Support Vector Machines (SVM) with Kernel Trick

1 Introduction

Support Vector Machines (SVMs) are supervised learning models used for classification and regression tasks. This chapter explains the mathematical foundations of SVMs, with a focus on kernel functions, and how nonlinear optimization—specifically quadratic programming in convex optimization—is used to solve these problems.

2 Problem Statement

We are given a dataset of n examples:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \tag{1}$$

where $x_i \in \mathbb{R}^d$ are input vectors and $y_i \in \{-1, 1\}$ are labels representing two classes. The goal is to find a function that correctly classifies the data points while maximizing the margin between the two classes.

3 Nonlinear SVM with Kernel Trick

3.1 Mathematical Foundation

For nonlinearly separable data, we map the data into a higher-dimensional space using a function $\phi(x)$:

$$x \mapsto \phi(x) \tag{2}$$

We define a **kernel function** $K(x_i, x_j)$ such that:

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{3}$$

Common kernel functions include:

- Polynomial: $K(x, x') = (x^\top x' + c)^d$
- Radial Basis Function (RBF): $K(x, x') = \exp(-\gamma \|x - x'\|^2)$

3.2 Primal Problem

The primal problem for a nonlinear SVM is given by:

$$\min_{\mathbf{w}, b, \boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \tag{4}$$

$$\text{subject to} \quad y_i \left(\langle \mathbf{w}, \phi(x_i) \rangle + b \right) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{5}$$

Here, $\mathbf{w}$ is the weight vector in the transformed feature space, and ξ_i are slack variables.

3.3 Dual Formulation

Using Lagrange multipliers α_i and μ_i, we form the Lagrangian:

$$\mathcal{L} = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum \xi_i - \sum \alpha_i \left[y_i(\langle \mathbf{w}, \phi(x_i)\rangle + b) - 1 + \xi_i\right] - \sum \mu_i \xi_i \tag{6}$$

Taking derivatives and setting them to zero, we eliminate $\mathbf{w}$ and ξ_i, obtaining the dual problem:

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{7}$$

$$\text{subject to} \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{8}$$

$$0 \leq \alpha_i \leq C \tag{9}$$

This dual problem is a convex quadratic program.

3.4 Karush-Kuhn-Tucker (KKT) Conditions

KKT conditions must be satisfied at optimality:

- Primal constraints hold.
- Dual variables satisfy $\alpha_i \geq 0$ and $\mu_i \geq 0$.
- Complementary slackness: $\alpha_i \left(y_i f(x_i) - 1 + \xi_i\right) = 0$

3.5 Final Classifier

After solving the dual, the decision function is:

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b \tag{10}$$

3.6 Algorithm Implementation

The algorithm for nonlinear SVM with kernel trick is as follows:

Algorithm 1 Nonlinear SVM with Kernel Trick

Input: Dataset (x_i, y_i), kernel function K, regularization parameter C, tolerance ϵ, maximum iterations N
Initialize $\alpha_i \leftarrow 0$
$k \leftarrow 0$
while $k < N$ and not converged **do**
    Solve the dual quadratic programming problem using K to update α_i
    $k \leftarrow k + 1$
end while
Return: α_i