data: tuples $\{x_i, y_i\}_{i=1}^n$, $x_i \in \mathbb{R}^d$

supervised learning (learning of functions) maps

$$\mathcal{X} \to \mathcal{Y} \qquad x \mapsto f(x) \qquad \mathcal{F}: \text{function space}$$

In RKHS, $f \in H_K$, infinite dim space.

computed by kernel trick
(general kernel matrix)

Gram matrix: $G = (K_{ij})_{ij}^n$, $K_{ij} = K(x_i, x_j)$

general kernel matrix: $G = (K(x_i, x_j))_{ij}^n$

$$\Delta: \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad \text{PSD kernel}$$

eg: $K(x, y) = x^T y$ the linear kernel

$K(x, y)$ generally may be non-linear

$K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$

feature map, feature space

Formal definitions:

$\mathcal{X}$ : domain of data $\quad \mathcal{X} \subset \mathbb{R}^d$.

- PSD kernel

def: Function $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called symmetric if $K(x, y) = K(y, x) \quad \forall x, y \in \mathcal{X}$

def: A symmetric function $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive-semi-definite (PSD) kernel if for any $x_1, \ldots, x_n \in \mathcal{X}$, $a_1, \ldots, a_n \in \mathbb{R}$,

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

Rk: It is called strictly positive definite (PD) if for distinct $x_1, \ldots, x_n$ equality holds iff $a_1 = \cdots = a_n = 0$.

- If $K$ is a PSD kernel, then general Gram matrix

$$G = (K(x_i, x_j))_{ij}^n \quad \text{for } \{x_i\}_i \in \mathcal{X}$$

is a (symmetric) PSD matrix.

- Hilbert space

inner product space $(V, \langle \cdot, \cdot \rangle)$

def: $V$ is a $\mathbb{C}$-vector space, a mapping $\langle \cdot, \cdot \rangle: V \times V \to \mathbb{C}$ is called an inner product if

(1) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$, $\langle \lambda x, y \rangle = \lambda \langle x, y \rangle$, $\forall x, y, z \in V$, $\lambda \in \mathbb{C}$

(2) $\langle x, y \rangle = \overline{\langle y, x \rangle}$

(3) $\langle x, x \rangle \geq 0$, $\langle x, x \rangle = 0 \Leftrightarrow x = 0$

Rmk: $\langle x, x \rangle$ is real, $\langle x, y \rangle$ is complex in general, $\langle x, y \rangle = \overline{\langle y, x \rangle}$

Cauchy-Schwarz: $|\langle x, y \rangle| \leq \sqrt{\langle x, x \rangle} \sqrt{\langle y, y \rangle}$

Hilbert space: An inner product space $(V, \langle \cdot, \cdot \rangle)$ is Hilbert if $V$ is complete under the induced norm $\|x\| = \sqrt{\langle x, x \rangle}$

eg: $L^2(\mathbb{R})$ : $\int_{-\infty}^{\infty} |f(x)|^2 dx < \infty$ (Lebesgue)

$L^2(\mathbb{R}, \mu)$ : $\int_{-\infty}^{\infty} |f(x)|^2 d\mu < \infty$ (Borel measure)

$\langle f, g \rangle = \int_{\mathbb{R}} f(x) \overline{g(x)} d\mu$

(finite) sequence, also space of equivalence classes of functions that differ on a set with measure zero (of zero measure zero)

Rk: There is $\mathbb{C}$-isomorphism bw $L^2(\mathbb{R})$ and $L^2(\mathbb{C}^N)$ by Fourier series.

$$f(x) = \sum_{n=-\infty}^{\infty} c_n e^{inx}$$

$$\|f\|^2 = \sum_{n=-\infty}^{\infty} |c_n|^2 \quad \in L^2(\mathbb{Z})$$

where $f \in L^2(\mathbb{C}, \mu)$, $\quad$ (Parseval's identity)

$$\langle f, g \rangle = \int_{\mathbb{R}} f(x) \overline{g(x)} d\mu = \sum_{n=-\infty}^{\infty} c_n \overline{d_n} \in L^2(\mathbb{Z})$$

- orthonormal set
$\{e_n = e^{inx}\}$

def: A countable collection of vectors $\{\phi_i\}_i$ in $V$ forms an orthonormal set if $\langle \phi_i, \phi_j \rangle = \delta_{ij}$ (or zero for $i \neq j$)

Rk: This implies that $\|\phi_i\|^2 = 1$.

- RKHS (Reproducing Kernel Hilbert Space)

def: An non-empty set $A$ RKHS over $\mathcal{X}$ is a Hilbert space $\mathcal{H} = (\mathcal{H}, \langle \cdot, \cdot \rangle)$ where there is a function $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying that

(1) $\forall x \in \mathcal{X}$, $K(\cdot, x) \in \mathcal{H}$

(2) $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, K(\cdot, x) \rangle_{\mathcal{H}} = f(x)$
(reproducing property)

The function $K$ is called the reproducing kernel of $\mathcal{H}$.

Rk: We call $\phi(x) = K(\cdot, x) \in \mathcal{H}$ a feature map and $\forall x, y \in \mathcal{X}$

$$K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}, \quad \text{Mercer}$$

thus $K$ is symmetric PSD.

Pf of (2): For any $f \in \mathcal{H}$, $f(x) = \langle K(\cdot, x), f \rangle$, $\forall x \in \mathcal{X}$, $\forall y \in \mathcal{X}$,

thus $K(x, y) = \langle K(\cdot, y), K(\cdot, x) \rangle$

$= \overline{\langle K(\cdot, x), K(\cdot, y) \rangle}$

$= \overline{\langle K(\cdot, y), K(\cdot, x) \rangle}$ ($\langle A, B \rangle$ Hermitian)

- This reproducing kernel $K$ of an RKHS is unique. (by Pf)

- $K$ is PSD kernel

def: $\forall x_1, \ldots, x_n \in \mathcal{X}$, $\forall a \in \mathbb{R}^n$,

$$\sum_{i,j=1}^n a_i a_j K(x_i, x_j) = \sum_{i,j=1}^n a_i a_j \langle \phi(x_i), \phi(x_j) \rangle$$

$$= \langle \sum_{i=1}^n a_i \phi(x_i), \sum_{j=1}^n a_j \phi(x_j) \rangle$$

$$= \|\sum_{i=1}^n a_i \phi(x_i)\|^2 \geq 0$$

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

Pf: $K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \langle \phi(y), \phi(x) \rangle_{\mathcal{H}}$ (by Def of $\mathcal{H}$ RKHS)

$= \overline{\langle \phi(x), \phi(y) \rangle} = \overline{K(x, y)} \geq 0$

The above gives that RKHS $\xrightarrow{\text{def}}$ $K$ is PSD

the converse PSD $K \xrightarrow{\text{Moore-Aronszajn}}$ RKHS, also holds

Thm (Moore-Aronszajn): Let $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, be a symmetric PSD kernel, then there exists an RKHS $H_K$ consisting of functions on $\mathcal{X}$ s.t.

1) $\langle K(\cdot, x), K(\cdot, y) \rangle_{H_K} = K(x, y)$ $\quad \forall x, y \in \mathcal{X}$

2) span $\{K(\cdot, x) : x \in \mathcal{X}\}$ dense in $H_K$

3) $K$ is a reproducing kernel on $H_K$

Rk: Strict correspondence bw PSD kernels and RKHS

$$K \quad \text{and} \quad H_K$$

Source: notes 2, p. 3a

Eg: $\mathcal{X} = \mathbb{S}^1 = \{x \in \mathbb{R}^2 : \|x\| = 1\}$ with periodic boundary condition.

Fourier basis $\{e^{inx}\}_{n=-\infty}^{\infty}$ $\quad$ orthonormal on $L^2(\mathbb{S}^1)$

$$K(x, y) = \sum_{n=-\infty}^{\infty} e^{in(x-y)} \quad \text{PSD}$$

$$\left( \int_{-\pi}^{\pi} e^{inx} \overline{e^{imx}} dx = \begin{cases} 0 & \text{if } n \neq m \\ 2\pi & \text{if } n = m \end{cases} \right)$$

$K(x, y) = \sum_{n=1}^{\infty} 2\pi^{-n} \cos(n(x-y))$. Rewritten

$\phi(x) = K(\cdot, x) = \sum_{n=1}^{\infty} 2\pi^{-n} \cos(nx) \in L^2(\mathbb{S}^1)$

Q: What is the RKHS associated with $K$?

$\langle K, \phi_n \rangle$ satisfy the reproducing property

Suppose $f = \sum_{n=1}^{\infty} f_n e^{inx}$, $f(x) = \sum_{n=1}^{\infty} f_n e^{inx}$

(convergence in $L^2$ sense)

$$\langle f, K(\cdot, x) \rangle_H = \langle \sum_n f_n e^{inx}, \sum_n 2\pi^{-n} e^{inx} \rangle_H$$

$$= \sum_{n,m} f_n \overline{2\pi^{-m} e^{imx}} \langle e^{inx}, e^{imx} \rangle_H$$

$$= f(x) = \sum_n f_n e^{inx}$$

$\Rightarrow \langle e^{inx}, e^{imx} \rangle_H = 2\pi^{n}$

$\Rightarrow$ for $f(x) = \sum_n f_n e^{inx}$, $\|g\|_H = \sum_n f_n e^{inx}$,

$$\langle f, g \rangle_H = \sum_n f_n \overline{g_n} 2\pi^{n}$$

and $\|f\|_H^2 = \sum_n |f_n|^2 2\pi^{n}$. $\sum_n |f_n|^2 < \infty$ for $f \in H_K$

Thus, $H_K$ consists of functions, in a subset of $L^2(\mathbb{S}^1)$

Rk: verify that $K(x, y) = \langle K(\cdot, x), K(\cdot, y) \rangle_{H_K}$

Eg: (Fourier basis on $\mathbb{R}$)

$$K(x, y) = e^{-\frac{(x-y)^2}{2\sigma^2}}, \quad x, y \in \mathbb{R},$$

$H_K = \{f \in L^2(\mathbb{R}, d\mu) : \int |\hat{f}(\omega)|^2 e^{\frac{\omega^2 \sigma^2}{2}} d\omega < \infty\}$

$\langle f, g \rangle_{H_K} = \int f(\omega) \overline{\hat{g}(\omega)} e^{\frac{\omega^2 \sigma^2}{2}} d\omega$

Rk: verify the reproducing property

- Integral operator and Mercer's Thm

$K(x, y)$ is a continuous kernel on $(\mathcal{X}, d\mu)$

$\mathcal{X}$ : finite measure on $\mathcal{X}$, $\text{supp}(\mu) = \mathcal{X}$,

$K$: operator, square-integrable

$$f_{K,\mu}: L^2(\mathcal{X}) \to L^2(\mathcal{X})$$

The integral operator of $K$:

$$T_K f(x) = \int_{\mathcal{X}} K(x, y) f(y) d\mu(y), \quad \forall f \in L^2(\mathcal{X}, \mu)$$

$T_K$ is compact and self-adjoint, positive. Thus applies to spectral theorem

$$T_K \phi_i = \lambda_i \phi_i, \quad \langle \phi_i, \phi_j \rangle = \delta_{ij}$$

$\{\phi_i\}_i$ are a CONS on $L^2(\mathcal{X}, \mu)$

Then (Mercer) $K$ is PSD kernel, square-integrable, on $\mathcal{X} \times \mathcal{X}$, and continuous, then $\{\lambda_i\}_i$ is real and $\geq 0$ and

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y), \quad \forall x, y \in \mathcal{X},$$

the convergence is absolute and uniform for $\{x, y\}$.

Rk: The inverse implies is false. $\{x, y\} \in \mathbb{R}^2, \{x, y\} \in \mathbb{R}$.

- The Kernel trick

$K(x, y) = \langle \phi(x), \phi(y) \rangle_{H_K}$.

$\forall x, y$ fixed, $\phi(x)$, $\phi(y)$ infinite dimensional

finite $\{x_i\}_i$ can be computed on finite samples.

The cost: MLE algorithm, only have product $\langle \phi(x), \phi(y) \rangle_{H_K}$ (general kernel matrix) is needed

$\Rightarrow$ infinite $\{x_i\}$ to be $\langle \phi(x_i), \phi(y) \rangle_{H_K}$

Examples of kernels:

- RBF $\quad K(x, y) = \exp(-\|x - y\|^2)$

$e^{-\|x-y\|^2}$ (Gaussian)

$e^{-\|x-y\|}$ (Laplace)

- Polynomial $\quad (\langle x, y \rangle + c)^d$

The benefit of non-linear features.

Solving XOR: XOR not linearly separable, but can be solved after mapping to 3 dimensions