Chapter 12

Support Vector Machines and Flexible Discriminants

Reproducing Kernel Hilbert Space

Definition 12.1 (Reproducing Kernel Hilbert Space, a.k.a. RKHS) *A Hilbert space of functions $\mathcal{H} = \{f|f : \mathcal{X} \to \mathbb{R}\}$ is a RKHS if every of its evaluation functional is continuous.*

Remark 12.1.1 *For any Hilbert space of functions $\mathcal{H} = \{f|f : \mathcal{X} \to \mathbb{R}\}$ and any $x \in \mathcal{X}$, a evaluation funtional is a mapping from $\mathcal{H}$ to $\mathbb{R}$, denoted as Ev_x, with $Ev_x(f) = f(x)$.*

Remark 12.1.2 *A evaluation functional Ev_x with respect to a Hilbert space of functions $\mathcal{H}$ is continuous if there exists $M_x > 0$ such that $|Ev_x(x)| \triangleq |f(x)| \leq M_x ||f||_{\mathcal{H}}$. In other words, if two functions f and g in $\mathcal{H}$ are close in norm, then they are point-wise close: $|f(x) - g(x)| \leq M_x ||f - g||_{\mathcal{H}}$. If we construction a sequence of functions f_n with $||f_n - g|| \to 0$ as $n \to \infty$, because of the fact that M_x is in general a function of x, the convergence of f to g is point-wise, not uniform.*

Remark 12.1.3 *The definition of RKHS has nothing to do with* reproducing kernel*, as it is just a Hilbert space of functions with continuous evaluation functional. The name prefix of "Reproducing Kernel" comes from the fact that the evaluation functional of RKHS can be expressed as inner product, a manifestation of Riesz representation theorem shown next.*

Theorem 12.2 (Riesz Representation Theorem) *Every continuous linear functional Φ defined on a Hilbert space of functions $\mathcal{H}$ can be written uniquely in the form $\Phi(f) = \langle f, g \rangle_{\mathcal{H}}$ for some appropriate element $g \in \mathcal{H}$.*

Remark 12.2.1 (Reproducing Kernel) *According to Riesz Representation Theorem, given a RKHS $\mathcal{H} = \{f|f : \mathcal{X} \to \mathbb{R}\}$, for any $x \in \mathcal{X}$, since the evaluation functional $Ev_x : \mathcal{H} \to \mathbb{R}$ is continuous, it can be expressed as $Ev_x(f) = \langle K_x, f \rangle_{\mathcal{H}} = f(x)$ for a unique $K_x \in \mathcal{H}$. With the set of functions $\{K_x, x \in \mathcal{X}\} \subset \mathcal{H}$, we can define a mapping from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as $K(x, y) = \langle K_x, K_y \rangle_{\mathcal{H}}$, which is called Reproducing Kernel.*

Remark 12.2.2 *If we have a Reproducing Kernel defined, then it is easy to see that the corresponding evaluation functional is continuous by resorting to Cauchy Schwarz:* $|f(x) - g(x)| = |\langle K_x, f\rangle_{\mathcal{H}} - \langle K_x, g\rangle_{\mathcal{H}}| = |\langle K_x, f - g\rangle_{\mathcal{H}}| \leq ||K_x||_{\mathcal{H}}||f - g||_{\mathcal{H}}$.

Definition 12.3 (Positive Definite Kernel, a.k.a. PD Kernel) *A symmetric function* $K : \mathcal{X} \times \mathcal{X}$ *is called a positive definite kernel on* $\mathcal{X}$ *if*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0,$$

for any $n \in \mathbb{N}$ *and any* $x_1, x_2, \ldots, x_n \in \mathcal{X}$ *and any* $c_1, c_2, \ldots, c_n \in \mathbb{R}$. *It can be thought of as a generalization of positive semi-definite matrix.*

Remark 12.3.1 *If* $K(\cdot, \cdot)$ *is a reproducing kernel associated to a RKHS* $\mathcal{H}$*, then we have*

$$\begin{aligned} \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(x_i, x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \left\langle K_{x_i}, K_{x_j}\right\rangle_{\mathcal{H}} \\ \textit{(bilinearity of inner product)} &= \sum_{i=1}^{n} c_i \left\langle K_{x_i}, \sum_{j=1}^{n} c_j K_{x_j}\right\rangle_{\mathcal{H}} \\ &= \left\langle \sum_{i=1}^{n} c_i K_{x_i}, \sum_{j=1}^{n} c_j K_{x_j}\right\rangle_{\mathcal{H}} \geq 0, \end{aligned}$$

meaning that it is a PD Kernel.

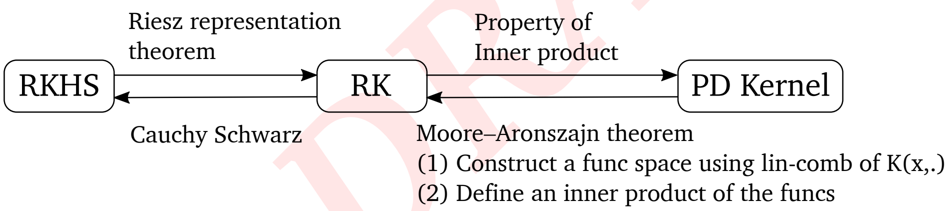

Figure 12.1: Relationship between Reproducing Kernel Hilbert Space, Reproducing Kernel, and Positive Definite Kernel. The mapping among the three concepts is unique.

Definition 12.4 (Translation Invariant Kernel) *For a RKHS* $\mathcal{H} \subset L^2$ *with the inner product defined as the following*

$$\begin{aligned} \langle f, g\rangle_{\mathcal{H}} &= \int \frac{F(\omega)G^*(\omega)}{Q(\omega)} d\omega \\ ||f||_{\mathcal{H}}^2 &= \int \frac{|F(\omega)|^2}{Q(\omega)} d\omega, \end{aligned}$$

with $F(\cdot)$ *being the Fourier transform of* f *and the real function* $Q(\omega) \to 0$ *as* $\omega \to \infty$*, then the reproducing kernel can be expressed as* $K(x, y) = q(||x - y||)$ *with* q *being the inverse Fourier* *transform of Q. Here's a short proof for 1-dimensional case:*

$$f(x) = \int F(\omega)e^{i\omega x}d\omega = \int \frac{F(\omega)Q(\omega)e^{i\omega x}}{Q(\omega)}d\omega$$
$$= \left\langle f, \mathcal{F}^{-1}(Q(\omega)e^{i\omega x})\right\rangle = \langle f, q(\cdot - x)\rangle .$$

Some typical examples on $Q(\omega)$ and the corresponding Kernels are provided next.

Remark 12.4.1 (Poisson/Abel Kernel)

$$Q(\omega) = \frac{2\gamma}{\gamma^2+\omega^2} \quad \Rightarrow q(x) = e^{-\gamma|x|} \quad \Rightarrow K(x,y) = e^{-\gamma|x-y|}$$
$$||f||_{\mathcal{H}_K} = \int \frac{|F(\omega)|^2}{Q(\omega)}d\omega = \int \frac{1}{2}\gamma|F(\omega)|^2 + \frac{1}{2\gamma}|\omega F(\omega)|^2 d\omega = \int \frac{1}{2}\gamma|f(x)|^2 + \frac{1}{2\gamma}|f'(x)|^2 dx$$

Remark 12.4.2 (Gaussian/Radio-Basis-Function Kernel)

$$Q(\omega) = \frac{2\gamma}{\gamma^2+\omega^2} \quad \Rightarrow q(x) = e^{-\gamma|x|} \quad \Rightarrow K(x,y) = e^{-\gamma|x-y|}$$
$$||f||_{\mathcal{H}_K} = \int \frac{|F(\omega)|^2}{Q(\omega)}d\omega = \int \frac{1}{2}\gamma|F(\omega)|^2 + \frac{1}{2\gamma}|\omega F(\omega)|^2 d\omega = \int \frac{1}{2}\gamma|f(x)|^2 + \frac{1}{2\gamma}|f'(x)|^2 dx$$

Problem 12.1

The original formulation of support vector classifier with margin is:

$$\arg\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t. } \xi_i \geq 0, y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \text{ for any } 1 \leq i \leq N.$$

From the constraint we know that

$$\xi_i \geq 1 - y_i(x_i^T\beta + \beta_0) \text{ and } \xi_i \geq 0$$
$$\Rightarrow \xi_i \geq \left(1 - y_i(x_i^T\beta + \beta_0)\right)^+ ,$$

which transforms the original problem to the following

$$\arg\min_{\beta,\beta_0} \frac{1}{2}||\beta||^2 + C\sum_{i=1}^{N}\left(1 - y_i(x_i^T\beta + \beta_0)\right)^+$$
$$= \arg\min_{\beta,\beta_0} \underbrace{\frac{1}{2C}||\beta||^2}_{\substack{\text{sum-of-squares}\\\text{penalty of parameters}}} + \underbrace{\sum_{i=1}^{N}\left(1 - y_i(x_i^T\beta + \beta_0)\right)^+}_{\text{SVM hinge loss}} .$$