Support Vector Machines - Derivations & Proofs

Paul F. Roysdon, Ph.D.

Contents

1 Mathematical Derivations & Proofs **1**
1.1 Introduction . . . 1
1.2 Data and Notation . . . 1
1.3 Model Formulation and Geometric Margin . . . 1
1.4 Hard-Margin SVM: Primal Problem (Separable Case) . . . 2
1.5 Lagrangian and Dual Derivation . . . 2
1.6 KKT Conditions and Support Vectors . . . 2
1.7 Soft-Margin SVM: Primal, Dual, and Hinge-Loss ERM . . . 3
1.8 Kernelization (Nonlinear SVM) . . . 3
1.9 Variants and Practical Notes . . . 3
1.10 Summary of Variables and Dimensions . . . 4
1.11 Summary . . . 4

1 Mathematical Derivations & Proofs

1.1 Introduction

Support Vector Machines are supervised learning models used for classification (and regression) tasks. In the case of classification, the goal is to find a hyperplane that maximizes the margin between data points of two classes. In this section, we derive the optimization formulation for the *hard-margin SVM* (assuming data is linearly separable) and *soft-margin SVM*, show the Lagrangian formulation, derive the corresponding dual problem, and explain the steps involved. All variables are introduced with their dimensions and properties.

1.2 Data and Notation

We are given n labeled samples

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \qquad \mathbf{x}_i \in \mathbb{R}^d, \;\; y_i \in \{-1, +1\}.$$

A linear decision function takes the form

$$f(\mathbf{x}) \;=\; \mathbf{w}^\top \mathbf{x} + b,$$

with $\mathbf{w} \in \mathbb{R}^d$ (weights) and $b \in \mathbb{R}$ (bias). We predict $\hat{y} = \operatorname{sign}\big(f(\mathbf{x})\big)$.

1.3 Model Formulation and Geometric Margin

The signed distance (geometric margin) of $(\mathbf{x}_i, y_i)$ to the hyperplane $\mathbf{w}^\top \mathbf{x} + b = 0$ is

$$\gamma_i \;=\; \frac{y_i\big(\mathbf{w}^\top \mathbf{x}_i + b\big)}{\|\mathbf{w}\|_2}. \tag{1}$$

The sample margin is $\gamma = \min_i \gamma_i$. By scaling $(\mathbf{w}, b)$ so that

$$y_i\big(\mathbf{w}^\top \mathbf{x}_i + b\big) \;\geq\; 1 \quad \forall i, \tag{2}$$

the margin equals $\gamma = 1/\|\mathbf{w}\|_2$. Thus, maximizing the margin is equivalent to minimizing $\|\mathbf{w}\|_2^2$.

1.4 Hard-Margin SVM: Primal Problem (Separable Case)

For linearly separable data, the *primal* problem is

$$\begin{aligned} \min_{\mathbf{w},\, b} \quad & \tfrac{1}{2}\,\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i\big(\mathbf{w}^\top \mathbf{x}_i + b\big) \geq 1, \quad i = 1, \ldots, n. \end{aligned} \tag{3}$$

1.5 Lagrangian and Dual Derivation

Introduce Lagrange multipliers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^\top$ with $\alpha_i \geq 0$ for the constraints Eqn. (2). The Lagrangian is

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) \;=\; \tfrac{1}{2}\|\mathbf{w}\|_2^2 \;-\; \sum_{i=1}^{n} \alpha_i \Big(y_i\big(\mathbf{w}^\top \mathbf{x}_i + b\big) - 1 \Big). \tag{4}$$

Stationarity (Step 1).

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \;\Rightarrow\; \mathbf{w} \;=\; \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i, \tag{5}$$

$$\frac{\partial L}{\partial b} = 0 \;\Rightarrow\; \sum_{i=1}^{n} \alpha_i y_i \;=\; 0. \tag{6}$$

Dual Objective (Step 2). Substitute Eqns. (5)–(6) into Eqn. (4) to eliminate $(\mathbf{w}, b)$:

$$\max_{\boldsymbol{\alpha}} \quad D(\boldsymbol{\alpha}) \;=\; \sum_{i=1}^{n} \alpha_i \;-\; \tfrac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j\, \mathbf{x}_i^\top \mathbf{x}_j, \quad \text{s.t. } \alpha_i \geq 0, \;\; \sum_{i=1}^{n} \alpha_i y_i = 0. \tag{7}$$

Decision Function and Bias Recovery. With an optimal $\boldsymbol{\alpha}^\star$, the classifier is

$$f(\mathbf{x}) \;=\; \sum_{i=1}^{n} \alpha_i^\star y_i\, \mathbf{x}_i^\top \mathbf{x} \;+\; b, \tag{8}$$

and b can be obtained from any support vector s with $\alpha_s^\star > 0$:

$$b \;=\; y_s - \sum_{j=1}^{n} \alpha_j^\star y_j\, \mathbf{x}_j^\top \mathbf{x}_s. \tag{9}$$

The margin is $1/\|\mathbf{w}\|_2$ with $\mathbf{w}$ given by Eqn. (5).

1.6 KKT Conditions and Support Vectors

At optimality, the KarushKuhnTucker conditions hold:

$$\alpha_i \geq 0, \qquad y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \geq 0, \qquad \alpha_i \big[y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \big] = 0 \quad \forall i.$$

Samples with $\alpha_i > 0$ are *support vectors* and lie on the margin: $y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1$.

1.7 Soft-Margin SVM: Primal, Dual, and Hinge-Loss ERM

For nonseparable data, introduce slack variables $\xi_i \geq 0$:

$$\begin{aligned} \min_{\mathbf{w},\, b,\, \boldsymbol{\xi}} \quad & \tfrac{1}{2}\,\|\mathbf{w}\|_2^2 \;+\; C\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n. \end{aligned} \tag{10}$$

KKT analysis yields box constraints $0 \leq \alpha_i \leq C$ in the dual:

$$\max_{\boldsymbol{\alpha}} \; \sum_{i=1}^{n}\alpha_i - \tfrac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j\, \mathbf{x}_i^\top \mathbf{x}_j, \quad \text{s.t. } 0 \leq \alpha_i \leq C, \;\; \sum_{i=1}^{n}\alpha_i y_i = 0. \tag{11}$$

Eliminating $\boldsymbol{\xi}$ from Eqn. (10) gives the equivalent unconstrained ERM with hinge loss:

$$\min_{\mathbf{w},\, b} \;\; \tfrac{1}{2}\|\mathbf{w}\|_2^2 \;+\; C\sum_{i=1}^{n}\max\big(0,\, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\big). \tag{12}$$

A subgradient of the hinge term for sample i is

$$\partial_{\mathbf{w}} = \begin{cases} -\,y_i\mathbf{x}_i, & y_i(\mathbf{w}^\top \mathbf{x}_i + b) < 1, \\ \{0\}, & y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 1, \end{cases} \qquad \partial_b = \begin{cases} -\,y_i, & y_i(\mathbf{w}^\top \mathbf{x}_i + b) < 1, \\ \{0\}, & y_i(\mathbf{w}^\top \mathbf{x}_i + b) > 1. \end{cases}$$

1.8 Kernelization (Nonlinear SVM)

Because $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$ (representer theorem), the solution depends on inner products $\mathbf{x}_i^\top \mathbf{x}_j$. Replace them by a positive semidefinite kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) \;=\; \boldsymbol{\phi}(\mathbf{x}_i)^\top \boldsymbol{\phi}(\mathbf{x}_j), \qquad K \succeq 0, \tag{13}$$

to obtain the nonlinear decision function

$$f(\mathbf{x}) \;=\; \sum_{i=1}^{n}\alpha_i y_i\, K(\mathbf{x}_i, \mathbf{x}) \;+\; b. \tag{14}$$

Common kernels: linear $K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \mathbf{z}$; polynomial $K = (\mathbf{x}^\top \mathbf{z} + c)^p$; RBF $K = \exp\big(-\|\mathbf{x} - \mathbf{z}\|_2^2/(2\sigma^2)\big)$.

1.9 Variants and Practical Notes

Regularization parameter C. Controls trade-off between margin width and violations (larger $C \Rightarrow$ fewer violations, potentially smaller margin).

Class imbalance. Use class-specific costs C_+, C_-, leading to $0 \leq \alpha_i \leq C_{y_i}$ in Eqn. (11).

ν-SVM. Alternative form

$$\min_{\mathbf{w},b,\boldsymbol{\xi},\rho} \;\; \tfrac{1}{2}\|\mathbf{w}\|_2^2 - \rho + \tfrac{1}{\nu n}\sum_{i=1}^{n}\xi_i \quad \text{s.t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq \rho - \xi_i, \;\; \xi_i \geq 0,$$

where $0 < \nu \leq 1$ lower-bounds the fraction of support vectors and upper-bounds the fraction of margin errors.

Multiclass. One-vs-rest (K classifiers) or one-vs-one ($K(K-1)/2$ classifiers); direct multiclass SVMs also exist.

Optimization. The dual QP Eqn. (11) has one linear equality and box constraints; SMO/coordinate-ascent methods update pairs of α_i efficiently. After solving, recover b via Eqn. (9) using any i with $0 < \alpha_i < C$.

1.10 Summary of Variables and Dimensions

- $\mathbf{x}_i \in \mathbb{R}^d$: feature vector; $y_i \in \{-1, +1\}$: label.
- $\mathbf{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$: hyperplane parameters.
- $\boldsymbol{\alpha} \in \mathbb{R}^n$: Lagrange multipliers; in soft-margin, $0 \le \alpha_i \le C$.
- $\boldsymbol{\xi} \in \mathbb{R}^n_{\ge 0}$: slack variables (soft margin).
- $K(\cdot, \cdot)$: PSD kernel replacing $\mathbf{x}_i^\top \mathbf{x}_j$.

1.11 Summary

From first principles: (i) define and maximize the geometric margin, yielding the hard-margin primal Eqn. (3); (ii) form the Lagrangian, eliminate $(\mathbf{w}, b)$ to get the dual Eqn. (7); (iii) enforce KKT to identify support vectors and compute b Eqn. (9); (iv) handle nonseparable data with the soft-margin primal Eqn. (10), dual Eqn. (11), and hinge-loss ERM Eqn. (12); (v) achieve nonlinear separation via kernelization Eqns. (13)–(14). These steps constitute a complete derivation and working procedure for linear and kernel SVMs.