Support Vector Machine

Primal-to-Dual Derivation

CS229: Machine Learning

Parth Bansal

June 11, 2025

Detailed Lagrangian-duality proof
with support-vector insights

From Primal SVM to Dual: A Detailed Lagrange-Duality Derivation

1 Primal Problem (Hard-Margin SVM)

We are given a labeled training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$, where $x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \{+1, -1\}$. The *primal* (hard-margin) SVM is:

$$\min_{w \in \mathbb{R}^d,\, b \in \mathbb{R}} \quad \frac{1}{2} \|w\|^2 \tag{1}$$

$$\text{s.t.} \quad y^{(i)} \big(w^\top x^{(i)} + b \big) \ \geq\ 1, \quad i = 1, \dots, n. \tag{2}$$

Goal: Find w, b that separate the two classes with maximal margin.

Notation.

- Let $f_0(w, b) = \frac{1}{2}\|w\|^2$ be the objective.
- Let $f_i(w, b) = 1 - y^{(i)}(w^\top x^{(i)} + b)$, so the constraints are $f_i(w, b) \leq 0$, $i = 1, \dots, n$.

2 Lagrangian and Dual Function

2.1 Constructing the Lagrangian

Introduce nonnegative multipliers (dual variables) $\alpha_i \geq 0$ for each constraint $f_i(w, b) \leq 0$. The *Lagrangian* is

$$\mathcal{L}(w, b, \alpha) = f_0(w, b) \ + \ \sum_{i=1}^n \alpha_i\, f_i(w, b) = \frac{1}{2}\, w^\top w + \sum_{i=1}^n \alpha_i \big[\, 1 - y^{(i)}(w^\top x^{(i)} + b) \big].$$

Rearranging signs,

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\, w^\top w - \sum_{i=1}^n \alpha_i \big[y^{(i)}(w^\top x^{(i)} + b) - 1 \big].$$

2.2 Definition of the Dual Function

The *Lagrange dual function* $g(\alpha)$ is defined by minimizing $\mathcal{L}$ over the primal variables (w, b):

$$g(\alpha) = \inf_{w, b}\ \mathcal{L}(w, b, \alpha).$$

Because $\mathcal{L}$ is *convex* in (w, b) for each fixed $\alpha \geq 0$, this infimum is attained by solving the stationary conditions (first-order necessary conditions).

2.3 Stationarity Conditions

$$\nabla_w \mathcal{L} = w \;-\; \sum_{i=1}^{n} \alpha_i \, y^{(i)} \, x^{(i)} \;=\; 0 \quad\Longrightarrow\quad w \;=\; \sum_{i=1}^{n} \alpha_i \, y^{(i)} \, x^{(i)}, \tag{3}$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{n} \alpha_i \, y^{(i)} \;=\; 0 \quad\Longrightarrow\quad \sum_{i=1}^{n} \alpha_i \, y^{(i)} = 0. \tag{4}$$

There is no constraint on (w, b) beyond convexity, so these are sufficient.

3 Evaluating $g(\alpha)$

Substitute the optimal w from (3) back into $\mathcal{L}$:

$$\mathcal{L}^{\star}(\alpha) := \mathcal{L}\big(w(\alpha),\, b(\alpha),\, \alpha\big).$$

Since $\sum_i \alpha_i y^{(i)} = 0$ makes the $-b \sum_i \alpha_i y^{(i)}$ term vanish, we get

$$\mathcal{L}^{\star}(\alpha) = \frac{1}{2} \big(\sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)}\big)^{\top} \big(\sum_{j=1}^{n} \alpha_j y^{(j)} x^{(j)}\big) \;-\; \sum_{i=1}^{n} \alpha_i \, (y^{(i)} w(\alpha)^{\top} x^{(i)} - 1).$$

Expand term by term:

$$\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \, y^{(i)} y^{(j)} \, \langle x^{(i)}, x^{(j)} \rangle \;-\; \sum_{i,j} \alpha_i \alpha_j \, y^{(i)} y^{(j)} \, \langle x^{(j)}, x^{(i)} \rangle \;+\; \sum_{i=1}^{n} \alpha_i.$$

Noting the bilinear form duplicates, the quadratic terms combine to $-\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y^{(i)} y^{(j)} \langle x^{(i)}, x^{(j)} \rangle$. Thus

$$g(\alpha) = \sum_{i=1}^{n} \alpha_i \;-\; \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \, y^{(i)} y^{(j)} \, \langle x^{(i)}, x^{(j)} \rangle.$$

4 The Dual Problem

By definition, the dual is

$$\max_{\alpha \geq 0} \; g(\alpha) \quad \text{s.t.} \quad \sum_{i=1}^{n} \alpha_i \, y^{(i)} = 0.$$

Putting everything together,

$$\max_{\alpha \in \mathbb{R}^n} \quad \sum_{i=1}^{n} \alpha_i \;-\; \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \, y^{(i)} y^{(j)} \, \langle x^{(i)}, x^{(j)} \rangle \tag{5}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i \, y^{(i)} = 0, \quad \alpha_i \geq 0, \; i = 1, \ldots, n. \tag{6}$$

5 Strong Duality and KKT Conditions

- The primal (1)–(2) is convex and there exists a strictly feasible point (Slater's condition), so *strong duality* holds: $\min(\text{primal}) = \max(\text{dual})$.
- The full KKT system is
$$\begin{cases} y^{(i)}(w^\top x^{(i)} + b) - 1 \ \geq 0, \\ \alpha_i \geq 0, \\ \alpha_i \big[y^{(i)}(w^\top x^{(i)} + b) - 1 \big] = 0, \\ \nabla_w \mathcal{L} = 0, \quad \partial_b \mathcal{L} = 0. \end{cases}$$
- *Complementary slackness* $\alpha_i \big[y^{(i)}(w^\top x^{(i)} + b) - 1 \big] = 0$ implies
  - If $\alpha_i > 0$, then $y^{(i)}(w^\top x^{(i)} + b) = 1$: these points *lie on the margin boundary.*
  - These points are the **support vectors**.

Brief Note on Support Vectors

Because $w = \sum_i \alpha_i y^{(i)} x^{(i)}$, *only* those $x^{(i)}$ with $\alpha_i > 0$ appear in the final classifier. All others have $\alpha_i = 0$ and hence *do not* affect $\operatorname{sign}(w^\top x + b)$. In practice, this yields a *sparse* model determined by the few training points closest to the decision boundary.