Lecture Notes: PRML Chapter 7.1 – Maximum Margin Classifiers

Prerequisites

- Convex optimization and quadratic programming
- Lagrange multipliers and Karush-Kuhn-Tucker (KKT) conditions
- Concepts of margin, hyperplanes, and kernel functions
- Inner product space and feature-space transformations

Key Terminology

- **Support vector**: A training data point with a non-zero Lagrange multiplier; lies on or inside the margin boundary and defines the decision surface.
- **Margin**: Perpendicular distance from the decision boundary to the closest training point.
- **Slack variable** (ξ_n): Permits misclassification in non-separable data by relaxing the margin constraint.
- **Hinge loss**: Loss function used in SVM, defined as $[1 - y_n t_n]_+$.
- **Box constraint**: Constraint limiting Lagrange multipliers: $0 \leq a_n \leq C$.
- **Hard margin**: Assumes perfect separability, no slack.
- **Soft margin**: Introduces slack variables to allow some violations of separability.
- **Kernel function**: Defines an inner product in high-dimensional space without explicitly mapping the data.

Why It Matters

The support vector machine (SVM) provides a principled and efficient approach to classification with strong theoretical underpinnings from convex optimization. Its margin-maximizing behavior improves generalization and results in sparse solutions dependent only on critical examples—the support vectors. By reformulating in terms of kernels, it extends naturally to non-linear boundaries in high-dimensional spaces.

Key Ideas

Definition: Linear Discriminant in Feature Space

$$y(x) = w^\top \phi(x) + b \tag{7.1}$$

- $\phi(x)$: Fixed transformation to feature space. - b: Bias term.

Geometric Margin

Distance of x_n to decision boundary:

$$\frac{t_n y(x_n)}{\|w\|} = \frac{t_n(w^\top \phi(x_n) + b)}{\|w\|} \tag{7.2}$$

3-Column Derivation: Dual Form of Maximum Margin Problem

Step	Equation	Reason
1	$\min_{w,b} \frac{1}{2}\|w\|^2$ subject to $t_n(w^\top \phi(x_n) + b) \geq 1$	Formulate primal optimization for hard-margin SVM
2	$\mathcal{L}(w, b, a) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n \left[t_n(w^\top \phi(x_n) + b) - 1\right]$	Lagrangian with multipliers $a_n \geq 0$
3	$\frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow w = \sum_{n=1}^{N} a_n t_n \phi(x_n)$	Stationarity wrt w
4	$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} a_n t_n = 0$	Stationarity wrt b
5	Substitute into $\mathcal{L}$ to get: $\tilde{\mathcal{L}}(a) = \sum_n a_n - \frac{1}{2} \sum_{n,m} a_n a_m t_n t_m k(x_n, x_m)$	Dual form (7.10) using kernel trick $k(x, x') = \phi(x)^\top \phi(x')$

Step	Equation	Reason
6	Constraints: $a_n \geq 0$, $\sum_n a_n t_n = 0$	KKT conditions, feasible set for dual

---

3-Column Derivation: Introducing Slack for Soft Margin

Step	Equation	Reason
1	Add slack: $\xi_n \geq 0$, new constraint $t_n(w^\top \phi(x_n) + b) \geq 1 - \xi_n$	Allow margin violations
2	Objective: $\min \frac{1}{2}\|w\|^2 + C \sum_n \xi_n$	Penalize margin violations (7.21)
3	Lagrangian: $\mathcal{L} = \frac{1}{2}\|w\|^2 + C \sum_n \xi_n - \sum_n a_n[t_n(w^\top \phi(x_n) + b) - 1 + \xi_n] - \sum_n \mu_n \xi_n$	Dual problem with dual variables $\mu_n \geq 0$
4	Stationarity: $\frac{\partial \mathcal{L}}{\partial w} = 0 \Rightarrow w = \sum_n a_n t_n \phi(x_n)$	Optimality
5	$\frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \sum_n a_n t_n = 0$	Optimality
6	$\frac{\partial \mathcal{L}}{\partial \xi_n} = 0 \Rightarrow a_n = C - \mu_n$	Complementary slackness
7	Dual becomes same as hard margin but with $0 \leq a_n \leq C$	Box constraints (7.33)

Prediction Function

$$y(x) = \sum_{n \in S} a_n t_n k(x, x_n) + b \tag{7.13}$$

- Only support vectors ($a_n > 0$) contribute.

Computing the Bias Term

Average over support vectors n with $0 < a_n < C$:

$$b = \frac{1}{N_S} \sum_{n \in S} \left(t_n - \sum_{m \in S} a_m t_m k(x_n, x_m) \right) \tag{7.18}$$

Relevant Figures from PRML

- **Figure 7.1**: Shows the geometric margin and the role of support vectors in determining the boundary.
- **Figure 7.2**: Example of SVM on nonlinearly separable 2D data with Gaussian kernel—shows decision boundary and margin.
- **Figure 7.3**: Visualizes slack variable regimes: correctly classified, margin violators, and misclassified.
- **Figure 7.5**: Compares hinge loss, logistic loss, misclassification, and squared loss—motivates sparsity from hinge loss.