11- Kernel Methods: Basics

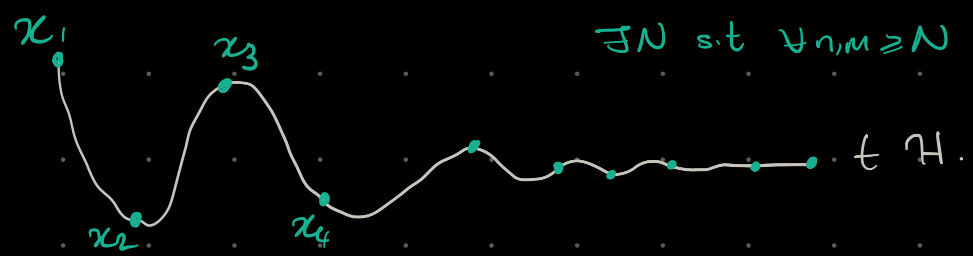
- Hilbert Space:

A Hilbert space $\mathcal{H}$ is a real (or complex) inner product space that is also a complete metric space wrt the norm induced by its inner product.

Remark: Two key properties: 1- inner product 2- completeness

- Complete: Every Cauchy sequence in $\mathcal{H}$ has a limit in $\mathcal{H}$.

$\exists N$ s.t $\forall n,m \geq N$ $d(x_n, x_m) \leq \varepsilon$.



Def (Inner prod): An inner product is a fnc $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ satisfying for $f, g, h \in \mathcal{H}$

1- Symmetry: $\langle f, g \rangle = \langle g, f \rangle$

2- Linearity: $a, b \in \mathbb{R}$ $\langle af + bg, h \rangle = a\langle f, h \rangle + b \cdot \langle g, h \rangle$

3- Non-negativity: i) $\langle f, f \rangle \geq 0$

ii) $\langle f, f \rangle = 0 \iff f = 0$.

Norm induced by the inner product $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle}$.

Ex (Euclidean space): $\mathcal{H} = \mathbb{R}^d$ and standard inner prod

$u, v \in \mathbb{R}^d$, $\langle u, v \rangle = \sum_{i=1}^{d} u_i v_i$ which defines a norm $\|u\|_2 = \sqrt{\sum_{i=1}^{d} u_i^2}$.

Ex (Square integrable fncs on $[0,1]$):

$$L^2([0,1]) = \left\{ f : [0,1] \to \mathbb{R} \text{ and } \int_0^1 f(x)^2 dx < \infty \right\}$$

with inner prod $\langle f, g \rangle = \int_0^1 f(x) g(x)\, dx$.

Def (Dual space): Dual space $\mathcal{H}^*$ of a Hilbert space $\mathcal{H}$ is the space of all cnts and linear fncs from $\mathcal{H}$ to $\mathbb{R}$.

It carries a norm, $F \in \mathcal{H}^*$ $\|F\|_* = \sup_{\|x\|_{\mathcal{H}}=1} |F(x)|$.

Def (Linear fnc): A fnc $f : X \to \mathbb{R}$ is linear if for $x, x' \in X$ and any $c \in \mathbb{R}$, it satisfies

$$f(x+x') = f(x) + f(x').$$

Remark: $f(x) = ax$ is linear, but $f(x) = ax+b$ is not for $b \neq 0$.

Ex (Euclidean space): $\mathcal{H} = \mathbb{R}^d$, then its dual $\mathcal{H}^*$

$$\mathcal{H}^* = \left\{ F : \mathbb{R}^d \to \mathbb{R} \text{ where } F \text{ is linear (and cnts)} \right\}$$

Fncs of the form $F(x) = \langle x, u \rangle$ for some $u \in \mathbb{R}^d$ satisfy the condition in $\mathcal{H}^*$. Are there any more? — $*$

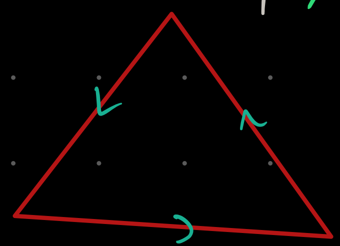
Thm (Riesz – Fréchet Representation): For every $f \in \mathcal{H}$, $\exists F_f \in \mathcal{H}^*$ unique st $F_f(g) = \langle f, g \rangle$. Also, for every $F \in \mathcal{H}^*$ $\exists f_F \in \mathcal{H}$ unique s.t. $F(g) = \langle f_F, g \rangle$.

→ No! $\quad \mathcal{H}^* = \left\{ F_u(x) = \langle u, x \rangle \quad \underbrace{u}_{f} \in \underbrace{\mathbb{R}^d}_{\mathcal{H}} \right\}$

Dual norm $\quad \|F\|_* = \sup_{\|x\|_2 = 1} |\langle u, x\rangle| = \|u\|_2 .$

– Kernels : Formal Definition

Feature map ϕ



→ Goal.

Kernel $k(\cdot,\cdot)$ $\qquad$ RKHS $\mathcal{F}$

Def (Feature map) A fnc $\phi : \mathcal{X} \longrightarrow \mathcal{H}$ where $\mathcal{X}$ is the input space and $\mathcal{H}$ is a Hilbert space.

Def (Kernel): A kernel is a fnc $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ s.t. for any $x_1, \dots x_n \in \mathcal{X}$, the matrix $K_{ij} = k(x_i, x_j)$ is PSD.

Ex: Linear kernel $\quad k(x, x') = \langle x, x' \rangle \qquad \mathcal{X} = \mathbb{R}^d$

Any $x_1 \dots x_n \in \mathbb{R}^d$, $\qquad K_{ij} = \langle x_i, x_j \rangle$

$$X = \begin{bmatrix} - x_1^T - \\ \vdots \\ - x_n^T - \end{bmatrix} \qquad K = X X^T \qquad \forall u \quad \langle u, Ku \rangle = \|X^T u\|_2^2 \geq 0.$$

Thm ($\phi \to k$): A feature map $\phi : X \to \mathcal{H}$ defines a kernel.

proof: – $k(x, x') = \langle \phi(x), \phi(x') \rangle$

– For any $x_1 \dots x_n$ $\quad K_{ij} = k(x_i, x_j)$ is PSD. $\boxtimes$

Thm ($k \to \phi$): For every kernel $k : X \times X \to \mathbb{R}$,
$\exists\, \mathcal{H}$ a Hilbert space and a feature map $\phi : X \to \mathcal{H}$ s.t.

$$k(x, x') = \langle \phi(x), \phi(x') \rangle.$$

proof (for finite X): Let $\mathcal{X} = \{x_1, \dots x_n\}$ and

$$K_{ij} = k(x_i, x_j) \Rightarrow K \text{ is PSD} \Rightarrow K = UDU^T = \Phi \Phi^T$$

$$\phi(x_i) = D^{1/2} u_i \quad \text{defines a feature map.} \quad \boxtimes$$

Remark: The choice of ϕ is not unique. $\phi'(x) = Q\phi(x)$ s.t. Q is orthogonal $Q^T Q = I$.

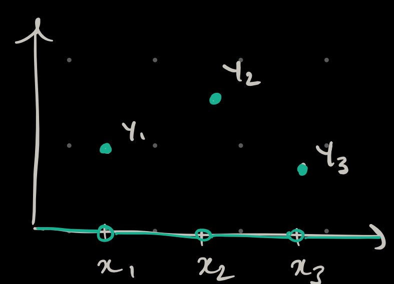
– Hilbert Space defined by the 'Reproducing Kernel'

For a dataset (y_i, x_i) $i = 1 \dots n$ and $\mathcal{F} = L^2([0,1])$.

We consider

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2$$

– Choose f



$$f(x) = \sum_{i=1}^{n} y_i \mathbb{1}_{\{x = x_i\}}$$

$$\|f\|_{\mathcal{F}} = 0.$$

$\Rightarrow$ 0 training error

$\Rightarrow$ overfitting

- Clearly, a Hilbert space is too complex (even has indicators).

Def (Evaluation functional): For a Hilbert space of fncs $h: \mathcal{X} \to \mathbb{R}$, $\forall x \in \mathcal{X}$, the evaluation functional $\mathcal{E}_x : \mathcal{H} \to \mathbb{R}$ is defined as $\mathcal{E}_x(h) = h(x)$.

Remark: Evaluation functionals are linear.

$h, h' \in \mathcal{H}$ $\quad \mathcal{E}_x(h+h') = (h+h')(x) = h(x) + h'(x) = \mathcal{E}_x(h) + \mathcal{E}_x(h')$

$c \in \mathbb{R}$ $\quad \mathcal{E}_x(c \cdot h) = (c \cdot h)(x) = c \cdot h(x) = c \cdot \mathcal{E}_x(h)$.

Ex (Euclidean input space): $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \{ h_\theta(x) = \langle \theta, x \rangle \ \theta \in \mathbb{R}^d \}$

$$\mathcal{E}_x(h_\theta) = h_\theta(x) = \langle \theta, x \rangle .$$

Def (RKHS): An RKHS $\mathcal{F}$ is a Hilbert space over fncs $f: \mathcal{X} \to \mathbb{R}$ s.t. evaluation fnc'als are Lipshitz.

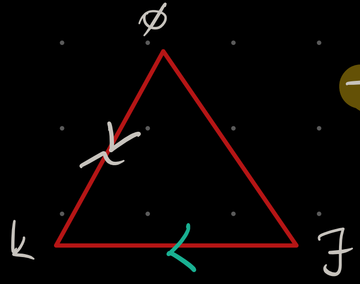
Remarks: - The constraint on the eval fnc'als restricts $\mathcal{F}$. For example, indicators no longer belong to $\mathcal{F}$.

- Eval fnc'els are cnts and linear $\Rightarrow \mathcal{E}_x \in \mathcal{F}^*$
- By Riesz-Fréchet, $\exists f_{\mathcal{E}_x} \triangleq \Psi_x \in \mathcal{F}$ s.t.

$$\forall g \quad \underbrace{g(x) = \mathcal{E}_x(g) = \langle \Psi_x, g \rangle}_{\text{dual}} .$$

! Ψ_x is called the representer

- Fnc evaluations can be written as inner products!



Thm $(\mathcal{F} \to k)$: Every RKHS $\mathcal{F}$ defines a unique kernel.

proof: - RKHS $\Rightarrow$ eval fnc'als $\mathcal{E}_x$ are Lipschitz cnts.
$\Rightarrow \mathcal{E}_x \in \mathcal{F}^*$

- By Riesz-Fréchet, $\exists \psi_x \in \mathcal{F}$ unique s.t.
$\forall f \in \mathcal{F}, \quad \mathcal{E}_x(f) = \langle f, \psi_x \rangle = f(x). \quad (*)$

(ψ_x is the representer, $(*)$ is the <u>reproducing property</u>)

- Since $\psi_x \in \mathcal{F}$, for $x' \in \mathcal{X}$, by $(*)$

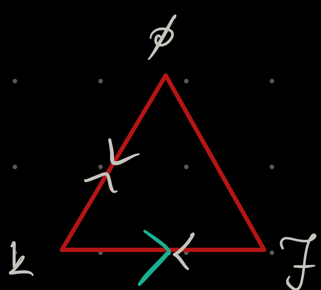
$$\psi_x(x') = \mathcal{E}_{x'}(\psi_x) = \langle \psi_x, \psi_{x'} \rangle \triangleq k(x, x').$$

- $\psi_x \in \mathcal{F}$, $\psi_x = \phi(x)$ defines a feature map, which defines a kernel. $\boxtimes$

Remark: RKHS $\mathcal{F}$ defines a unique kernel k called the <u>reproducing kernel.</u>

$$f(x) = \mathcal{E}_x(f) = \langle f, \psi_x \rangle = \langle f, k(x, \cdot) \rangle$$

representer



Thm $(k \to \mathcal{F}$; Moore-Aronszajn$)$:
Every kernel corresponds to a unique RKHS.

proof: – Basic idea: Use $k(x,\cdot)$ as basis for RKHS.

– Let $n \in \mathbb{N}$, $x_1 \dots x_n \in \mathcal{X}$, $\alpha_i, \theta_i \in \mathbb{R}$.

– Consider $f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$ and $g(x) = \sum_{i=1}^{n} \theta_i k(x, x_i)$

– $\mathcal{F} = \left\{ f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) : n \in \mathbb{N}, x_1 \dots x_n \in \mathcal{X}, \alpha_i \in \mathbb{R} \right\}$

(this is a vector space, but not necessarily complete)

– Define the fnc $\langle \cdot, \cdot \rangle : \mathcal{F} \times \mathcal{F} \longrightarrow \mathbb{R}$ as

$$\langle f, g \rangle \triangleq \sum_{ij} \alpha_i \theta_j k(x_i, x_j)$$

This defines an inner product:

1. Symmetry ✓
2. Linearity ✓
3. Non-negativity: i) $\langle f, f \rangle = \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) = \alpha^T K \alpha \geq 0$.

ii) $\langle f, f \rangle = 0 \iff f = 0$.

trivial

$\longrightarrow$ Define $c(x) = [k(x, x_1) \dots k(x, x_n)]^T \in \mathbb{R}^n$

Augmented kernel matrix for $\{x_1 \dots x_n, x\}$

$$K' = \begin{bmatrix} K & c(x) \\ c(x)^T & k(x,x) \end{bmatrix} \succeq 0$$

Assume $\langle f, f \rangle = \alpha^T K \alpha = 0$ but $f \neq 0$ $(\Rightarrow \alpha \neq 0)$

For any scalar $b \in \mathbb{R}$, let $u = \begin{bmatrix} \alpha \\ b \end{bmatrix} \in \mathbb{R}^{n+1}$

$$u^T K' u = \begin{bmatrix} \alpha \\ b \end{bmatrix}^T \begin{bmatrix} K & c(x) \\ c(x)^T & k(x,x) \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \underbrace{\alpha^T K \alpha}_{=0} + 2b\alpha^T c(x) + b^2 k(x,x).$$

$$= 2b\alpha^T c(x) + b^2 k(x,x) \geq 0.$$

But for any ξ_1 and $\xi_2 \in \mathbb{R}$, $\exists b$, $b\xi_1 + b^2 \cdot \xi_2 < 0$, so contradiction.

– Need to show all eval. fnc'als are Lipschitz.

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

Notice $k(x, \cdot) = \sum_{j=1}^{n} \theta_j k(x_j, \cdot) = 1 \cdot k(x, \cdot)$ $\begin{pmatrix} n=1 \\ \theta_1 = 1 \\ x_1 = x \end{pmatrix}$

$$\langle f, k(x, \cdot) \rangle = \sum_{ij} \alpha_i \theta_j k(x_i, x) = \sum_i \alpha_i k(x_i, x).$$

– $\mathcal{E}_x$ be an evaluation fnc'al $\forall f, g \in \mathcal{F}$

$$|\mathcal{E}_x(f-g)| = |\langle f-g, k(x, \cdot) \rangle|$$

$$\leq \|f-g\|_{\mathcal{F}} \|k(x, \cdot)\|_{\mathcal{F}} \quad (\text{by Cauchy-Schwartz})$$

$$= \|f-g\|_{\mathcal{F}} \sqrt{k(x,x)}$$

since $\|k(x, \cdot)\|_{\mathcal{F}}^2 = \langle k(x, \cdot), k(x, \cdot) \rangle = k(x,x).$

- To complete the proof, one needs to consider the completion of $\mathcal{F}$ including all limit points. This is skipped. $\boxtimes$

Remark: The main take away property of RKHS:

$f \in$ RKHS then $f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$ for some $x_i \in \mathcal{X}$ and $\alpha_i \in \mathbb{R}$.