Support Vector Machine (SVM) Theory

1 Introduction

Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. The main goal of SVM is to find the optimal hyperplane that maximizes the margin between different classes in the feature space.

2 Objective Function

Given a training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector and $y_i \in \{-1, +1\}$ is the class label, the SVM aims to find a hyperplane that separates the classes with the maximum margin. The equation of the hyperplane can be written as:

$$\mathbf{w}^\top \mathbf{x} + b = 0 \tag{1}$$

where $\mathbf{w} \in \mathbb{R}^d$ is the weight vector and b is the bias term.

The objective is to maximize the margin between the two classes, which can be expressed as minimizing the following objective function:

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

subject to the constraints:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 \quad \text{for } i = 1, \ldots, n \tag{3}$$

3 Lagrangian Formulation

To handle the constraints, we use Lagrange multipliers. The Lagrangian for the SVM problem is:

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(\mathbf{w}^\top \mathbf{x}_i + b) - 1 \right] \tag{4}$$

where $\alpha_i \geq 0$ are the Lagrange multipliers. The dual problem is then formulated as:

$$\max_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^\top \mathbf{x}_j) \right\} \tag{5}$$

subject to:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{6}$$

$$\alpha_i \geq 0 \tag{7}$$

4 Support Vectors

The support vectors are the data points that lie closest to the hyperplane and are critical in determining the optimal hyperplane. These points satisfy:

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) = 1 \tag{8}$$

These support vectors are the only data points that influence the position and orientation of the hyperplane.

5 Kernel Trick

In cases where the data is not linearly separable, SVMs can be extended to handle non-linear classification problems using the kernel trick. The kernel function implicitly maps the data into a higher-dimensional space where a linear hyperplane can be used to separate the data. Common kernels include:

- **Polynomial Kernel**: $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^\top \mathbf{x}_j)^d$
- **Radial Basis Function (RBF) Kernel**: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$

where d is the degree of the polynomial and γ is a parameter for the RBF kernel.