Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 9 — September 11

Instructor: Stas Minsker *Scribe: Mose Wintner*

9.1 SVM and Reproducing Kernel Hilbert Spaces (RKHS)

Assume that we have as before $(X, Y) \in \mathbb{S} \times \{\pm 1\}$. Assume now that $\mathbb{S}$ is not a Hilbert space, or it is a low-dimensional Hilbert space.

Kernel trick: Let $\mathbb{H}$ be a separable Hilbert space. Let $\phi : \mathbb{S} \to \mathbb{H}$ be injective. ϕ induces

$$(X_1, \ldots, X_n) \mapsto (\phi(X_1), \ldots, \phi(X_n)).$$

We can define

$$f_{w,b}(\phi(x)) = \langle w, \phi(x) \rangle + b,$$

which is a linear functional on $\mathbb{H}$. According to the representer theorem, the solution of the SVM optimization problem can be represented by $w^* = \sum_{j=1}^n \alpha_j \phi(X_j)$. Then for any $x \in \mathbb{S}$,

$$g^*(x) = \text{sign}(f_{w^*,b^*}(\phi(x)) = \text{sign}(\langle w^*, \phi(x^*) \rangle + b^*).$$

Now

$$\langle w^*, \phi(x) \rangle = \sum_{j=1}^n \alpha_j^* \langle \phi(X_j), \phi(X) \rangle.$$

Observation: g^* depends on ϕ only through inner products in $\mathbb{H}$ of the form $\langle \phi(x), \phi(y) \rangle$. Let

$$K(x, y) = \langle \phi(x), \phi(y) \rangle,$$

the "reproducing kernel"; then we only need $K(x, y)$, not ϕ. The kernel K must satisfy the following:

1. $K(x_1, x_2) = K(x_2, x_1)$
2. $K(x, y)$ is positive semidefinite: for all $k \geq 1$, $x_1, \ldots, x_k \in \mathbb{S}$, $\alpha_1, \ldots, \alpha_k \in \mathbb{R}$,

$$\sum_{i,j=1}^k \alpha_i \alpha_j K(x_i, x_j) \geq 0.$$

Soft-margin SVM minimizes

$$\frac{1}{n} \sum_{j=1}^n (1 - Y_j f_{w,b}(X_j))_+ + \lambda ||w||^2$$

over $w, b \in \mathbb{H} \times \mathbb{R}$. Writing $\tilde{x_j} = \phi(x_j)$, then $w^* = \sum_{j=1}^n \alpha_j^* \phi(x_j)$, $||w^*||^2 = \sum_{i,j=1}^n \alpha_i^* \alpha_j^* K(x_i, x_j)$.

9.1.1 Reproducing Kernel Hilbert Spaces

Let $K : \mathbb{S} \times \mathbb{S} \to \mathbb{R}$ be a kernel that satisfies the above properties 1 and 2. Define

$$\mathbb{H}_0 = \left\{ \sum_{j=1}^{k} \alpha_j K(x_j, \cdot) : \, k \geq 1, \, \alpha_1, \dots \alpha_k \in \mathbb{R}, \, x_1, \dots, x_k \in \mathbb{S} \right\}.$$

Define the inner product on $\mathbb{H}_0$

$$\left\langle \sum_i \alpha_i K(x_i, \cdot), \sum_j \beta_j K(x_k, \cdot) \right\rangle := \sum_{i,j} \alpha_i \beta_j K(x_i, x_j)$$

and let $\mathbb{H}$ be the completion of $\mathbb{H}_0$. Then $\mathbb{H}$ is called the *RKHS* with kernel K. The term "reproducing" comes from the following: for any $f \in \mathbb{H}$, $\langle f, K(x, \cdot) \rangle = f(x)$. Indeed, if $f \in \mathbb{H}_0$, so $f = \sum_j \alpha_j K(x_j, \cdot)$, then

$$\langle f, K(x, \cdot) \rangle = \sum_j \alpha_j \langle K(x_j, \cdot) K(x_j, \cdot) \rangle = \sum_j \alpha_j K(x_j, x) = f(x).$$

From this point of view, the soft-margin SVM is a penalized hinge risk minimization in an RKHS $\mathbb{H}$.