CS281B/Stat241B (Spring 2006) Statistical Learning Theory — Lecture: 10

Kernel Methods for Regression

Lecturer: Peter Bartlett — *Scribe: Jiajun Wang*

Problem description: Given a probability distribution on $\mathcal{X}\times\mathbb{R}$, we would like to choose a function $f : \mathcal{X} \to \mathbb{R}$ to minimize the risk $L(f) = \mathbb{E}l(Y, f(X))$.

Examples of loss function $l(\cdot)$ include:

- $l(y, \hat{y}) = (y - \hat{y})^2$
- $l(y, \hat{y}) = |y - \hat{y}|$
- $l(y, \hat{y}) = (|y - \hat{y}| - \epsilon)_+$ (ϵ-insensitive, see Figure 1)
- $l(y, \hat{y}) = [(|y - \hat{y}| - \epsilon)_+]^2$

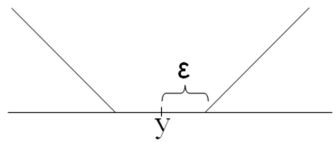

Figure 1: ϵ-insensitive loss function

The choice of function f that minimizes the risk is related to the conditional distribution $P(Y|X)$. If we have $Y = f(X) + Z$ and $X_1, \cdots, X_n$, then choosing a function f to maximize the likelihood is equivalent to

$$\max_f \prod_{i=1}^n p(Y_i|f(X_i)) \Leftrightarrow \min_f -\sum_{i=1}^n \log p(Y_i|f(X_i)).$$

This is equivalent to minimizing sample average of $l(Y_i, f(X_i))$ where $l(Y, f(X))$ matches $-\log p(Y|f(X))$.

1 Quadratic loss: kernel ridge regression

Choose $f \in \mathcal{H}$, where $\mathcal{H}$ is a RKHS, to minimize

$$\lambda\|f\|^2_{\mathcal{H}} + \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2.$$

This is equivalent to

$$\min \lambda\|f\|^2 + \frac{1}{n}\sum_{i=1}^n \xi_i^2 \quad \text{such that} \quad \xi_i = y_i - \langle f, k(x_i, \cdot)\rangle$$

Forming the Lagrangian and eliminating the primal variables, we obtain $f = \frac{1}{2\lambda}\sum_{i=1}^n \alpha_i k(x_i, \cdot)$ and $\xi_i = \frac{n\alpha_i}{2}$.

The dual is then $g(\alpha) = y^T\alpha - \frac{1}{4\lambda}\alpha^T K\alpha - \frac{n}{4}\alpha^T\alpha$ where $K_{ij} = k(x_i, x_j)$.

Maximizing the dual we get $\alpha = 2\lambda(K + n\lambda I)^{-1}y$. Therefore $f(x) = y^T(K + n\lambda I)^{-1}k$ where $k_i = k(x, x_i)$.

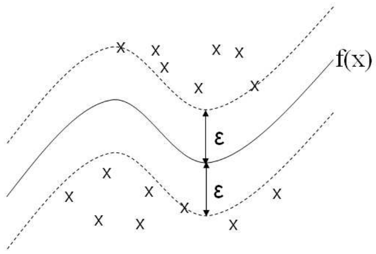

Figure 2: Only points on or outside the boundaries affect the solution.

2 Bayesian Analysis

Suppose $f : \mathcal{X} \to \mathbb{R}$ is chosen from a prior such that $\{(x, f(x)) : x \in \mathcal{X}\}$ is a Gaussian process, i.e. each finite set $\{x_1, \cdots, x_n\} \subseteq \mathcal{X}$ gives zero-mean Gaussian marginals $\{f(x_1), \cdots, f(x_n)\} \sim \mathcal{N}(0, \Sigma)$ where Σ is the covariance matrix and $\Sigma_{ij} = \mathbb{E}[f(x_i)f(x_j)]$. Notice that Σ completely specifies the Gaussian process and it is positive semidefinite and can thus be thought of as a kernel function. Therefore choosing the Σ to specify a Gaussian process is analogous to choosing the kernel function to specify the RKHS.

Also, suppose $y = f(x) + \xi$ where $\xi \sim \mathcal{N}(0, \sigma^2)$. Then the posterior distribution of $f(x_0)$ given $x_0, x_1, \cdots, x_n$ and $y_1, \cdots, y_n$ is

$$p(t_0, t|x_0, x, y) \quad \text{where} \quad t_0 = f(x_0) \text{ and } t = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$$

$$p(t_0, t|x_0, x, y) \propto p(y|t)p(t_0, t|x_0, x)$$

The first term $p(y|t) \propto \exp\{-\frac{(y-t)^T(y-t)}{2\sigma^2}\}$. The second term $p(t_0, t|x_0, x) \propto \exp\{-\frac{1}{2} \begin{bmatrix} t_0 & t^T \end{bmatrix} \Sigma^{-1} \begin{bmatrix} t_0 \\ t \end{bmatrix}\}$, where Σ is the covariance matrix for $x_1, \cdots, x_n$.

We can show that the mean of $t_0 (= f(x_0))$ is equal to $k^T(K + \sigma^2 I)^{-1}y$ where $k_i = \langle x_0, x_i \rangle$ and $K_{ij} = \langle x_i, x_j \rangle$.

3 ϵ-insensitive loss

Consider the ϵ-insensitive loss function $l(y, \hat{y}) = (|y - \hat{y}| - \epsilon)_+$ and the corresponding risk minimization problem:

$$\min_f \|f\|_{\mathcal{H}}^2 + \frac{C}{n}\Sigma(|y_i - f(x_i)| - \epsilon)_+.$$

From representer theorem, $f^*(\cdot) = \sum_{i=1}^n \alpha_i k(x_i, \cdot)$.

Using ν-parameterization, the problem becomes

$$\min_{f, \epsilon \geq 0} \lambda \|f\|_{\mathcal{H}}^2 + (\nu\epsilon + \frac{1}{n}\Sigma(|y_i - f(x_i)| - \epsilon)_+)$$

The solution can be expressed as the sum of support vectors lying on or outside the ϵ-boundary from $f(x)$ (see Figure 2).