Duality in Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are powerful supervised learning models used for classification and regression tasks. A fundamental concept in SVMs is duality, which plays a crucial role in optimizing the SVM problem. This document provides an overview of duality in SVMs, explaining its importance and the steps involved in deriving the dual problem.

1. Introduction to SVMs

Support Vector Machines aim to find the hyperplane that best separates data into different classes. The optimal hyperplane maximizes the margin, which is the distance between the hyperplane and the nearest data points from each class, known as support vectors.

Primal Problem

For a linearly separable dataset, the primal problem can be formulated as:

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

subject to:

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

where:

- $\mathbf{w}$ is the weight vector,
- b is the bias term,
- y_i is the class label of the i-th sample,
- $\mathbf{x}_i$ is the i-th sample.

Regularized Primal Problem for Non-Separable Data

To handle non-separable data, slack variables ξ_i are introduced, leading to the following regularized primal problem:

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

subject to:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i$$
$$\xi_i \geq 0, \quad \forall i$$

where C is a regularization parameter controlling the trade-off between maximizing the margin and minimizing the classification error.

2. Duality in SVMs

The concept of duality allows us to transform the primal problem into its dual form. Solving the dual problem can be more efficient, especially when dealing with high-dimensional data or when the kernel trick is applied.

Lagrangian Formulation

To derive the dual problem, we introduce Lagrange multipliers $\alpha_i \geq 0$ for the constraints:

$$\mathrm{L}(\mathbf{w}, b, \xi, \alpha) = \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - (1-\xi_i)] - \sum_{i=1}^{n}\eta_i\xi_i$$

where $\eta_i \geq 0$ are Lagrange multipliers for the non-negativity constraints on ξ_i.

Karush-Kuhn-Tucker (KKT) Conditions

The optimal solution must satisfy the KKT conditions:

1. Stationarity:
   $\frac{\partial \mathrm{L}}{\partial \mathbf{w}} = 0, \quad \frac{\partial \mathrm{L}}{\partial b} = 0, \quad \frac{\partial \mathrm{L}}{\partial \xi_i} = 0$
2. Primal feasibility:
   $y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$
3. Dual feasibility:
   $\alpha_i \geq 0, \quad \eta_i \geq 0$
4. Complementary slackness:
   $\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - (1-\xi_i)] = 0, \quad \eta_i\xi_i = 0$

Dual Problem

By substituting the KKT conditions into the Lagrangian, we eliminate $\mathbf{w}$, b, and ξ_i to obtain the dual problem:

$$\max_\alpha \sum_{i=1}^{n}\alpha_i - \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j(\mathbf{x}_i\cdot\mathbf{x}_j)$$

subject to:

$$\sum_{i=1}^{n}\alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq C, \quad \forall i$$

Solution to the Dual Problem

The dual problem is a quadratic programming problem, which can be solved using optimization techniques like Sequential Minimal Optimization (SMO). The solutions α_i allow us to compute the weight vector $\mathbf{w}$ and the bias term b:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$
$$b = y_k - \sum_{i=1}^{n} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}_k)$$

for any support vector k with $0 < \alpha_k < C$.

3. Kernel Trick

The dual formulation enables the use of the kernel trick, which allows SVMs to perform classification in high-dimensional feature spaces without explicitly computing the coordinates. By replacing the dot product $(\mathbf{x}_i \cdot \mathbf{x}_j)$ with a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$, the SVM can handle non-linear decision boundaries.

Common kernel functions include:

- Linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$
- Polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d$
- Radial Basis Function (RBF): $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$
- Sigmoid: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\kappa \mathbf{x}_i \cdot \mathbf{x}_j + \theta)$