CPSC 536F Dec 5, 2025

By the end of this class, you should know what is meant by:

- A **feature map**: $\Phi : \mathcal{X} \to H$

Hilbert space

$\mathbb{R}^d$, inner product: $\bullet$, d large

- **Ridge regression**:

$$\min_{w} \|Xw - y\|^2 + \lambda \|w\|_H^2$$

normal eqs: $X^*Xw + \lambda w = X^*y$

- **Kernel ridge regression**: (essentially) solving ridge regression via

① $XX^*\alpha + \lambda\alpha = y$

② $w = X^*\alpha$

Thm "Representer theorem" this works

- **Representer**: The $\vec{\alpha}$ above

- **The Gram kernel/matrix**: $K = XX^*$

above, so ① becomes

$$K\alpha + \lambda\alpha = y$$

Reproducing kernel Hilbert space

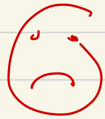
RKHS:

① $H_0 = \mathrm{Span}(\vec{\mathbb{1}})$ in $\mathbb{R}^3$

projection matrix $P = \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$

② $H_1 = H_0^{\perp}$ in $\mathbb{R}^3$

projection $Q = I - P = \begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix}$

Hilbert space that doesn't have a reproducing kernel is ... 

=

Any finite dimension inner product space, H_0 + product that looks like dot product

$v_1, v_2 \in H_0$: dot product

$$\langle v_1, v_2 \rangle_{H_0}$$

e.g. $H_0 = \mathbb{R}^d$, standard example

$$\langle \vec{u}, \vec{v} \rangle_{H_0} = u_1 v_1 + u_2 v_2 + u_3 v_3$$
$$= \vec{u} \cdot \vec{v}$$

Another $\langle \, , \, \rangle$

$$\langle \vec{u}, \vec{v} \rangle_{H_0} = 5 \vec{u} \cdot \vec{v}$$

$$+ (u_1 - u_2)(v_1 - v_2)$$

$$+ 3 u_3 \cdot v_3$$

<u>Def:</u> $\langle \vec{u}, \vec{v} \rangle_{H_0}$ is an inner product if

(1) $\langle \vec{u}, \vec{v} \rangle_{H_0}$ is bilinear

$$\langle \vec{u}, \vec{v} \rangle_{H_0} : \quad H_0 \times H_0 \to \mathbb{R}$$

$$\langle \vec{u}_1 + \vec{u}_2, \vec{v} \rangle = \langle \vec{u}_1, \vec{v} \rangle + \langle \vec{u}_2, \vec{v} \rangle$$

$$\langle \alpha \vec{u}, \vec{v} \rangle = \alpha \langle \vec{u}, \vec{v} \rangle$$

and

(2) (i) $\langle \vec{u}, \vec{u} \rangle_{H_0} \geq 0$,

equality iff $\vec{u} = 0$

(ii) $\langle \vec{u}, \vec{v} \rangle_{H_0} = \langle \vec{v}, \vec{u} \rangle_{H_0}$

(iii) $\| \vec{u} \|_2 \stackrel{\text{def}}{=} \sqrt{\langle \vec{u}, \vec{u} \rangle_{H_0}}$

$$\| \vec{u}_1 + \vec{u}_2 \| \leq \| \vec{u}_1 \| + \| \vec{u}_2 \|$$

E.g. $H_0 = C_0^\infty(\mathbb{R})$

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

is an inner product.

E.g. $H_0 = L^2[0,1]$,

$$\langle f, g \rangle = \int_0^1 f(x)\,g(x)\,dx$$

is an inner product, here f, g

$L^2[0,1] =$ "functions on $[0,1] \to \mathbb{R}$

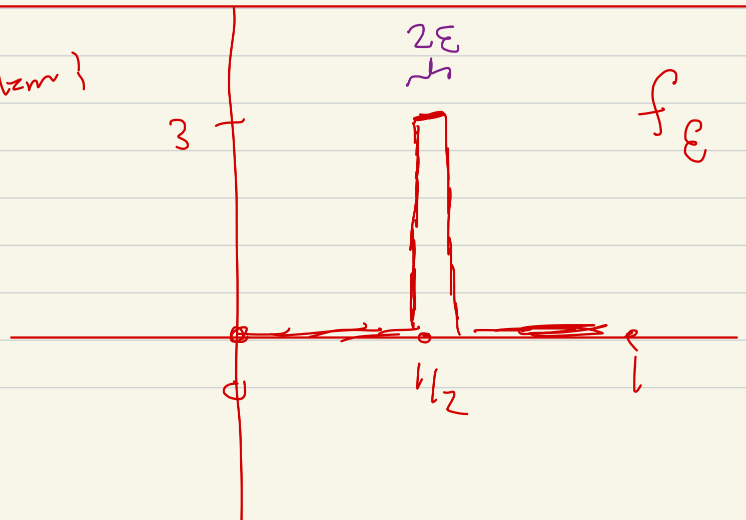
s.t. $\underbrace{\int_0^1 f^2(x)\,dx < \infty}_{\text{these days, we speak of measure theory}}$"

Does it make sense to speak of

$f(1/2)$? Data point $(1/2, 3)$

$$\min_{f \in L^2[0,1]} \left(f(1/2) - 3\right)^2 + \lambda \|f\|_{L^2}^2$$

Problem:



$$f_\varepsilon(x) = \begin{cases} 3 & \text{if } |x - \tfrac{1}{2}| \le \varepsilon \\ 0 & \text{elsewhere} \end{cases}$$

$$\int_0^1 |f_\varepsilon(x)|^2 \, dx = 9 \cdot 2\varepsilon$$

$$f_\varepsilon(1/2) = 3$$

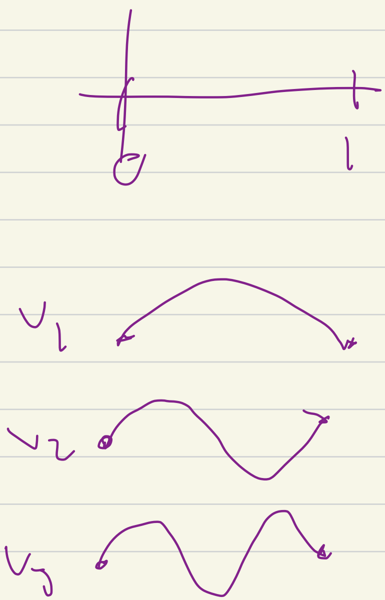
$$\min \underbrace{\left(f(1/2) - 3\right)^2}_{0} + \underbrace{\lambda}_{\lambda} \underbrace{\| f \|_{L^2}^2}_{9 \cdot 2\varepsilon}$$

$\varepsilon \to 0$, $\min \to 0$.

$$\min \left(\left(\int_{.45}^{.55} f(x)\, dx \right)^{-2.7} \frac{1}{.55-.45} - 3 \right)^2$$

$$+ \lambda \|f\|_{L^2}^2$$

$$\left[\left| \int_a^b f(x)\, dx \right|^2 \le \int_a^b f^2(x)\, dx \int_a^b 1\, dx \right]$$



ON basis:

$$v_n(x) = \sqrt{2} \sin(\pi n x)$$

for $L^2[0,1]$

$H_0 \subset L^2[0,1]$,

$H_0 = \mathrm{Span}(v_1, \dots, v_m)$

So

RKHS $\longmapsto$ what

are the questions,

the quantities to

minimize that make sense...

$$f(x) = \int f(y)\, \delta_x(y)\, dy$$

but $\delta_x(y) \notin L^2[0,1]$ or $L^2(\mathbb{R})$

In $L^2[0,1]$: $\langle f, g \rangle = \int_0^1 f(x) g(x)\, dx$

Take ON basis, e.g. $v_n(x) = \sqrt{2} \sin(\pi n x)$.

Let

$$H_0 = \operatorname{Span}\langle v_1(x), \dots, v_m(x) \rangle$$

Then projection

$$K(x,y) = \sum_{j=1}^{m} v_j(x) v_j(y)$$

here $\sum_{j=1}^{m} \sin(\pi j x) \sin(\pi j y)$

Rem: $\sum_{j=1}^{\infty} v_j(x) v_j(y)$ "$=$" $\delta_x(y)$

is problematic.

But ... $L^2[0,1]$ is not a RKHS,

since although

$$f(x) = \int \delta_x(y) f(y)\, dy$$

$\delta_x(y) \notin L^2[0,1]$.

So while any finite dimensional subspace of $\mathbb{R}^{\mathcal{X}}$ is a RKHS, $L^2[0,1]$ isn't;

also the representer theorem fails!

The real point here is not what is a RKHS, but what is not...

And that the representer theorem fails there... Indeed, $(f(x_i) - y_i)^2$ doesn't make sense, as $f(x_i)$ doesn't.

E.g. one data point: $(x_1, y_1) = (1/2, 3)$,

$$\min_f \left(\left(f(1/2) - 3 \right)^2 + \lambda \|f\|^2_{L^2} \right)$$

can be arbitrarily small.

Need to replace

$f(1/2)$ with continuous

$H \to \mathbb{R}$

Most of all: we should know <u>why</u> kernels and kernel tricks arise from feature maps:

$$\Phi : \mathcal{X} \to H \qquad \left(\begin{array}{l}\text{Hilbert space,} \\ \text{e.g. } \mathbb{R}^d, \ d \text{ large}\end{array}\right)$$

Data:

$$(x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathbb{R}$$

Model: Pick $w \in H$, or $v \in H$, or $f \in H$

$y = y(x)$ modeled by:

$$\Phi(x) \cdot w = y$$

S_6

$$\min \quad \lambda \|w\|_H^2 + \sum_{i=1}^n \Big(\underbrace{\Phi(x_i)\cdot w}_{} - y_i \Big)^2$$

$\Phi(x_i)\cdot w$

becomes

$X w$, where

$$X = \begin{bmatrix} -\Phi(x_1)^T- \\ \vdots \\ -\Phi(x_n)^T- \end{bmatrix}$$

$\langle \Phi(x_i), w \rangle_H$

when we want to emphasize Hilbert space

Then: back to variational argument, and write $K = XX^T$

Now we generalize: $\min_\omega$

$$G(\langle \Phi(x_1), \omega\rangle_H, \ldots, \langle \Phi(x_n), \omega\rangle_H)$$

$$+ \lambda\, m(\|\omega\|_H^2) \quad , \qquad G, m \text{ mild conditions}$$

Obscure or shorten

$$\min_{f \in H} G(f(x_1), \ldots f(x_n)) + \lambda \|f\|_H^2$$

$$f(x_i) \stackrel{\text{def}}{=} \langle \Phi(x_i), f\rangle_H$$

Use kernel language

Since $XX^T = \begin{bmatrix} \Phi(x_1)\cdot\Phi(x_1) - \\ \vdots \\ \\ \end{bmatrix} = K$

and introduce

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

as

$$k(x, x') = \Phi(x) \cdot \Phi(x')$$

or $\langle \Phi(x), \Phi(x') \rangle_H$

Also

$$w = X^* \vec{\alpha}$$

means

$$w = \begin{bmatrix} | & & | \\ \Phi(x_1) & \cdots & \Phi(x_n) \\ | & & | \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$$

$$= \sum_{i=1}^{n} \alpha_i \Phi(x_i)$$

and future predictions:

$$y(x) \overset{\text{predict}}{\approx} \Phi(x) \cdot w$$

become

$$y(x) = \Phi(x) \cdot \left(\sum_{i=1}^{n} \alpha_i \Phi(x_i) \right)$$

$$= \sum_{i=1}^{n} \alpha_i \Phi(x) \cdot \Phi(x_i)$$

$$= \sum_{i=1}^{n} \alpha_i K(x, x_i)$$