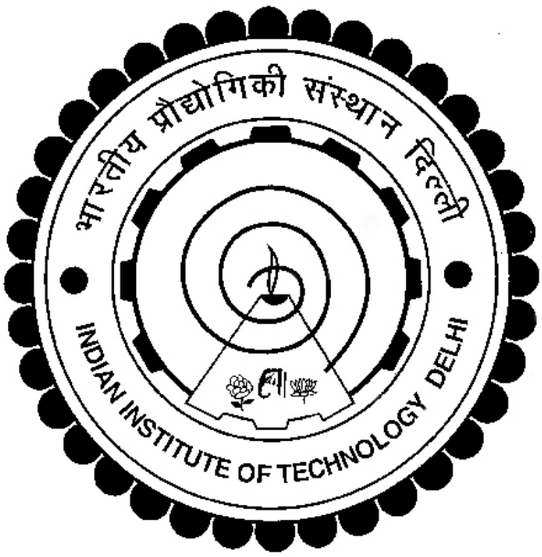


An Introduction to the theory of Reproducing Kernel Hilbert Spaces

Supervisor: Prof. Aparajita Dasgupta

Report by: Rohit Jain (2019MT10721)

Nishant Kumar (2019MT10708)

February 24, 2023

MTD421 - B.Tech Project

Contents

1 Introduction **2**

1.1 Background about Kernels 2

1.2 Kernel Methods: An overall picture 3

2 Important definitions and theorems **4**

2.1 Inner Product Space 4

2.2 Bounded Linear Functionals 4

2.3 Reisz Representation Theorem 4

3 Reproducing Kernel Hilbert Space **5**

3.1 Kernel implies embedding 5

3.2 Formal Definition 6

4 Kernel Matrix **7**

5 Theory of Kernel methods and its applications **8**

5.1 Representer Theorem 8

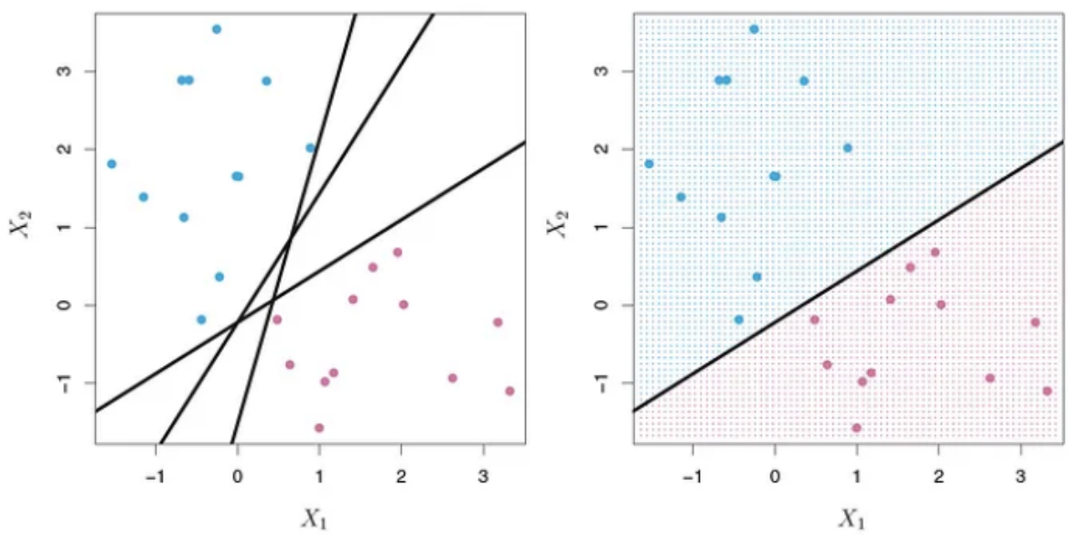
5.2 Kernel Methods: General Form 9

5.3 Hyperplanes in RKHS 9

5.4 Future goal 10

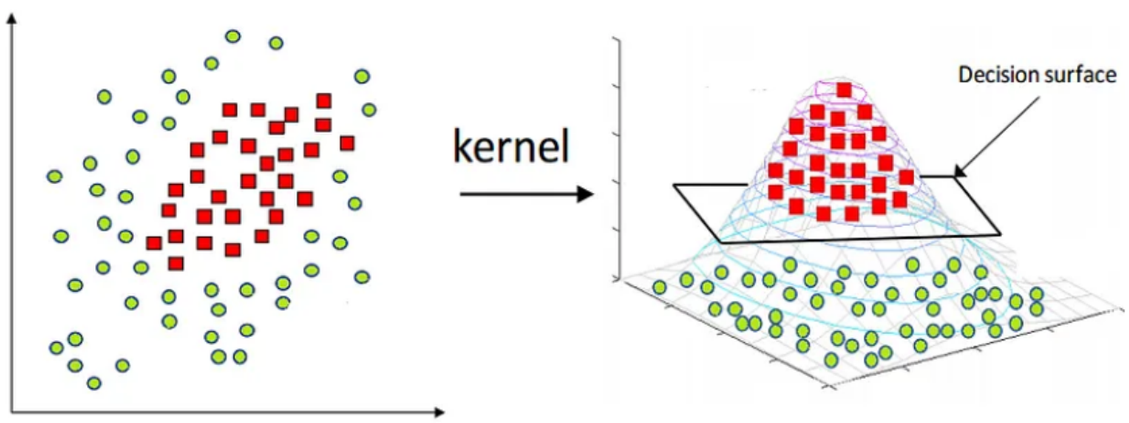
1 Introduction

1.1 Background about Kernels



In the graph above, we notice that there are two classes of observations: the blue points and the purple points. There are tons of ways to separate these two classes as shown in the graph on the left. However, we want to find the "best" hyperplane that could maximize the margin between these two classes, which means that the distance between the hyperplane and the nearest data points on each side is the largest.

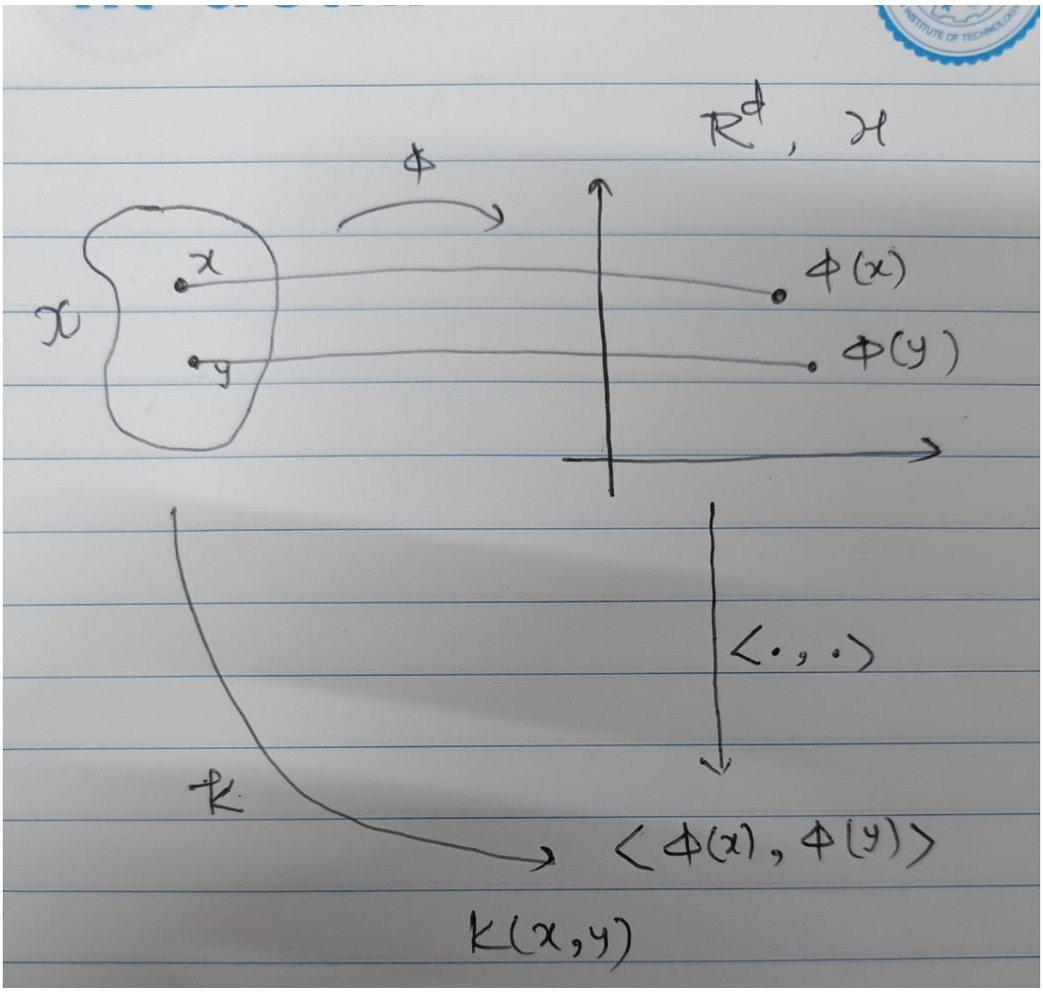
It sounds simple in the example above. However, not all data are linearly separable. In fact, in the real world, almost all the data are randomly distributed, which makes it hard to separate different classes linearly.



As you can see in the above picture, if we find a way to map the data from 2-dimensional space to 3-dimensional space, we will be able to find a decision surface that clearly divides between different classes.

However, when there are more and more dimensions, computations within that space become more and more expensive. This is when the kernel trick comes in. It allows us to operate in the original feature space without computing the coordinates of the data in a higher dimensional space.

1.2 Kernel Methods: An overall picture



In the above figure, we have input space X, which basically maps to points to a higher dimensional space using defined map ϕ. In general Map ϕ can be an infinite dimension and exactly predicting ϕ map is a difficult task (or sometimes impossible in case of non-deterministic/noisy tasks), also we don't need to know exact ϕ function, at the end of classification/regression method calculation all ϕ can be converted to $\phi\phi^T$ which is basically a pairwise inner product of all mapped points. These calculations have really high complexity and can be bypassed by using Kernelization function K which is equivalent to $K(x, y) = \langle \phi(x), \phi(y) \rangle$. In doing so we have to only deal with input space.

2 Important definitions and theorems

2.1 Inner Product Space

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 x Xinto the scalar field K of X; that is every pair of vectors x and y there is associated a scalar, which is written as $\langle x, y\rangle$ and is called the inner product of x and y, such that for all vectors x, y and scalars α we have

- $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$
- $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$
- $\langle x, y\rangle = \overline{\langle y, x\rangle}$
- $\langle x, x\rangle \geq 0, \langle x, x\rangle = 0, \iff x = 0$

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

A space X is said to be a complete metric space if every Caushy sequence converges in X.

A Hilbert space is a complete inner product space

2.2 Bounded Linear Functionals

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

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

2.3 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 $L_x : F \to R$ such that $L_x(f) = f(x); \forall f \epsilon F$.

3 Reproducing Kernel Hilbert Space

To see the beauty of RKHS before we dive into the theory of Reproducing kernel Hilbert spaces, We will look into some more properties of kernels for which they have become the go-to in many Machine learning applications.

3.1 Kernel implies embedding

Theorem: A function $k : X$ x $X \to \mathbb{R}$ is a kernel if and only if there exists a Hilbert space H and a map ϕ: $X \to H$ such that

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

proof: The proof can be broken down into two parts, We will prove the above theorem by showing that the existence of ϕ implies the existence of a kernel k; and the existence of a kernel k implies the existence of a transformation ϕ.

$\Leftarrow$ We define the function k as $k : X$ x X to $\mathbb{R}$ such that, $k(x, y) = \langle \phi(x), \phi(y) \rangle$, we can check the function is well defined because ϕ is well defined as well as satisfy the properties of a kernel.

$\Rightarrow$ **(Construction of RKHS)** For this part of the proof, we will construct a Hilbert space of functions and show the existence of a transformation ϕ corresponding to the given kernel k.

We define a mapping $\phi : X \to \mathbb{R}^x$ (where R^x denotes the space of all real-valued functions from $X \to \mathbb{R}$)

$k_x : X \to \mathbb{R}$, $k_x(y) = k(x, y)$.

Now Since the space of input points has finite real-world data points, we convert the space of points after the transformation into a Vector space Consider the spanning set of all image points in higher dimension

$$G := \{\sum_{i=1}^{r} \alpha_i k(x_i, .) \mid \alpha_i \epsilon \mathbb{R},\ r \epsilon \mathbb{N},\ x_i \epsilon X\}$$

Now we define the Inner Product in our Vector space as:

For points from the dataset: $\langle k_x, k_y \rangle = \langle k(x, .), k(y, .) \rangle := k(x, y)$

In general, the inner product becomes

$$\text{If } g = \sum_i \alpha_i k(x_i, .) \text{ and } f = \sum_j \beta_j k(y_i, .)$$

$$\text{then } \langle f, g \rangle_G := \sum_{i,j} \alpha_i \beta_j k(x_i, y_j)$$

Now We have a space of functions but it does not guarantee the completeness of the space thus we take Closure of the space to include all the limit points of the Cauchy Sequences in the space. **Hence We have created a Hilbert space that has the property** $\langle k_x, k_y \rangle = \langle \phi(x), \phi(y) \rangle$ **(By Construction) This constructed space is called "Reproducing Kernel Hilbert Space"**

3.2 Formal Definition

An RKHS F, is a Hilbert space of functions on some set X in which all the point evaluations are bounded linear functionals.

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 $L_{x_i} : \mathcal{F} \to \mathbb{R}$ 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 $\{k_{x_i}\} \subseteq \mathcal{F}$ 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: $\phi : \mathcal{X} \to \mathcal{F}$ such that $\phi(x) = k_x$ and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

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

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. The reproducing kernel is symmetric, that is $k(x, y) = k(y, x), x, y \in \mathcal{X}$.

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.

4 Kernel Matrix

Definition (Semi Positive Definite function) A function k: X x $X \rightarrow R$ is semi-positive definite if

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

for all $c_i, c_j \in R$. The reproducing kernel k is semi positive definite on X x X, since for any $x_1, x_2, ... \in X$ and $a_1, a_2, ... \in R$

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

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

Definition (Kernel matrix) Given a kernel k and points $x_1, ..., x_n \in X$, the N x X matrix

$$k = [k(x_i, x_j)]_i j$$

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

Definition (Semi Positive definite matrix) A real N x N symmetric matrix K satifying

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

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

A function k:X x X is a reproducing kernel if and only if for all N $\in N, x_i \in X$, the corresponding kernel matrix is semi-positive definite.

Examples of Kernel functions :

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

5 Theory of Kernel methods and its applications

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

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

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 optimum $f^* = \arg\, min_{f\epsilon F} J(f)$

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

It can be proved using the representer theorem that the above minimization problem gives the solution to 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:

5.1 Representer Theorem

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

5.2 Kernel Methods: General Form

The solution to minimising the above cost function has the general form

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

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

5.3 Hyperplanes in RKHS

The equation of the hyperplane in $\mathbb{R}^n$ is $\langle w, x \rangle + b = 0$ ‘w and b can be considered as its parameters.

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

$$\tilde{f}(x) = f(x) + b = \langle f, k_x \rangle + b, \forall x \epsilon X f, k_x \epsilon F, b \epsilon \mathbb{R}$$

Now $H(.) = \langle f, . \rangle i + b$ is a hyperplane in RKHS with parameter f and b. As $\tilde{f}(x) = H(k_{x)}$, **finding $\tilde{f}$ in input domain is equivalent to finding a hyperplane in RKHS.** (see the figure)

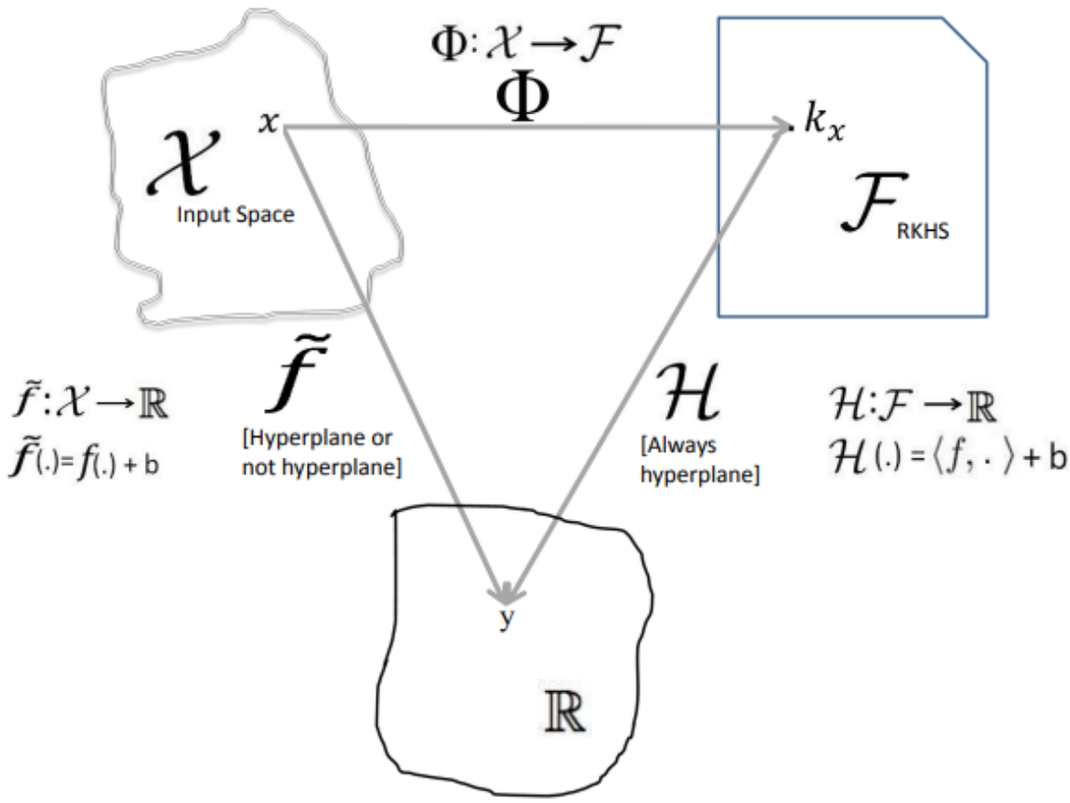


Figure 1: RKHS Mapping

5.4 Future goal

We have done theoretical study of RKHS method, now we'll move towards its applications, here are some important applications of RKHS method

- Kernel Methods in Machine Learning: Kernel methods are a popular approach in machine learning that uses RKHS as a tool for learning and prediction. We can explore the use of RKHS in support vector machines (SVMs) or kernel-based clustering algorithms to classify data points or identify patterns in datasets.
- Image Analysis and Computer Vision: RKHS can be used in image analysis and computer vision to extract useful features from images or videos. We can investigate the use of RKHS in image denoising, image segmentation, object detection, and recognition.
- Control Theory and Robotics: RKHS can also be applied to control theory and robotics to design control systems that are robust to disturbances and uncertainties. We can explore the use of RKHS in adaptive control, robust control, and nonlinear control systems.

- Time Series Analysis: RKHS can be used in time series analysis to model and predict time-dependent data. We can investigate the use of RKHS in time series forecasting, anomaly detection, and signal processing.

- Functional Data Analysis: RKHS can be used in functional data analysis to analyze datasets where the observations are functions rather than discrete data points. We can explore the use of RKHS in functional regression, classification, and clustering.

These are just a few examples of real-life applications that we can explore in the RKHS domain.