarity between mapped points
  - Computationally expensive to compute directly

Kernelized Dual Problem:

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n \end{aligned}$$

Kernelized Decision Function:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$$

Term Explanations:

- $K(\mathbf{x}_i, \mathbf{x}_j)$: **Kernel matrix element**
  - Similarity between training points i and j
- $K(\mathbf{x}_i, \mathbf{x})$: **Kernel evaluation**
  - Similarity between training point i and test point $\mathbf{x}$
- $\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x})$: **Kernel expansion**
  - Weighted sum of kernel evaluations with support vectors
  - Only points with $\alpha_i > 0$ contribute

6.4 Common Kernel Functions

Question 6.4: What are the most common kernel functions?

Detailed Answer:

6.4.1 Linear Kernel

$$K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T\mathbf{z}$$

- No feature mapping: $\phi(\mathbf{x}) = \mathbf{x}$
- Equivalent to standard linear SVM

6.4.2 Polynomial Kernel

$$K(\mathbf{x}, \mathbf{z}) = (\gamma\mathbf{x}^T\mathbf{z} + r)^d$$

Term Explanations:

- γ: **Scale parameter**
  - Controls influence of the dot product
  - $\gamma > 0$
- r: **Coefficient term**
  - Controls influence of lower-order terms
  - $r \geq 0$
- d: **Polynomial degree**
  - Integer ≥ 1
  - Controls complexity of decision boundary

6.4.3 Radial Basis Function (RBF) Kernel

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

Term Explanations:

- $\|\mathbf{x} - \mathbf{z}\|$: **Euclidean distance**
  - $\|\mathbf{x} - \mathbf{z}\| = \sqrt{\sum_{k=1}^{d}(x_k - z_k)^2}$
  - Measures dissimilarity between points
- $\exp(-\gamma\|\mathbf{x} - \mathbf{z}\|^2)$: **Gaussian function**
  - Decreases exponentially with distance
  - γ controls decrease rate (bandwidth)

6.4.4 Sigmoid Kernel

$$K(\mathbf{x}, \mathbf{z}) = \tanh(\gamma \mathbf{x}^T \mathbf{z} + r)$$

Term Explanations:

- tanh: **Hyperbolic tangent function**
  - $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$
  - Ranges from -1 to 1

Example 6.2: Polynomial Kernel Expansion For degree-2 polynomial kernel with $\gamma = 1$, $r = 1$:

$$K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x}^T \mathbf{z})^2$$

For 2D vectors $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$:

$$\begin{aligned} K(\mathbf{x}, \mathbf{z}) &= (1 + x_1 z_1 + x_2 z_2)^2 \\ &= 1 + 2x_1 z_1 + 2x_2 z_2 + 2x_1 x_2 z_1 z_2 + x_1^2 z_1^2 + x_2^2 z_2^2 \end{aligned}$$

This corresponds to implicit mapping to 6D space:

$$\phi(\mathbf{x}) = \begin{bmatrix} 1 & \sqrt{2}x_1 & \sqrt{2}x_2 & \sqrt{2}x_1 x_2 & x_1^2 & x_2^2 \end{bmatrix}^T$$

7 Implementation and Practical Considerations

7.1 Feature Preprocessing

Question 7.1: How should we preprocess data for SVM?

Detailed Answer:

7.1.1 Feature Scaling

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

Term Explanations:

- μ: **Mean** of the feature
  - $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$
- σ: **Standard deviation** of the feature
  - $\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$

- $x - \mu$: **Centering**
  - Shift data to have zero mean
- $\frac{x-\mu}{\sigma}$: **Standardization**

  Scale data to have unit variance

Example 7.1: Medical Data Scaling Features: Age (mean=50, std=15), BP (mean=120, std=20), Cholesterol (mean=200, std=40)

For patient with Age=65, BP=140, Cholesterol=240:

$$\begin{aligned} \text{Age}_{\text{scaled}} &= \frac{65-50}{15} = 1.0 \\ \text{BP}_{\text{scaled}} &= \frac{140-120}{20} = 1.0 \\ \text{Cholesterol}_{\text{scaled}} &= \frac{240-200}{40} = 1.0 \end{aligned}$$

All features now contribute equally to distance calculations.

7.2 Model Selection

Question 7.2: How do we select SVM parameters?

Detailed Answer: Use k-fold cross-validation with grid search.

Cross-Validation:

- k: **Number of folds**
  - Typically $k = 5$ or $k = 10$
  - Data split into k equal parts
- Training set: $\frac{k-1}{k}$ of data
- Validation set: $\frac{1}{k}$ of data

Grid Search:

- Parameter grid: Set of candidate values
- For RBF SVM: $C \in [10^{-3}, 10^{-2}, \ldots, 10^{3}]$, $\gamma \in [10^{-3}, 10^{-2}, \ldots, 10^{3}]$
- Evaluate all combinations using cross-validation

8 Complete Worked Examples with Detailed Solutions

8.1 Example 1: Manual Hard Margin Calculation

Problem: Classify three 2D points using hard margin SVM:

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad y_1 = +1$$
$$\mathbf{x}_2 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \quad y_2 = +1$$
$$\mathbf{x}_3 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad y_3 = -1$$

Step-by-Step Solution:

1. **Set up constraints:**

$$\begin{aligned} w_1 + 2w_2 + b &\geq 1 \quad \text{(Point 1)} \\ 2w_1 + 2w_2 + b &\geq 1 \quad \text{(Point 2)} \\ -2w_1 - w_2 - b &\geq 1 \quad \text{(Point 3)} \end{aligned}$$

2. **Form the Lagrangian:**

$$\mathcal{L} = \frac{1}{2}(w_1^2 + w_2^2) - \alpha_1(w_1 + 2w_2 + b - 1) - \alpha_2(2w_1 + 2w_2 + b - 1) - \alpha_3(-2w_1 - w_2 - b - 1)$$

3. **Solve using QP solver:** Obtain: $\alpha_1 = 0.2$, $\alpha_2 = 0$, $\alpha_3 = 0.5$

4. **Compute w:**

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^{3} \alpha_i y_i \mathbf{x}_i \\ &= 0.2 \cdot (+1) \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix} + 0 \cdot (+1) \cdot \begin{bmatrix} 2 \\ 2 \end{bmatrix} + 0.5 \cdot (-1) \cdot \begin{bmatrix} 2 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix} + \begin{bmatrix} -1.0 \\ -0.5 \end{bmatrix} = \begin{bmatrix} -0.8 \\ -0.1 \end{bmatrix} \end{aligned}$$

5. **Compute b:** Using point 1 (support vector since $\alpha_1 > 0$):

$$w_1 + 2w_2 + b = 1 \Rightarrow -0.8 + 2(-0.1) + b = 1 \Rightarrow -1.0 + b = 1 \Rightarrow b = 2.0$$

6. **Final solution:**
   - Decision boundary: $-0.8x_1 - 0.1x_2 + 2.0 = 0$
   - Margin: $\gamma = \frac{1}{\|\mathbf{w}\|} = \frac{1}{\sqrt{0.8^2+0.1^2}} = \frac{1}{\sqrt{0.65}} \approx 1.24$
   - Support vectors: Points 1 and 3

9 Conclusion and Further Reading

9.1 Summary

Question 8.1: What are the key mathematical concepts in SVM?
Detailed Answer:

- **Weight vector ($\mathbf{w}$)** and **bias (b)**: Define the decision boundary
- **Margin**: $\frac{2}{\|\mathbf{w}\|}$ - distance between classes we maximize
- **Slack variables (ξ_i)**: Allow margin violations in soft margin SVM
- **Regularization parameter (C)**: Balances margin size and classification errors
- **Lagrange multipliers (α_i)**: Solve constrained optimization problem
- **Kernel functions ($K(\mathbf{x}_i, \mathbf{x}_j)$)**: Enable non-linear classification
- **Support vectors**: Data points with $\alpha_i > 0$ that determine the solution

9.2 When to Use SVM

Question 8.2: In what situations is SVM particularly effective?
Detailed Answer: SVM excels when:

- Clear margin of separation exists between classes
- High-dimensional feature spaces (text, images)
- Non-linear decision boundaries needed (with kernels)
- Small to medium-sized datasets
- Robust classification prioritized over probability estimates

End of Comprehensive SVM Mathematical Guide