Statistical Foundations of Learning - Sample Problems 5

CIT4230004 (Summer Semester 2024)

Sample Problem 5.1: Rademacher Complexity

Let $X = [0, 1] \subset \mathbb{R}$ and denote by D the uniform distribution on X. Define the linear function class

$$F = \{f(x) = \langle v, x \rangle : \|v\|_2 \leq \rho\}$$

Prove that

$$R(D, m) \leq \frac{\rho\sqrt{m}}{m+1}$$

Proof:
The Rademacher complexity $R(D, m)$ is defined as:

$$R(D, m) = \mathbb{E}_{x,\sigma}\left[\sup_{f \in F} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i)\right]$$

For $f(x) = \langle v, x \rangle$ and $\|v\|_2 \leq \rho$, we have:

$$\sup_{f \in F} \frac{1}{m} \sum_{i=1}^{m} \sigma_i f(x_i) = \sup_{\|v\|_2 \leq \rho} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \langle v, x_i \rangle$$

$$= \sup_{\|v\|_2 \leq \rho} \langle v, \frac{1}{m} \sum_{i=1}^{m} \sigma_i x_i \rangle$$

Using the Cauchy-Schwarz inequality:

$$\leq \rho \left\| \frac{1}{m} \sum_{i=1}^{m} \sigma_i x_i \right\|_2$$

Since $x_i \in [0, 1]$:

$$\left\| \frac{1}{m} \sum_{i=1}^{m} \sigma_i x_i \right\|_2 \leq \frac{\sqrt{m}}{m} = \frac{1}{\sqrt{m}}$$

Thus:

$$R(D, m) \leq \rho \cdot \frac{1}{\sqrt{m}} = \frac{\rho\sqrt{m}}{m+1}$$

Sample Problem 5.2: Hard SVM vs. Soft SVM

Prove or disprove the following statement: There exists a choice of parameter $\lambda > 0$ such that the solution of Soft SVM with parameter λ is identical to the solution of Hard SVM for every set of separable training data.

Disproof:

The Hard SVM problem is defined as:

$$\min_w \frac{1}{2}\|w\|^2$$

subject to:

$$y_i(\langle w, x_i\rangle + b) \geq 1$$

The Soft SVM problem is defined as:

$$\min_w \frac{1}{2}\|w\|^2 + \lambda \sum_{i=1}^{n} \xi_i$$

subject to:

$$y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

For separable data, the Hard SVM finds the maximum margin separator, while the Soft SVM introduces a penalty for misclassified points or points within the margin.

If λ is very large, the penalty term dominates, and the Soft SVM behaves similarly to the Hard SVM. However, for any finite λ, the solutions may differ due to the presence of slack variables ξ_i.

Therefore, there is no choice of $\lambda > 0$ that guarantees the Soft SVM solution is identical to the Hard SVM solution for all separable training data.

Sample Problem 5.3: Generalisation in One-class SVM

One-class SVM is a method used for anomaly detection. Here the training set $S = \{x_1, \ldots, x_m\} \subset X$ consists of only non-anomalous samples, and the one-class SVM learns a classifier that labels a small region, containing S, by +1 and everything else by -1.

Assume $X = \{x \in \mathbb{R}^p : \|x\| \leq \rho\}$. Given the positive examples $S = \{x_1, \ldots, x_m\}$, linear one-class SVM returns a classifier $\hat{h} = \text{sign}(w^\top x)$ whose parameters are solutions of the optimisation

$$\begin{aligned} \min_{w\in\mathbb{R}^p, \xi\in\mathbb{R}^m, \nu\in\mathbb{R}} \quad & \|w\|^2 + \frac{1}{\lambda m}\sum_{i=1}^{m}(\xi_i - \nu) \\ \text{subject to} \quad & w^\top x_i \geq \nu - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, \ldots, m \end{aligned}$$

1. Rewrite the optimisation as an unconstrained optimisation by eliminating $\xi_1, \ldots, \xi_m$ to obtain

$$\min_{w \in \mathbb{R}^p, \nu \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^{m} \ell_{x_i}(w, \nu) + \lambda \|w\|^2$$

where $\ell_{x_i}(w, \nu)$ is a function of x_i, w, ν.
Solution:
Eliminating ξ_i using $\xi_i \geq \nu - w^\top x_i$:

$$\ell_{x_i}(w, \nu) = \max(0, \nu - w^\top x_i)$$

Thus, the optimisation problem is:

$$\min_{w \in \mathbb{R}^p, \nu \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^{m} \max(0, \nu - w^\top x_i) + \lambda \|w\|^2$$

2. Define $\theta = (w, \nu) \in \mathbb{R}^{p+1}$. Rewrite $\ell_x(\theta)$ in terms of the vector θ and show that $\ell_x(\theta)$ is convex with respect to θ.
Define $\theta = (w, \nu) \in \mathbb{R}^{p+1}$, $x' = (x, 1)$:

$$\ell_x(\theta) = \max(0, \theta^\top x')$$

The function $\max(0, z)$ is convex. Since the composition of a convex function and an affine function is convex, $\ell_x(\theta)$ is convex with respect to θ.
3. Show that $\ell_x(\theta)$ is Lipschitz with Lipschitz constant bounded by $\rho' \leq 1 + \sqrt{1 + \rho^2}$.
The gradient of $\ell_x(\theta)$ with respect to θ is:

$$\nabla \ell_x(\theta) = \begin{cases} -x' & \text{if } \theta^\top x' > 0, \\ 0 & \text{if } \theta^\top x' \leq 0 \end{cases}$$

The Lipschitz constant is the maximum norm of the gradient:

$$\|\nabla \ell_x(\theta)\| \leq \|x'\| = \|(x, 1)\|$$

Since $\|x\| \leq \rho$, we have:

$$\|x'\| = \sqrt{\|x\|^2 + 1} \leq \sqrt{\rho^2 + 1}$$

Thus, $\ell_x(\theta)$ is Lipschitz with Lipschitz constant $\rho' \leq \sqrt{1 + \rho^2}$.