Introduction to Kernel Methods

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

Hypothesis Function: Kernel Methods

$$\tilde{f} : \mathcal{X} \to \mathbb{R}$$

- $\tilde{f}(x_i) = f(x_i) + b$
- f lies in Reproducing Kernel Hilbert Space $\mathcal{F}, b \in \mathbb{R}$

RKHS: Basic Concepts

- Functional Analysis
- metric space
    - complete metric space
- normed space
    - Banach Space
- inner product space
    - Hilbert space

Metric Spaces

Definition

A metric space is a pair (X, d), where X is a set and d is a metric on X (or distance function on X), that is, a function defined on $X \times X$ 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)

Examples

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. $L^p[a, b] : d(x, y) = (\int_a^b (|x(t) - y(t)|)^p dt)^{1/p}$
5. l_∞ with the metric $d(x, y) = \sup_{j \in N} |\eta_j - \psi_j|$, where l_∞ is a bounded sequence of complex numbers,
   $x = (\eta_1, \eta_2, \eta_3, \ldots), y = (\psi_1, \psi_2, \psi_3, \ldots)$.

Sequence

- $(x_1, x_2, x_3, \dots)$
- $(1, 1/2, 1/3, \dots)$
- $(1, 1 + x, 1 + x + x^2/2!, \dots)$
- Every element has a particular position

Open and Closed Set

- $B(x_0, \epsilon) = \{x \in \mathcal{X}; d(x, x_0) < \epsilon)\}$
- A subset M of $\mathcal{X}$ is said to be open if for every $x_0 \in M, \exists\ \epsilon > 0$, such that, $B(x_0, \epsilon) \subset M$
- A subset M of $\mathcal{X}$ is said to be closed, if M^c is open

Questions

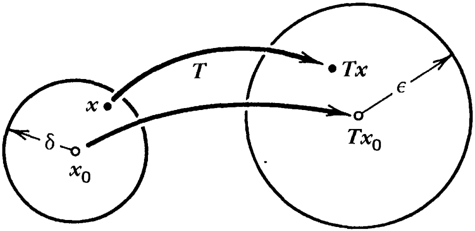
- (0,1)
- $[0, 1]$
- Set of all polynomials defined on $[a, b] \subseteq \mathbb{R}$

Continuous Function

Definition

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

Continuous Function



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 $d(x_n, x_0) < \epsilon$

Questions

- Limit points of (0,1)
- Limit points of set of all rational numbers
- Limit points of set of all irrational numbers

Dense Subset

Definition

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

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

Questions

- Closure of (0,1)
- Closure of set of all rational numbers
- Closure of set of all irrational numbers
- Dense subset of $\mathbb{R}$

Convergent Sequence

Definition

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

Caushy Sequence

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

Every convergent sequence is Caushy.

Proof.

Let (x_n) is a convergent sequence in $\mathcal{X}$ and x_0 be its limit. Choos $\epsilon > 0$. By definition of convergence, $\exists N$ such that $d(x_m, x_0) < \frac{\epsilon}{2}, \forall m > N$. 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. □

Caushy Sequence: Example

$(x_1, x_2, \dots x_N, x_{N+1}, \dots)$ is Caushy
If $\epsilon = 0.5$ and $N = 100$ then,
$d(x_{150}, x_{1000}) < 0.5$

Complete Metric Space

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, C[a,b] 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.

Example

- Let X be the set of all polynomials on some finite closed interval $J = [a, b]$
- $a_1(t) = t + t^2$, $b_1(t) = t + 2t$
- $d(a_1(t), b_1(t)) = \max_{t\in[a,b]}(|a_1(t) - b_1(t)|)$
- $(t, t - \frac{t^3}{3!}, t - \frac{t^3}{3!} + \frac{t^5}{5!}, \dots)$

$L^p[a,b]$

$C[a,b]$ with metric

$d(x(t),y(t) = \left(\int_a^b (|x(t)-y(t)|)^p dt\right)^{\frac{1}{p}}$, $p \geq 1$ is an incomplete metric space

The completion of $C[a,b]$ with the above metric is the space $L^p[a,b]$

Limit Point and Sequence

Theorem

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

Proof.

Let $x \in \overline{M}$. To prove that there exists a sequence $(x_n) \in M$, such that $x_n \to x$: 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. □

- Set of polynomials is dense in C[a,b]
- Set of polynomials is dense in $L^p[a, b]$
- the Weierstrass approximation theorem states that every continuous function defined on a closed interval $[a, b]$ can be uniformly approximated as closely as desired by a polynomial function

Limit Point and Sequence

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

Vector Space

A set V is a vector space over a field K if there exists a structure $\{V, K, +, *)$ consisting of V , K, a vector addition operation + and a scalar multiplication $*$. This structure must obey the following axioms for any $u, v, w \in V$ and $\alpha, \beta \in K$:

- Associative Law: $(u + v) + w = u + (v + w)$.
- Commutative Law: $u + v = v + u$.
- Additive identity: for any vector v in V, 0 + v = v and v + 0 = v.
- Inverse: $\forall u \in V, \exists s \in V$ such that $u + s = 0$
- Unitarity: $1u = u, 1 \in K, u \in V$
- Multiplication by scalars: $\alpha * v \in V$
- Distributive Laws $\alpha * (u + v) = \alpha * u + \alpha * v$ and $(\alpha + \beta) * u = \alpha * u + \beta * v$

Normed Space

Definition

A normed space is a vector space with a norm defined on it. A norm on a vector space X is a function $|| \; || \rightarrow R$ such that:

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

- $\mathbb{R}^n, \mathbb{C}^n$
- $C[a, b] : \|x\| = \max_{t \in [a,b]} |x(t)|$
- $L^p[a, b] : \|x\| = \left(\int_a^b |x(t)|^p dt \right)^{1/p}, p \geq 1$

Banach Space

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

- All normed spaces are metric spaces. But the converse is not true.
- A Banach space is a complete normed space

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 $\alpha \in K$.

Supremum

The supremum (sup) of a set is its least upper bound. $sup(0, 1)$ is 1.

Supremum is called maximum when the least upper bound is a member of the set. Maximum of [0,1] is 1.

Bounded Linear Operator

Definition

Let X and Y be normed spaces and $T : X \rightarrow Y$ a linear operator. The operator T is said to be bounded if there is a real number $c > 0$ such that for all $x \in X$, $||Tx|| \leq c||x||$.

$$||T|| = sup_{x \in X, x \neq 0} \frac{||Tx||}{||x||} \text{ or} ||T|| = sup_{x \in X, ||x||=1} ||Tx||.$$

- $||T||$ is called the operator norm. Operator norm satisfies all the properties of a norm.

Bounded Linear Functional

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

$$f : X \to K$$

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

or else

$$||f|| = sup_{x \in X), ||x|| = 1} ||f(x)||$$

.

Theorem

A linear operator T is continuous iff it is bounded.

Operator Norm

- The operator norm can be defined only for bounded linear operator or functionals
- Operator norm satisfies all the properties of the norm

$$||T|| = sup_{x \in X, x \neq 0} \frac{||Tx||}{||x||} \text{ or} ||T|| = sup_{x \in X, ||x||=1} ||Tx||.$$

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

or else

$$||f|| = sup_{x \in X), ||x||=1} ||f(x)||$$

.

Inner Product Space

Definition

An inner product space is a vector space X with an inner product defined on X. An inner product is a mapping $\langle , \rangle : X \times X \to K$

- $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$
- $\langle x, y \rangle = \overline{\langle y, x \rangle}$
- $\langle x, x \rangle \geq 0$, $\langle x, x \rangle = 0$ *iff* $x = 0$
- $\mathbb{R}^n, \mathbb{C}^n$
- $L^2[a, b]$

$$\langle f, g \rangle = \int_a^b f(t) \overline{g(t)} \, dt$$

Caushy-Schwartz Inequality

$\|\langle x, y\rangle\| \leq \|x\|\|y\|, \forall x, y \in$ inner product space X

Relation Between Metric Space, Normed Space and Inner Product Space

- $||x|| = \sqrt{\langle x, x \rangle}$, $d(x, y) = ||x - y|| = \sqrt{\langle x - y, x - y \rangle}$
- All inner product spaces are normed spaces. The converse is not true.

Hilbert Space

Definition

A Hilbert space is a complete inner product space

Projection Theorem

Theorem

Let Y be a closed subspace of a Hilbert space H. Then $H = Y \oplus Y^{\perp}$.

Every $x \in H$ can be uniquely expressed as $x = y + y', y \in Y, y' \in Y^{\perp}$

The scalar field K of any vector space is taken to be $\mathbb{R}$ for the rest of the slides.

Reisz Representation Theorem

- Every bounded linear functional T on a Hilbert space H can be represented in terms of innerproduct, namely, $\exists z \in H$, such that

  $$T(x) = \langle x, z \rangle, \forall x \in H$$

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

- $T : \mathcal{H} \rightarrow \mathbb{R}$, where $\mathcal{H}$ is a Hilbert space

- T bounded and linear

- $T(x) = \langle x, z \rangle$, z is unique and depends only on T

Checking Reisz Theorem

- Find $\|f\|$ using the formula of operator norm
- The operator norm is defined by

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

$$\frac{|f(x)|}{\|x\|} \leq \|w\| \forall x \in \mathcal{X}$$

Therefore,

$$sup_{x \in \mathbb{R}^n, x \neq 0} \frac{||f(x)||}{||x||} \leq \|w\|$$

$$\|f\| \leq \|w\| \tag{1}$$

Now,

$$f(w) = \langle w, w \rangle = \|w\|^2$$

Therefore, $|f(w)| = \|w\|^2$ and $\dfrac{|f(w)|}{\|w\|} = \|w\|$.

As $\|f\|$ is the supremum of the set ,$\left\{ \dfrac{||f(x)||}{||x||} : x \in \mathbb{R}^n, x \neq 0 \right\}$

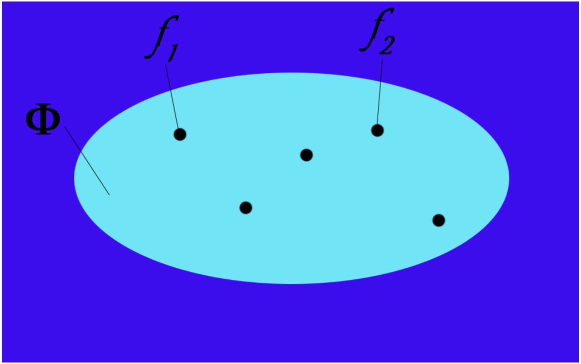
$$\frac{\|f(w)\|}{\|w\|} \leq \|f\|$$

$$\|w\| \leq \|f\| \tag{2}$$

From (1) and (2)

$$\|f\| = \|w\|$$

Space of Functions



Evaluation Functional

Consider $\mathcal{H} = \{f : f : \mathcal{X} \to \mathbb{R}\}$, a Hilbert space of functions.

Definition

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

Point Evaluation Functional or Evaluation Functional

Let $\mathcal{H}$ be a Hilbert space of real-valued functions defined on a domain $\mathcal{X}$.

Definition

The **point evaluation functional** at a fixed point $x \in \mathcal{X}$ is the linear functional

$$L_x : \mathcal{H} \to \mathbb{R}, \quad L_x(f) = f(x), \quad \forall f \in \mathcal{H}.$$

Properties:

- **Linearity:** For any $f, g \in \mathcal{H}$ and scalars a, b,

$$L_x(af + bg) = af(x) + bg(x).$$

Reproducing Kernel Hilbert Space (RKHS)

Definition

A **Reproducing Kernel Hilbert Space (RKHS)** $\mathcal{F}$ is a Hilbert space of functions defined on some set $\mathcal{X}$, in which all point evaluation functionals are bounded linear functionals.

Riesz Theorem on RKHS $\mathcal{F}$

- $L_{x_i} : \mathcal{F} \to \mathbb{R}, \quad x_i \in \mathcal{X}$
  - $L_{x_i}(f) = f(x_i)$
  - By the RKHS property, L_{x_i} is a bounded linear functional on $\mathcal{F}$. By the Riesz theorem:
    - $L_{x_i}(f) = f(x_i) = \langle f, z \rangle_{\mathcal{F}}, \quad \forall f \in \mathcal{F}$, for some $z \in \mathcal{F}$.
    - Define $z = k_{x_i}$, where k_{x_i} is the reproducing kernel function.

Representor of Evaluation

Consider the set of evaluation functionals defined on the RKHS $\mathcal{F}$:

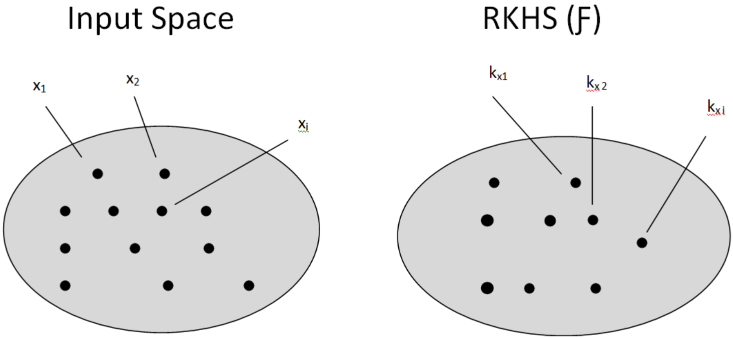
$$\{L_{x_i} : L_{x_i} : \mathcal{F} \to \mathbb{R}\}_{x_i \in \mathcal{X}}.$$

By the properties of RKHS, $\{L_{x_i}\}_{x_i \in \mathcal{X}}$ are bounded linear functionals defined on $\mathcal{F}$. Therefore, by the Riesz representation theorem, there exists a unique $k_{x_i} \in \mathcal{F}$ such that

$$L_{x_i}(f) = f(x_i) = \langle f, k_{x_i} \rangle, \quad \forall f \in \mathcal{F},\ i = 1, 2, \ldots \tag{3}$$

Here, k_{x_i} is called the representor of evaluation at x_i. The representor k_{x_i} depends only on L_{x_i} and the point x_i. Thus, for each $x_i \in \mathcal{X}$, there exists a $k_{x_i} \in \mathcal{F}$.

Association of Input Space and RKHS



- $\phi : \mathcal{X} \to \mathcal{F}$ is a mapping such that $\phi(x) = k_x$.
- The function ϕ is called the **feature mapping.**
- Each point in the input space $\mathcal{X}$ is mapped to a point in the feature space (RKHS) $\mathcal{F}$.

RKHS: Reproducing Property

Consider $f \in \mathcal{F}$. From (3),

$$f(x) = \langle f, k_x \rangle, x \in \mathcal{X}, k_x \in \mathcal{F}, \forall x \in \mathcal{X}$$

This equation is called the reproducing property because the function values are "reproduced" via inner products in the Hilbert space.

RKHS: Reproducing Property

Consider $f \in \mathcal{F}$. From Equation (3), we have:

$$f(x) = \langle f, k_x \rangle_{\mathcal{F}}, \quad \forall x \in \mathcal{X}, \quad k_x \in \mathcal{F}.$$

This equation is called the **reproducing property** because the function values are "reproduced" via inner products in the Hilbert space.

Equivalence of a function defined on $\mathcal{X}$ with a hyperplane equation defined on RKHS

Consider $f \in \mathcal{F}$. Define

$$\tilde{f} : \mathcal{X} \to \mathbb{R}$$

where

$$\tilde{f}(x) = f(x) + b, b \in \mathbb{R}$$

Therefore

$$\tilde{f}(x) = \langle f, k_x \rangle + b = \langle f, \phi(x) \rangle + b$$

Define $H_f : \mathcal{F} \to \mathbb{R}$ where $H_f(g) = \langle f, g \rangle$. H_f is a linear function in RKHS. Now,

$$\tilde{f}(x) = H_f(\phi(x)) + b$$

Classification and Regression in RKHS Settings

- Classification:$y \in \{0, 1\}$
  - $\tilde{f}(x) \geq 0$ implies $H_f(\phi(x)) + b \geq 0$
  - $H_f(g) + b = 0$ is the equation of a hyperplane in RKHS
  - Decision boundary is a hyperplane in RKHS

- Regression:$y \in \mathbb{R}$
  - $\tilde{f}(x) = f(x) + b = y$ implies $H_f(\phi(x)) + b = y$
  - $H_{f,b}(g) + b = \alpha, \alpha \in \mathbb{R}$ is the equation of a hyperplane in RKHS
  - $(\phi(x), y)$ lies in a hyperplane

- Finding $\tilde{f}$ in input space is equivalent to finding H_f and b.

Classification and Regression in RKHS Settings

- **Classification:** $y \in \{0, 1\}$
  - $\tilde{f}(x) \geq 0$ implies $H_f(\phi(x)) + b \geq 0$
  - The decision boundary is given by $H_f(\phi(x)) + b = 0$, which defines a hyperplane in RKHS.

- **Regression:** $y \in \mathbb{R}$
  - $\tilde{f}(x) = f(x) + b = y$ implies $H_f(\phi(x)) + b = y$
  - The equation $H_f(\phi(x)) + b = \alpha, \quad \alpha \in \mathbb{R}$ defines a hyperplane in RKHS.
  - Each data point $(\phi(x), y)$ lies on this hyperplane.

- Finding $\tilde{f}$ in the input space is equivalent to finding H_f and b in RKHS.

Value of k_{x_i}

$$f(x_i) = \langle f, k_{x_i} \rangle \tag{4}$$

$$k_{x_i}(x_j) = \langle k_{x_i}, k_{x_j} \rangle \tag{5}$$

Reproducing Kernel

For each $x_i \in \mathcal{X}$, there exists a unique function $k_{x_i} \in \mathcal{F}$. The values of k_{x_i} are determined using the inner product structure in the RKHS.

The **reproducing kernel** (r.k.) k is defined as:

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$$

such that:

$$k(x, y) = \langle k_x, k_y \rangle = \langle \phi(x), \phi(y) \rangle$$

From Equation (5), we obtain:

$$k(x, y) = k_x(y) = k_y(x)$$

Reproducing Kernel & RKHS

- Associated with every RKHS there exists a unique r.k and vice versa.

Span of a Subset

$M = \{v_1, v_2, \ldots\} \subseteq H$, where H is a Hilbert space.
$Span(M) = \{\sum_i \alpha_i v_i, \alpha_i \in \mathbb{R}, v_i \in M\}$
$M^\perp = \{v' \in H : \langle v, v' \rangle = 0, \forall v \in M\}$

Spanning Property

Lemma

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

That is,

$$\overline{\{span(M)\}} = H, \text{ iff } M^{\perp} = \{0\}$$

Spanning Property of $\{k_{x_i}\}$

Theorem

$\overline{span(M)} = \mathcal{F}$, *where* $M = \{k_{xi}, i = 1, 2, \ldots\} \subseteq$ *RKHS* $\mathcal{F}$.

Proof.

Let $f \in M^\perp$. Therefore

$$\langle f, k_x \rangle = 0, \forall k_x \in M$$

This implies,

$$f(x) = 0 \ \forall x \in \mathcal{X}$$

Hence $f \equiv 0$.Therefore $M^\perp = \{0\}$. Hence by the above lemma, $\overline{span(M)} = \mathcal{F}$. □

Spanning Property of $\{k_{x_i}\}$

Theorem

Every $f \in \mathcal{F}$ can be expressed as

$$f = \sum_{i=1}^{\infty} \alpha_i k_{x_i}, \ \alpha_i \in \mathbb{R}.$$

For proof, we make use of the following result:
$x \in \overline{M}$, if and only if there exists a sequence $(x_n) \in M$, such that $x_n \to x$.

Proof: Spanning Property of $\{k_{x_i}\}$

Proof.

$M = \{k_{xi}, i = 1, 2, \dots\}$. Therefore $\text{span}(M) = \{\sum_i \alpha_i k_{x_i}, \alpha_i \in \mathbb{R}, k_{x_i} \in \mathcal{F}\}$. Now $\overline{\text{span}(M)} = \mathcal{F}$. $f \in \mathcal{F}$. Therefore , there exists a sequence $(f_1, f_2, \dots f_n, \dots) \in \text{span}(M)$ where $f_n = \sum_i^n \alpha_i k_{x_i}$, such that $f_n \to f$. As $f = \lim_{n\to\infty} f_n$

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

□

Kernel Trick

$$f(x) = \langle f, k_x \rangle = \langle \sum_{i=1}^{\infty} \alpha_i k_{x_i}, k_x \rangle = \sum_{i=1}^{\infty} \alpha_i \langle k_{x_i}, k_x \rangle = \sum_{i=1}^{\infty} \alpha_i \langle \phi(x_i), \phi(x) \rangle$$

Therefore

$$f(x) = \sum_{i=1}^{\infty} \alpha_i k(x_i, x)$$

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

Semi Positive definite function

Definition

A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is semi positive-definite if*

$$\sum_{i,j} a_i a_j k(x_i, x_j) \geq 0$$

for all $a_i, a_j \in \mathbb{R}$

Properties of reproducing kernel

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

$$\begin{aligned}
\sum_{i,j} a_i a_j k(x_i, x_j) &= \sum_{i,j} a_i a_j \langle k_{x_i}, k_{x_j} \rangle \\
&= \left\langle \sum_i a_i k_{x_i}, \sum_i a_i k_{x_i} \right\rangle \\
&= || \sum_{i,j} a_i k_{x_i} ||^2 \geq 0
\end{aligned}$$

Semi Positive definite Kernel & RKHS

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.

Kernel Matrix

- Kernel matrix: Given a kernel k and points $x_1, \ldots, x_N \in \mathcal{X}$, the $N \times N$ matrix

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

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

Kernel Matrix

$$K = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & \cdots & k(x_1, x_N) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_N) \\ \vdots & \vdots & \vdots & \vdots \\ k(x_N, x_1) & k(x_N, x_2) & \cdots & k(x_N, x_N) \end{pmatrix}$$

Semi Positive definite matrix

Definition

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

$$c^T Kc = \sum_i \sum_j c_i c_j K_{ij} \geq 0 \tag{6}$$

for all $c \in R^N$ is called semi positive definite.[K_{ij} is the *ijth* element of K]. If equality in (6) only occurs when c is a zero vector, then the matrix is called as positive definite.

Kernel Matrix

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.

Semi Positive Definite Function and Reproducing Kernel

A function $k : \mathcal{X} \times \mathcal{X}$ is a kernel iff it is semi positive definite function.

Set of all linear functionals defined on $\mathbb{R}^n$

- Set of all linear functionals defined on $\mathbb{R}^n$: $\mathcal{H}^* = \{f : \mathbb{R}^n \to \mathbb{R} : f(x) = \langle w_f, x \rangle\}$
- $\mathcal{H}^*$ consists of bounded linear functionals
- Corresponding to each $f \in \mathcal{H}^*$ exists a hyperplane : $\{x \in \mathbb{R}^n : f(x) = \langle w_f, x \rangle\}\}$. That is $f(x) = \langle w_f, x \rangle$ is a equation to a hyperplane in $\mathbb{R}^n$
- $\mathcal{H}^*$ is a Hilbert space with $\|f\| = \|f\|_{operator}$ where $\|f\|_{operator}$ is the operator norm of f.
- $\mathcal{H}^*$ is called the dual space of $\mathbb{R}^n$

$\mathcal{H}^*$

- $L_x : \mathcal{H}^* \rightarrow \mathbb{R}$, $L_x f = f(x)$. Is L_x bounded?

By Reisz theorem $\|f\| = \|w_f\|$.
$L_x(f + g) = (f + g)(x) = f(x) + g(x),$
$L_x(\alpha f) = (\alpha f)(x) = \alpha f(x) = \alpha L_x(f)$
To prove L_x is bounded for all x,
$\|L_x(f)\| = \|f(x)\| = \|\langle w_f, x\rangle\| \leq \|w_f\|\|x\| = \|f\|\|x\|, \forall f \in \mathcal{H}^*$.
$L_x : x \in \mathbb{R}^n$ are bounded linear functionals. $\mathcal{H}^*$ is a RKHS space.

Linear Kernel

Consider $k(x_i, x_j) = \langle x_i, x_j \rangle$, Prove that k is positive semi definite.

Proof.

$$\begin{aligned} \sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) &= \sum_i \sum_j \alpha_i \alpha_j \langle x_i, x_j \rangle \\ &= ||\alpha_i x_i||^2 \geq 0 \end{aligned}$$

k is positive semidefinite. □

RKHS Corresponding to Linear Kernel

Linear Kernel: $k(x_i, x_j) = \langle x_i, x_j \rangle$
As k is positive semidefinite, there exists a RKHS corresponding to k.
By definition $k(x_i, x_j) = k_{x_i}(x_j) = k_{x_j}(x_i)$
$k_{x_i}(x_j) = \langle x_i, x_j \rangle, j = 1, 2, \ldots$ implies k_{x_i} is linear function defined on R^n.
Therefore the RKHS corresponding to the linear kernel consists of all the linear functions defined on R^n, that is the dual space of R^n.

Procedure for finding Affine function (Hyperplane) in input space: Classification & Regression

$\tilde{f}(x) = f(x) + b, b \in \mathbb{R}$ where $f(x) = \langle w, x \rangle$.
$f \in \mathcal{H}^*$ whose kernel is the linear kernel. Therefore

$$f(x) = \langle f, x \rangle = \sum_i \alpha_i k(x_i, x) = \sum_i \alpha_i \langle x_i, x \rangle$$

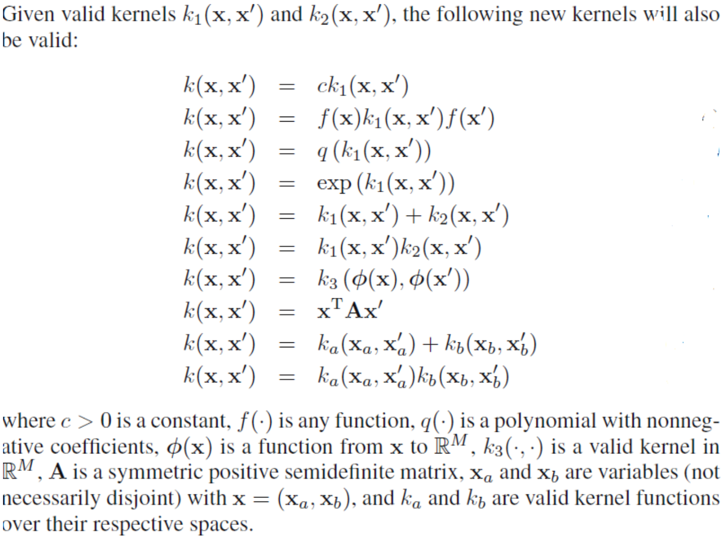
Therefore

$$\tilde{f}(x) = \sum_i \alpha_i \langle x_i, x \rangle + b$$

,

Construction of Kernels

(from Bishop's book)



Given valid kernels $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$, the following new kernels will also be valid:

$$
\begin{aligned}
k(\mathbf{x}, \mathbf{x}') &= ck_1(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= q\left(k_1(\mathbf{x}, \mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= \exp\left(k_1(\mathbf{x}, \mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= k_3\left(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}' \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}_a') + k_b(\mathbf{x}_b, \mathbf{x}_b') \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}_a')k_b(\mathbf{x}_b, \mathbf{x}_b')
\end{aligned}
$$

where $c > 0$ is a constant, $f(\cdot)$ is any function, $q(\cdot)$ is a polynomial with nonnegative coefficients, $\boldsymbol{\phi}(\mathbf{x})$ is a function from $\mathbf{x}$ to $\mathbb{R}^M$, $k_3(\cdot,\cdot)$ is a valid kernel in $\mathbb{R}^M$, $\mathbf{A}$ is a symmetric positive semidefinite matrix, $\mathbf{x}_a$ and $\mathbf{x}_b$ are variables (not necessarily disjoint) with $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$, and k_a and k_b are valid kernel functions over their respective spaces.

Polynomial Kernel

$k(x, y) = (\langle x, y\rangle + c)^d, c \geq 0, d \in \mathbb{N}$
We will look into the RKHS $\mathcal{F}$ corresponding with
$k(x, y) = (\langle x, y\rangle)^2$, $x = (x^{(1)}, x^{(2)})^T$, $y = (y^{(1)}, y^{(2)})^T \in \mathbb{R}^2$.

$$\begin{aligned} k(x,y) &= (\langle x, y\rangle)^2 \\ &= \left(x^{(1)}y^{(1)} + x^{(2)}y^{(2)}\right)^2 \\ &= {x^{(1)}}^2{y^{(1)}}^2 + {x^{(2)}}^2{y^{(2)}}^2 + 2x^{(1)}y^{(1)}x^{(2)}y^{(2)} \\ &= \left\langle \left({x^{(1)}}^2, {x^{(2)}}^2, \sqrt{2}x^{(1)}x^{(2)}\right), \left({y^{(1)}}^2, {y^{(2)}}^2, \sqrt{2}y^{(1)}y^{(2)}\right) \right\rangle \end{aligned}$$

If we define $\tilde{\phi} : \mathbb{R}^2 \to \mathbb{R}^3$ by $\tilde{\phi}(x) = (\left({x^{(1)}}^2, {x^{(2)}}^2, \sqrt{2}x^{(1)}x^{(2)}\right)$, then

$$k(x, y) = k_x(y) = \langle \tilde{\phi}(x), \tilde{\phi}(y)\rangle, \forall x, y \in \mathbb{R}^2 \tag{7}$$

Define

$$H_{\tilde{\phi}(x)} : \mathbb{R}^3 \to \mathbb{R}$$

by

$$H_{\tilde{\phi}(x)}(x') = \langle \tilde{\phi}(x), x' \rangle$$

$H_{\tilde{\phi}}$ is a linear function defined on $\mathbb{R}^3$ and hence $H_{\tilde{\phi}(x')} = \langle \tilde{\phi}(x), x' \rangle$ is the equation of a hyperplane having the parameter $\tilde{\phi}(x)$.

$$k_x(y) = \langle \tilde{\phi}(x), \tilde{\phi}(y) \rangle = H_{\tilde{\phi}(x)}(\tilde{\phi}(y))$$

Therefore, corresponding to each $k_x \in \mathcal{F}$, there exists a linear function defined on $\mathbb{R}^3$.

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

$$\begin{aligned} f(x) &= \sum \alpha_i k(x_i, x) \\ &= \sum_i \alpha_i \langle \tilde{\phi}(x_i), \tilde{\phi}(x) \rangle \text{(from (7))} \\ &= \left\langle \sum_i \alpha_i \tilde{\phi}(x_i), \tilde{\phi}(x) \right\rangle \end{aligned}$$

Define

$$H_{\sum_i \alpha_i \tilde{\phi}(x_i)} : \mathbb{R}^3 \to \mathbb{R}$$

by

$$H_{\sum_i \alpha_i \tilde{\phi}(x_i)}(x') = \left\langle \sum_i \alpha_i \tilde{\phi}(x_i), x' \right\rangle$$

Thus

$$f(x) = H_{\sum_i \alpha_i \tilde{\phi}(x_i)}(\tilde{\phi}(x))$$

$H_{\sum_i \alpha_i \tilde{\phi}(x_i)}$ is a linear function defined on $\mathbb{R}^3$ and hence $H_{\sum_i \alpha_i \tilde{\phi}(x_i)}(x') = \langle \sum_i \alpha_i \tilde{\phi}(x_i), x' \rangle$ is the equation of a hyperplane having the parameter $\sum_i \alpha_i \tilde{\phi}(x_i)$. Therefore corresponding to f there exists a linear function defined on $\mathbb{R}^3$. Hence

$$\tilde{f}(x) = f(x) + b = H_{\sum_i \alpha_i \tilde{\phi}(x_i)}(\tilde{\phi}(x)) + b$$

.

- Classification
  - $$V_1 = \{x' \in \mathbb{R}^3 : H_{\sum_i \alpha_i \tilde{\phi}(x_i)}(x') + b \geq 0\}$$
  - $$V_2 = \{x' \in \mathbb{R}^3 : H_{\sum_i \alpha_i \tilde{\phi}(x_i)}(x') + b < 0\}$$
  - If $\tilde{f}(x_i) \geq 0, \tilde{\phi}(x_i) \in V_1$, $\tilde{f}(x_i) < 0, \tilde{\phi}(x_i) \in V_2$
  - Hence the points that is mapped using $\tilde{\phi}$ can be separated by the hyperplane $H_{\sum_i \alpha_i \tilde{\phi}(x_i)}(x') + b = 0$ in $\mathbb{R}^3$.
- Regression
  - If $\tilde{f}(x) = y$, then $H_{\sum_i \alpha_i \tilde{\phi}(x_i)}(\tilde{\phi}(x)) + b = y$. Therefore $(\tilde{\phi}(x)), y)$ lies on the hyperplane $H_{\sum_i \alpha_i \tilde{\phi}(x_i)}(x') + b = y$

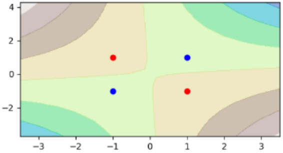


Figure: Non Linear in $\mathbb{R}^2$

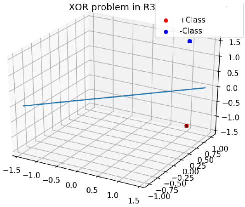


Figure: Hyperplane in $\mathbb{R}^3$

Examples of Kernel Functions

Linear	$k(x, x') = \langle x, x' \rangle$
Gaussian RBF$(\beta \in \mathbb{R}_+)$	$k(x, x') = \exp\left(-\beta \|\|x - x'\|\|^2\right)$
Polynomial $(d \in N, \theta \geq 0)$	$k(x, x') = [(x.x') + \theta]^d$
Inverse Multiquadratic $(c > 0)$	$k(x, x') = \dfrac{1}{\sqrt{\|\|x - x'\|\|^2 + c}}$

Supervised Learning: Kernel Theory

Given data $\{(x_1, y_1), x_2, y_2), \ldots\ldots(x_N, y_N)\}, x_i \in \mathbb{R}^n, y_i \in \mathbb{R}$. Assume $\tilde{f}(x) = f(x) + b, b \in \mathbb{R}, x \in \mathbb{R}^n$ be the function that generates the data. Assume f belings to RKHS $\mathcal{F}$ with k.

$$\tilde{f}(x) = \langle f, k_x \rangle + b = \left\langle \sum_i \alpha_i k_{x_i}, k_x \right\rangle + b \tag{8}$$

The RHS of (8) corresponds to a hyperplane in RKHS

Overfitting and Underfitting

Example: Polynomial curve fitting

$y(x,\mathbf{w}) = w_0 + w_1x + w_2x^2 + \ldots + w_Mx^M$

Regression: Learn parameters $\mathbf{w} = (w_1, \ldots, w_M)$

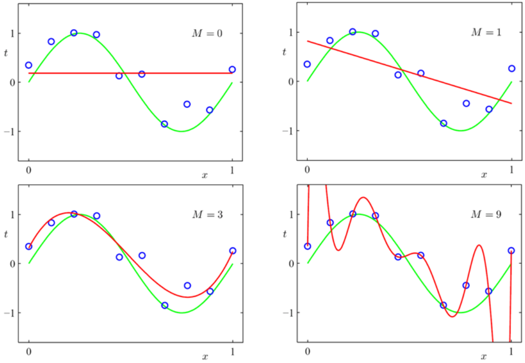


Image taken from Bishop’s Book

Smooth Function

- Small change in input corresponds to small change in output
- $f \in \mathcal{F}$

$$||f(x) - f(x')|| = \|\langle f, k_x \rangle - \langle f, k_{x'} \rangle\| = \|H_f(k_x) - H_f(k_{x'})\|$$

$$\|H_f(k_x) - H_f(k_{x'})\| = ||\langle f, k_x - k_{x'} \rangle|| \leq ||f|| \; ||k_x - k_{x'}||$$

- ||f|| small guarantees function to be smooth

Optimisation

$$\min_{f \in \mathcal{F}} \quad \frac{1}{N} \sum_{i=1}^{N} V(y_i, f(x_i))$$

subject to$\|f\|^2 \leq k$

Cost Function

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

  $$J(f) = \frac{1}{N}\sum_{i=1}^{N} V(y_i, f(x_i)) + \lambda ||f||^2 \tag{9}$$

  where V is the loss function, which is differentiable, and $\lambda > 0$ 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 $f^* = \arg\min_{f \in \mathcal{F}} J(f)$. The cost function J is convex. Therefore there exists a unique minimiser.
- Kernel methods can be divided into different types depending upon the loss function they are using.

Representation

Using the representer theorem that the minimization problem (9) gives the solution of the learning problem in terms of the number of training points. That is

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

$$f(x) = \sum_{i=1}^{N} \alpha_i k(x_i, x)$$

Representor Theorem

The Representer theorem can be stated as follows:

Theorem

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

$$V((x_1, y_1, f(x_1)), \ldots, (x_N, y_N, f(x_N))) + \Omega(||f||)$$

admits a representation of the form

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

Representor Theorem

Theorem

Any $f \in \mathcal{F}$ that minimizes

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

is of the form

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

Proof: Representor Theorem

Given f is the minimiser of the regularized risk functional. It is unique as J is convex. 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 a closed subspace of $\mathcal{F}$. 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

$$\begin{aligned} f(x_i) &= \langle f, k_{x_i} \rangle \\ &= \langle f_y, k_{x_i} \rangle \end{aligned}$$

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

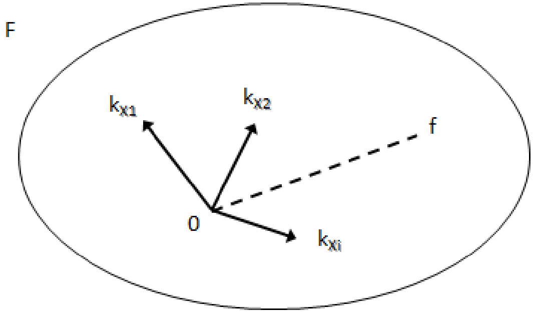
$$\begin{aligned} ||f||^2 &= ||f_y + f_{y^\perp}||^2 \\ &= (||f_y||^2 + ||f_{y^\perp}||^2) \\ &\geq |f_y||^2 \end{aligned}$$

Therefore $||f|| \geq ||f_y||$. Thus f_y satisfies the given points and also has norm less than or equal to f. Therefore f_y is a better solution of J than f. Given f is the unique minimizer. Therefore, $f \equiv f_y$. Therefore

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

Representation of the solution

By using *N* points, the projection of *f* onto the subspace of $\{k_{x_1}, k_{x_2}, \ldots k_{x_n}\}$ is determined using the regularized cost function.



Significance of Representor Theorem

- 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.

Inclusion of Bias

- 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.

Optimisation

$$\min_{f\in\mathcal{F},b\in\mathbb{R}} \quad \frac{1}{N}\sum_{i=1}^{N} V(y_i,\tilde{f}(x_i))$$

subject to $\|f\|^2 \leq k$

Semiparametric 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 $(\Psi(x_i))_{ip}$ has rank M. Then any $\tilde{f} := f + h$ with $f \in \mathcal{F}$ and $h \in \mathrm{span}\{\Psi_p\}$ minimizing the regularized risk functional

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

admits a representation of the form

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

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

Semiparametric Representor Theorem

Theorem

Consider span $\{\Psi\}$*, where* $\Psi(x) = c, \forall x \in \mathcal{X}$*. Any* $\tilde{f} := f + h$ *with* $f \in \mathcal{F}$ *and* $h \in$ *span*$\{\Psi\}$ *that minimizing the regularized risk functional*

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

is of the form

$$\tilde{f} = \sum_{i=1}^{N} \alpha_i k_{x_i} + \beta \psi$$

where β *is uniquely determined.*

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 $\tilde{f} = f_y + f_{y^\perp} + h, f_y \in Y, f_{y^\perp} \in Y^{\perp}$.
Also

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

$$\begin{aligned}
\tilde{f}(x_i) &= f_y(x_i) + f_{y^\perp}(x_i) + h(x_i) \\
&= \langle f_y, k_{x_i} \rangle + \langle f_{y^\perp}, k_{x_i} \rangle + h(x_i) \\
&= \langle f_y, k_{x_i} \rangle + h(x_i)
\end{aligned}$$

Hence $f_{y^\perp}$ has no role in determining the value of $\tilde{f}$. That is $f_y + h$ also satisfies the given points.

Now

$$\begin{aligned} ||f||^2 &= (||f_y + f_{y^\perp}||^2 \\ &= (||f_y||^2 + ||f_{y^\perp}||^2) \geq |f_y||^2 \end{aligned}$$

Therefore $||f|| \geq ||f_y||$. Thus $f_y + h$ satisfies the given points and f_y has the norm less than or equal to f. Therefore $f_y + h$ is a better solution of J than $\tilde{f}$. Given $\tilde{f}$ is the minimiser. Therefore $\tilde{f} = f_y + h$. Hence $\tilde{f} = \sum_{i=1}^{N} \alpha_i k_{x_i} + h$

As h lies in a one dimensional space spanned by ψ, there exists a unique $\alpha \in \mathbb{R}$ such that $h = \alpha\psi$ Hence

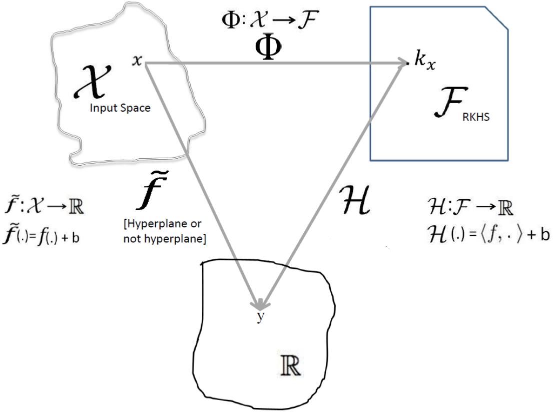
$$\tilde{f} = \sum_{i=1}^{N} \alpha_i k_{x_i} + \beta\psi$$

$$f(\tilde{x}) = \sum_{i=1}^{N} \alpha_i k(x_i, x) + b$$

where $b = \beta\psi(x)$

Frove the above two proofs, it is clear that the number of functions in RKHS that satisfies the given data points is equal to the cardinality of $Y^{\perp}$. That is $f_y + f', f' \in Y^{\perp}$, satisfies the given points. Among that $f_y + f_{y^{\perp}}$ has the smallest norm. Thus adding $||f||^2$ help to get a unique solution.

Kernel Methods



$\Phi: \mathcal{X} \rightarrow \mathcal{F}$
Φ
x
$\mathcal{X}$
Input Space
k_x
$\mathcal{F}$
RKHS
$\tilde{f}: \mathcal{X} \rightarrow \mathbb{R}$
$\tilde{f}(.) = f(.) + b$
$\tilde{f}$
[Hyperplane or not hyperplane]
$\mathcal{H}$
$\mathcal{H}: \mathcal{F} \rightarrow \mathbb{R}$
$\mathcal{H}(.) = \langle f, . \rangle + b$
y
$\mathbb{R}$