RKHS Methods in Machine Learning

S. Sumitra
Department of Mathematics
Indian Institute of Space Science and Technology

First we discuss the background material required for the formal development of the RKHS formulation. The concepts of metric spaces, vector spaces, normed spaces and inner product spaces are essential for understanding the concepts of RKHS.

1 Metric space

We all aware of determining the distance between two real numbers using Euclidean distance formula. In the same way, is that possible to find the distance between two real valued functions defined on $[a, b]$? The answer is yes, if they are members of a metric space. Given below is the definition of a metric space:

Definition A space X is called a metric space if a metric(distance function) d is defined on $\mathcal{X} \times \mathcal{X}$ such that for all $x, y, z \in X$ we have:

- d is real-valued, finite and non-negative
- $d(x, y) = 0$ iff $x = y$
- $d(x, y) = d(y, x)$ (Symmetry)
- $d(x, y) \leq d(x, z) + d(z, y)$ (Triangle inequality)

1.1 Examples of metric spaces

1. R^n with the metric $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots (x_n - y_n)^2}$
2. C^n with the metric $d(x, y) = \sqrt{|x_1 - y_1|^2 + |x_2 - y_2|^2 + \ldots |x_n - y_n|^2}$
3. $C[a, b]$ with the metric $d(x, y) = \max_{t \in [a,b]} |x(t) - y(t)|$

4. l_∞ with the metric $d(x,y) = \sup_{j\in N} |\eta_j - \psi_j|$, where l_∞ is a bounded sequence of complex numbers.

Next we will look into how the concept of continuous functions can be introduced in metric spaces.

Definition 1.1 *A mapping* $T\colon (\mathcal{X}, d) \to (\mathcal{Y}, \tilde{d})$ *is said to be continuous at* x_0 *if for every* $\epsilon > 0, \exists\ \delta > 0$ *such that* $d(x,y) < \delta \to d(Tx, Ty) < \epsilon$

2 Limit Point

Let $x_0 \in \mathcal{X}$, where $\mathcal{X}$ a metric space. Then x_0 is said to be a limit point of a subset M of $\mathcal{X}$, if $\forall \epsilon > 0$, $\exists x_n \neq x_0 \in M$ such that $d(x_n, x_0) < \epsilon$

Definition 2.1 *The closure of a subset* M *of a metric space* X *is the set consisting of* M *and all the limit points of* M *and it is represented as* $\overline{M}$.

Definition 2.2 *A subset* M *of a metric space* X *is dense in* X *if* $\overline{M} = X$.

2.1 Convergence of a Sequence

The concept of convergence of sequence can be introduced only in metric spaces.

Definition 2.3 *A sequence* (x_n) *in* $(X.d)$ *is said to be a convergent sequence if there exists a* $x_0 \in \mathcal{X}$, *such that* $\forall \epsilon > 0$, $\exists N$ *such that* $d(x_n, x_0) < \epsilon\ \forall n > N$.

The sequence $(\frac{1}{n}, n \in \mathbb{N})$ converges to 0. This is a convergent sequence in $[0,1] \subset \mathbb{R}$, but a divergent sequence in $(0,1) \subset \mathbb{R}$, as the limit $0 \in [0,1]$ and $0 \notin (0,1]$.

A sequence (x_n) in a metric space is said to be Caushy sequence if for every $\epsilon > 0$ there is a N such that $d(x_m, x_n) < \epsilon\ \forall m, n > N$.

Theorem 2.4 *Every convergent sequence is Caushy.*

Proof Let (x_n) is a convergent sequence in $\mathcal{X}$. Therefore $\exists\ x_0 \in \mathcal{X}$ such that $d(x_n, x_0) < \frac{\epsilon}{2}, \forall n > N$. Now $d(x_m, x_n) < d(x_m, x_0) + d(x_0, x_n) = \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon\ \forall m, n > N$. Therefore (x_n) is a Caushy sequence.

The converse of theorem 2.4 is not true for all metric spaces. There are some spaces where every Caushy sequence converges. The space $\mathcal{X}$ is said to be a complete metric space if every Caushy sequence converges.

R^n, C^n are complete metric spaces with usual metric. The set of rational numbers Q with Euclidean metric is not a complete metric space, as every irrational number can be represented as the limit of rational numbers.

Theorem 2.5 *$x \in \overline{M}$, if and only if there exists a sequence $(x_n) \in M$, such that $x_n \to x$.*

Proof Let $x \in \overline{M}$. If $x \in M$, then $(x, x, \ldots x) \to x$. If $x \notin M$, then also we can find a sequence $(x_n) \in M$ that converges to x, by taking $x_n \in B(x, 1/n)$, as x is the limit point of M. To prove the converse, assume there exists a sequence $(x_n) \in M$ that converges to x. Then every neighbor hood of x contains atleast a x_n, that is atleast one element of M. Therefore x is a limit point of M.

If M is closed, then $M = \overline{M}$ and therefore for every $x \in M$, there exists a sequence $(x_n) \in M$, such that $x_n \to x$.

3 Vector Spaces

A vector space over a field K is a non empty set V on which are defined two operations, vector addition and scalar multiplication such that the following conditions are satisfied $\forall u, v, w \in V$:

- Closed under vector addition: $u + v \in V$
- Associative under vector addition: $(u + v) + w = u + (v + w)$
- Commutative under vector addition: $u + v = v + u$
- Existence of additive identity: $\exists\, 0 \in V$, such that $0 + u = u$
- Existence of additive inverse: $\exists s \in V$ such that $u + s = 0$
- Closed under scalar multiplication: $\forall \alpha \in K, \alpha v \in V$
- Associative under scalar multiplication: $\alpha(\beta v) = (\alpha\beta)v, \alpha, \beta \in K$
- Distributive: $\alpha(u + v) = \alpha u + \alpha v$, $(\alpha + \beta)u = \alpha u + \beta u, \alpha, \beta \in K$
- Unitarity: $1u = u, 1 \in K$

3.1 Normed Space

A normed space is a vector space with a norm defined on it. A Banach space is a complete normed space (complete in the metric defined by the norm). Here a norm on a vector space X is a real-valued function X whose values at an $x \in \mathcal{X}$ is denoted by $||x||$ which has properties

- $||x|| \geq 0$
- $||x|| = 0 \iff x = 0$
- $||\alpha x|| = |\alpha|||x||$
- $||x + y|| \leq ||x|| + ||y||$

where $x, y \in X$ and α is any scalar.

A norm on X defines a metric d on X which is given by

$$d(x, y) = ||x - y||$$

is called the metric induced by the norm.

A complete normed space is called a Banach space.

3.2 Linear Operator

Definition A linear operator T is an operator such that

- the domain $\mathcal{D}(T)$ and the range $\mathcal{R}(T)$ of T are vector spaces over the same field
- $T(x + y) = T(x) + T(y)$; $T(\alpha(x)) = \alpha T(x)$ where, $x, y \in \mathcal{D}(T)$ and $\alpha \in K$.

3.3 Bounded Linear Operator

Definition Let X and Y be normed spaces and $T : \mathcal{D}(T) \to Y$ a linear operator, where $\mathcal{D}(T) \subset X$. The operator T is said to be bounded if there is a real number c such that for all $x \in \mathcal{D}(T)$, $||Tx|| \leq c||x||$.

$||T|| = sup_{x \in \mathcal{D}(\mathcal{T}), x \neq 0} \dfrac{||Tx||}{||x||}$ or$||T|| = sup_{x \in \mathcal{D}(\mathcal{T}), ||x|| = 1} ||Tx||$.

3.4 Bounded Linear Functional

A bounded linear functional f is a bounded linear operator with the range lies on the scalar field of its domain.

$$f : \mathcal{D}(f) \to K$$

$$||f|| = sup_{x \in \mathcal{D}(\mathcal{T}), x \neq 0} \frac{||f(x)||}{||x||}$$

or else

$$||f|| = sup_{x \in \mathcal{D}(\mathcal{T}), ||x||=1} ||f(x)||$$

.

Theorem 3.1 *A linear operator T is continous iff it is bounded.*

3.5 Inner Product Space, Hilbert Space

Definition An inner product space is a vector space X with an inner product defined on X. An inner product on X is a mapping of $X \times X$ into the scalar field K of X; that is every pair of vectors x and y there is associated a scalar, which is written $\langle x, y \rangle$ and is called the inner product of x and y, such that for all vectors x,y and scalars α we have

- $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$
- $\langle x, y \rangle = \overline{\langle y, x \rangle}$ (if $K = \mathbb{R}, \overline{\langle y, x \rangle} = \langle y, x \rangle$)
- $\langle x, x \rangle \geq 0$, $\langle x, x \rangle = 0, \iff x = 0$

$||x|| = \sqrt{\langle x, x \rangle}$, $d(x, y) = ||x - y|| = \sqrt{\langle x - y, x - y \rangle}$

Definition A Hilbert space is a complete inner product space.

Theorem 3.2 *(Projection Theorem) Let Y be a closed subspace of a Hilbert space H. Then $H = Y \oplus Y^{\perp}$.*

Lemma 3.3 *For any subset $M \neq \phi$ of a Hilbert space H, the span of M is dense in H iff $M^{\perp} = \{0\}$.*

[**Refer Kreyszig book for examples for the above given spaces and for more information**]. (Iin the rest of this article, the scalar field K of any vector space is taken to be $\mathbb{R}$.)

4 Reisz Representation Theorem

Definition Reisz representation theorem: Every bounded linear functional f on a Hilbert space H can be represented in terms of the inner product

$$f(x) = \langle x, z \rangle$$

where z depends on f, is uniquely determined by f and has norm $||f|| = ||z||$.

Definition An evaluation functional over the Hilbert space of functions F is a linear functional $L_x : \mathcal{F} \rightarrow \mathbb{R}$ such that $L_x(f) = f(x), \forall f \in \mathcal{F}$.

5 Hyperplanes in Hilbert Sapce

Let H be a Hilbert space. Let f be a bounded linear functional defined on $\mathcal{F}$. Therefore by Reisz representation theorem, $\exists w \in \mathcal{H}$, such that $f(x) = \langle w, x \rangle, \forall x \in \mathcal{H}$. Then $\Pi_w = \{x \in \mathcal{F} : f(x) - b = 0, b \in \mathbb{R}\}$ is called the hyperplane associated with f, having the parameter w and b. The hyperplane divides H in two half spaces: $\Pi_1 = \{x \in \mathcal{F} : f(x) - b \geq 0, b \in \mathbb{R}\}$ and $\Pi_2 = \{x \in \mathcal{F} : f(x) - b < 0, b \in \mathbb{R}\}$. In this case the equation to the hyperplane is $\langle w, x \rangle - b = 0$.

6 Reproducing Kernel Hilbert spaces

Definition A RKHS, $\mathcal{F}$, is a Hilbert space of functions on some set $\mathcal{X}$ in which all the point evaluations are bounded linear functionals.

Let $\mathcal{X} \subseteq \mathbb{R}^n$. For each $x_i \in \mathcal{X}$, if we define $L_{x_i} : \mathcal{F} \rightarrow \mathbb{R}$ such that,

$$L_{x_i}(f) = f(x_i), \tag{1}$$

where $f \in \mathcal{F}$, then by the definition of RKHS, $\{L_{x_i}\}_{x_i \in \mathcal{X}}$ are bounded [since $L'_{x_i}s$ are point evaluation functionals]. Hence by the Reisz representation theorem there exists a set of functions $\{k_{x_i} : x_i \in \mathcal{X}\} \subseteq \mathcal{F}$ such that

$$L_{x_i} f = \langle f, k_{x_i} \rangle, \forall f \in \mathcal{F} \tag{2}$$

where k_{x_i} depends only on L_{x_i}. Therefore, corresponding to every $x \in \mathcal{X}$, $\exists k_x \in \mathcal{F}$. Hence the following are well defined functions: $\phi : \mathcal{X} \rightarrow \mathcal{F}$ such that $\phi(x) = k_x$ and $k \colon \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$

$$k(x, y) = \langle k_x, k_y \rangle = \langle \phi(x), \phi(y) \rangle \tag{3}$$

The function k is called the reproducing kernel (r.k.) and ϕ is called its feature map. k_x is called the representer of evaluation at x.

Substituting $k(x, y)$ in place of $\langle \phi(x), \phi(y) \rangle$ is known as kernel trick, in the field of machine learning community.

Theorem 6.1 *If* $M = \{k_{xi}, i = 1, 2, \dots\}$, *then* $\overline{span(M)} = \mathcal{F}$.

Proof Let $f \in M^{\perp}$. Therefore $\langle f, k_x \rangle = 0, \forall x \in \mathcal{X}$. Therefore, $f(x) = 0 \ \forall x \in \mathcal{X}$. Hence $f \equiv 0$. Hence $M^{\perp} = \{0\}$. Hense by lemma (3.3), $\overline{span(M)} = \mathcal{F}$.

By virtue of the above theorem and theorem 2.5, every $f \in \mathcal{F}$ can be expressed as

$$f = \sum \alpha_i k_{x_i}, \ \alpha_i \in \mathbb{R}. \tag{4}$$

Hence,

$$f(x) = \sum_i \alpha_i k_{x_i}(x) = \sum_i \alpha_i k(x_i, x) \tag{5}$$

$[\langle f, k_{x_i} \rangle = f(x_i). \text{Hence} \langle k_{x_i}, k_x \rangle = k_{x_i}(x) = k(x_i, x)]$.

Definition (Semi Positive definite function) A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is semi positive-definite if

$$\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0 \tag{6}$$

for all $c_i, c_j \in \mathbb{R}$.

The reproducing kernel k is semi positive definite on $\mathcal{X} \times \mathcal{X}$, since, for any $x_1, x_2, \dots \in \mathcal{X}$ and $a_1, a_2, \dots \in \mathbb{R}$

$$\sum_{i,j} a_i a_j k(x_i, x_j) = \sum_{i,j} a_i a_j \langle k_{x_i}, k_{x_j} \rangle = || \sum a_i k_{x_i} ||^2 \geq 0 \tag{7}$$

The Moore-Aronszajn-Theorem states that for every semi positive definite kernel on $\mathcal{X} \times \mathcal{X}$, there exists a unique RKHS and vice versa.

6.1 Kernel Matrix

Definition (Kernel matrix) Given a kernel k and points $x_1, \dots, x_N \in \mathcal{X}$, the $N \times N$ matrix

$$K = [k(x_i, x_j)]_{ij} \tag{8}$$

is called the kernel matrix (Gram matrix) of k with respect to $x_1, \dots, x_N$.

6.2 Semi Positive Definite Kernels

Definition (Semi Positive definite matrix) A real $N \times N$ symmetric matrix K satisfying

$$c^T K c = \sum_i \sum_j c_i c_j K_{ij} \geq 0 \tag{9}$$

for all $c \in R^N$ is called semi positive definite.[K_{ij} is the $ijth$ element of K]. If equality in (9) only occurs when c is a zero vector, then the matrix is called as positive definite.

A function $k : \mathcal{X} \times \mathcal{X}$ is a reproducing kernel if and only for all $N \in \mathbb{N}, x_i \in \mathcal{X}$, the corresponding kernel matrix K is semi positive definite. A function $k : \mathcal{X} \times \mathcal{X}$ is a kernel iff semi positive definite function.

7 Reproducing Kernel

In this section we'll look into some common kernels.

7.1 Linear Kernel

Let $\mathcal{F}$ be the set of all bounded linear functionals defined on $\mathbb{R}^n$ with kernel k. What is the form of k?

Let $f \in \mathcal{F}$. Therefore

$$f(x) = \langle w_f, x \rangle$$

where w_f is the parameter associated with f. Corresponding to f, there exists a hyperplane $\Pi_{w_f} = \{x \in \mathbb{R}^n : f(x) - \langle w_f, x \rangle = 0\}$. By defining $\langle f, g \rangle = \langle w_f, w_g \rangle$, where w_g is the parameter associated with g, $\mathcal{F}$ is a RKHS.

Consider $k_{x_j} \in \mathcal{F}, j = 1, 2, ...$ Now

$$k_{x_j}(x_i) = \langle w_{k_{x_j}}, x_i \rangle = k_{x_i}(x_j)$$

where $w_{k_{x_j}}$ is the parameter associated with k_{x_j}. This implies for finding the image of a point using k_{x_i} one of the arguments in the inner product should be x_i, $i = 1, 2, \ldots$ Hence $k_{x_i}(x_j) = \langle x_i, x_j \rangle = k(x_i, x_j)$ which is the linear kernel.

7.2 Polynomial Kernel

$k(x, y) = (\langle x, y\rangle + c)^d, c \geq 0, d \in \mathbb{N}$

We will look into the RKHS corresponding with $k(x, y) = (\langle x, y\rangle)^2$, x, $y \in \mathbb{R}^2$.

$$\begin{aligned} k(x,y) &= (\langle x, y\rangle)^2 \\ &= (x_1y_1 + x_2y_2)^2 \\ &= {x_1}^2{y_1}^2 + {x_2}^2{y_2}^2 + 2x_1y_1x_2y_2 = \langle (x_1^2, x_2^2, \sqrt{2}x_1x_2), (y_1^2, y_2^2, \sqrt{2}y_1y_2)\rangle \end{aligned}$$

If we define $\phi : \mathbb{R}^2 \to \mathbb{R}^3$ by $\phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$, then $k_x(y) = \langle \phi(x), \phi(y)\rangle = H_{\phi(x)}(\phi(y))$ where $H_{\phi(x)}$ is a hyperplane in $\mathbb{R}^3$ with parameter $\phi(x)$.

Let $\mathcal{F}$ is the RKHS corresponding to k. Let $f \in \mathcal{F}$. $f(x) = \sum \alpha_i k(x_i, x) = \sum_i \langle \phi(x_i), \phi(x)\rangle = H_{\sum_i \phi(x_i)}(\phi(x))$. Therefore corresponding to f there exists a hyperplane in $\mathbb{R}^3$. Hence $\tilde{f}(x) = f(x) + b = H_{\sum_i \phi(x_i)}(\phi(x)) + b$. Hence the points that is mapped using ϕ can be seperated by a hyperplane in $\mathbb{R}^3$.

Other examples of kernel functions are

$$\begin{aligned} \text{Linear} \qquad & k(x,y) = \langle x, y\rangle \\ \text{Gaussian RBF}(\beta \in \mathbb{R}_+) \qquad & k(x,y) = \exp\left(-\beta||x-y||^2\right) \\ \text{Polynomial } (d \in N, \theta \in \mathbb{R}_+) \qquad & k(x,y) = [(x.y) + \theta]^d \\ \text{Inverse Multiquadratic } (c > 0) \qquad & k(x,y) = \frac{1}{\sqrt{||x-y||^2 + c}} \end{aligned}$$

With K a Gaussian, the dimensionality of the RKHS is infinite, while when K is a polynomial of degree k (eg $K(x, y) = (1 + \langle x, y\rangle)^k$), the dimensionality of the RKHS is finite.

8 Theory of Kernel Methods

As discussed earlier, associated with every RKHS there exists a symmetric semi positive definite function called the kernel function, k. Algorithms that use the concept of the kernel are called kernel methods.

The cost function used in kernel methods is the regularized cost function:

$$J(f) = \frac{1}{N}\sum_{i=1}^{N} V(y_i, f(x_i)) + \lambda||f||_k^2 \tag{10}$$

where V is the loss function, which is differentiable, and λ is the regularization parameter. The loss function $V(y_i, f(x_i))$ measures the error between the predicted value $f(x_i)$ and given output y_i.

The solution $f^* = \arg\min_{f \in \mathcal{F}} J(f)$.

Kernel methods can be divided into different types depending upon the loss function they are using.

It can be proved using the representer theorem that the minimization problem (10) gives the solution of the learning problem in terms of the number of training points. That is

$$f = \sum_{i=1}^{N} \alpha_i k_{x_i}$$

The Representer theorem can be stated as follows:

Theorem 8.1 *Denote $\Omega : [0, \infty) \to \mathbb{R}]$ a strictly a monotonically increasing function, by $\mathcal{X}$ a set, by $c : (\mathcal{X} \times \mathbb{R}^2)^N$ an arbitrary loss function. Then any $f \in$ RKHS $\mathcal{F}$ minimizing the regularized risk functional*

$$c((x_1, y_1, f(x_1)), \ldots, (x_N, y_N, f(x_N))) + \Omega(||f||) \tag{11}$$

admits a representation of the form

$$f(.) = \sum_{i=1}^{N} \alpha_i k_{x_i}. \tag{12}$$

Proof Given f is the minimiser of the regularized risk functional. Let $Y = span(k_{x_i})_{i=1}^N$. As every finite dimensional subspace of a normed space $\mathcal{X}$ is closed in $\mathcal{X}$, Y is closed. Therefore by projection theorem,

$$\mathcal{F} = Y \oplus Y^{\perp}$$

. Hence $f = f_y + f_{y^\perp}, f_y \in Y, f_{y^\perp} \in Y^{\perp}$. Now $f(x_i) = \langle f, k_{x_i} \rangle = \langle f_y, k_{x_i} \rangle$. As $f_y \in Y, f_y = \sum_{i=1}^{N} \alpha_i k_{x_i}$. Therefore $f(x) = f_y(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x)$. Hence $f_{y^\perp}$ has no role in determining the value of f.

Now $||f||^2 = ||f_y + f_{y^\perp}||^2 = (||f_y||^2 + ||f_{y^\perp}||^2) \geq |f_y||^2$ Therefore $||f|| \geq ||f_y||$. Therefore $\Omega(||f||) \geq \Omega(||f_y||)$. Thus f_y satisfies the given points and also has the least value for Ω. Therefore $f = f_y = \sum_{i=1}^{N} \alpha_i k_{x_i}$.

Any function of the form $f = \sum_{i=1}^{N} \alpha_i k_{x_i} + f', f' \in Y^{\perp}$ satisfies the given points, of which $\sum_{i=1}^{N} \alpha_i k_{x_i}$ has the least norm. The significance of the representor theorem is that the number of terms in the minimiser of regularized risk functional depends only of the number of training points, that is, it is independent of the dimensionality of RKHS space.

If $f \in \mathcal{F}, f(x) = \langle f, k_x \rangle$. Is that possible to model a function that generates the data of the form $\tilde{f}(x) = \langle f, k_x \rangle + b, b \in \mathbb{R}$ by making use of kernel theory?. For that we make use of semi parametric representor theorem.

Theorem 8.2 *(Semiparametric Representer Theorem) Suppose that in addition to the assumptions of the previous theorem we are given a set of M real valued functions* $\{\Psi_p\}_{p=1}^{M}$ *on* $\mathcal{X}$ *with the property that the* $N \times M$ *matrix* $(\Psi_p(x_i))_{ip}$ *has rank M. Then any* $\tilde{f} := f + h$ *with* $f \in \mathcal{F}$ *and* $h \in span\{\Psi_p\}$ *minimizing the regularized risk functional*

$$c((x_1, y_1, \tilde{f}(x_1)), \ldots, (x_N, y_N, \tilde{f}(x_N))) + \Omega(||f||) \tag{13}$$

admits a representation of the form

$$\tilde{f}(.) = \sum_{i=1}^{N} \alpha_i k_{x_i} + \sum_{i=1}^{M} \beta_p \Psi_p. \tag{14}$$

where $\beta_p, p = 1, 2 \ldots M$ *are uniquely determined.*

Proof Given $\tilde{f} = f + h$ is the minimiser of the regularized risk functional. Let $Y = span(k_{x_i})_{i=1}^{N}$. As every finite dimensional subspace of a normed space $\mathcal{X}$ is closed in $\mathcal{X}$, Y is closed. Therefore by projection theorem,

$$\mathcal{F} = Y \oplus Y^{\perp}$$

Hence $\tilde{f} = f_y + f_{y^{\perp}} + h, f_y \in Y, f_{y^{\perp}} \in Y^{\perp}$. Now $\tilde{f}(x_i) = \langle f, k_{x_i} \rangle + h(x_i) = \langle f_y, k_{x_i} \rangle + h(x_i)$. As $f_y \in Y, f_y = \sum_{i=1}^{N} \alpha_i k_{x_i}$. Therefore $\tilde{f}(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x) + h(x)$. Hence $f_{y^{\perp}}$ has no role in determining the value of $\tilde{f}$.

Now $||f||^2 = (||f_y + f_{y^{\perp}}||^2 = (||f_y||^2 + ||f_{y^{\perp}}||^2) \geq |f_y||^2$ Therefore $||f|| \geq ||f_y||$. Therefore $\Omega(||f||) \geq \Omega(||f_y||)$. Thus $f_y + h$ satisfies the given points and f_y has the least value for Ω. Therefore $f = f_y = \sum_{i=1}^{N} \alpha_i k_{x_i}$. Hence $\tilde{f} = \sum_{i=1}^{N} \alpha_i k_{x_i} + h$

Now $h(x_i) = \sum_{p=1}^{M} \beta_p \psi_p(x_i), i = 1, 2, \ldots N$. This is a set of N linear equations with M unknowns. This can be represented as $v = \beta_1 v_1 + \beta_2 v_2 + \ldots \beta_M v_m$ where $v_l = (\psi_l(x_1), \psi_l(x_2), \ldots \psi_l(x_N))^T, l = 1, 2, \ldots M$ and $v = (h(x_1), h(x_2), \ldots h(x_N))^T$. By the given conditions $\{v_1, v_2, \ldots v_M\}$ are linearly independent. Therefore the $\beta_i, i = 1, 2, \ldots M$ can be uniquely determined.

Examples

1. For SV regression with the ϵ insensitive loss $c((x_1, y_1, f(x_1)), \ldots, (x_N, y_N, f(x_N))) = \frac{1}{\lambda}\sum_i \max(0, |f(x_i) - y_i| - \epsilon)$ and the regularisor $\Omega(||f||) = ||f||^2$ where $\lambda > 0$ and $\epsilon \geq 0$ are fixed parameters which determine the tradeoff between regularization and fit to the training set. In addition a single $M = 1$ constant function $\Psi_1(x) = c, c \in \mathbb{R}$ is used as an offset that is not regularized by the algorithm.

2. In SVM classification $c((x_1, y_1, f(x_1)), \ldots, (x_N, y_N, f(x_N))) = \frac{1}{\lambda}\sum_i \max(0, 1 - y_i f(x_i)$, the regularisor $\Omega(||f||) = ||f||^2$ and $\Psi_1(x) = c, c \in \mathbb{R}$ is used as an offset that is not regularized by the algorithm.

Hence , for both the above cases, $\tilde{f} := f + b$.

[Refer: **A Generalized Representer Theorem by Bernhard Scholkopf, Ralf Herbrich and Alex J. Smola, Robert Williamson**]

8.1 Kernel Methods: General Form

The solution to (10) has the general form

$$\tilde{f}(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x) + b, \ \alpha_i, b \in \mathbb{R}, x_i, x \in \mathcal{X}. \tag{15}$$

Training a model requires the choice of few relevant quantities:

- the kernel function, that determines the shape of the decision surface;
- a parameter in the kernel function (eg: for gaussian kernel:variance of the Gaussian, for polynomial kernel: degree of the polynomial)
- the regularization parameter λ.

8.2 Hyperplane Models in RKHS

If $\tilde{f}$ is the unknown function of a datamodeling problem, then by kernel theory

$$\tilde{f}(x) = f(x) + b = \langle f, k_x \rangle + b, \forall x \in \mathcal{X}; \ f, k_x \in \mathcal{F}, b \in \mathbb{R} \tag{16}$$

Consider $H_{f,b} : \mathcal{F} \to \mathbb{R}$ where

$$H_{f,b}(g) = \langle f, g \rangle + b, \forall g \in \mathcal{F} \tag{17}$$

Comparing (16) and (17),

$$\tilde{f}(x) = H_{f,b}(k_x), x \in \mathcal{X}, k_x \in \mathcal{F}$$

Thus finding $\tilde{f}$ in input space is equivalent in finding $H_{f,b}$ in RKHS.

Now $\Pi = \{k_x \in \mathcal{F} : H_{f,b}(k_x) = \langle f, k_x \rangle + b\}$ is a hyperplane in RKHS with parameters f and b and its equation can be written as $H_{f,b}(k_x) - \langle f, g \rangle - b = 0$, that is $\tilde{f}(x) - \langle f, g \rangle - b = 0$. Hence Π divides $\mathcal{F}$ into two halves: $\Pi^+ = \{k_x \in \mathcal{F} : \tilde{f}(x) - \langle f, g \rangle - b \geq 0\}$ and $\Pi^- = \{k_x \in \mathcal{F} : \tilde{f}(x) - \langle f, g \rangle - b < 0\}$.

In the case of classification, if x is in positive class in input domain, then k_x is in Π^+ and if x is in negative class in input domain, then k_x is in Π^-. Therefore corresponding to the decision boundary in input space, there exists a decision boundary in RKHS which a hyperplane. Similarly in the case of regression, $(k_x, H_{f,b}(k_x))$ lies in hyperplane Π in RKHS.