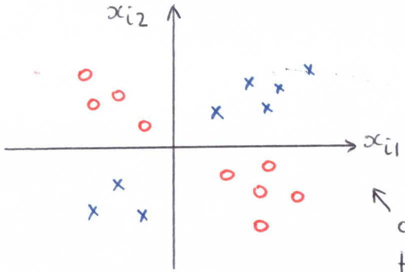
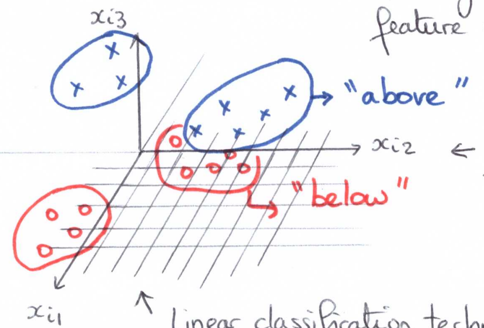
SL = REPRODUCING KERNEL HILBERT SPACES (RKHS)

- Motivating example: Binary Classification

Learning sample $\mathcal{L}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, $X_i \in \mathbb{R}^2$, $Y_i \in \{-1, 1\}$



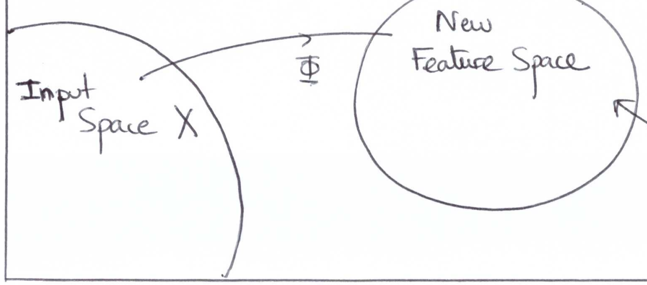
data is not linearly separable here. However, if we introduce a third variable $x_3 = x_1 x_2$, the points are linearly separable in the enlarged feature space.



New features = $\Phi(x) = (x_1 \;\; x_2 \;\; x_1 x_2)$

Linear classification techniques can be applied in the enlarged feature space.

- General idea:



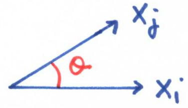
Perform linear Regression & Classification here.

A general strategy to predict the label of a new observation x is to choose a value y such that (x, y) is similar in some sense to some training examples.

↳ Similarity measure for input points $x_1, \dots, x_n$ is needed.
A natural candidate is the inner product:
Assuming that $\|x_i\| = 1$, $i = 1, \dots, n$, then

$$\langle x_i, x_j \rangle = x_i^t x_j = \cos\theta,$$

where
θ = angle between x_i and x_j.



$\langle x_i, x_j \rangle \geq 0 \iff x_i$ and x_j are pointing in the same direction.

In addition, vectors x_i & x_j such that $\langle x_i, x_j \rangle$ is close to 1 "look more alike" that vectors for which the inner product is small, or negative.

↳ The new feature space should support an inner product structure to allow us to evaluate similarity between new features; and then apply any machine learning algorithm there.

⇒ Natural candidates are Hilbert Spaces.
As we shall see, not all Hilbert Spaces are good candidates. We need nice ones, called Reproducing Kernel Hilbert Spaces.

- Outline.

↘ Elements of Functional Analysis (Hilbert Spaces / Operators)
↘ Reproducing Kernel Hilbert Spaces
↘ Constructing kernels
↘ Mercer Representation
↘ Applications in Machine Learning

I. ELEMENTS OF FUNCTIONAL ANALYSIS

I.1. HILBERT SPACES

Definition (Norm). Let $\mathcal{H}$ be a vector space over $\mathbb{R}$.
A function $\|.\|_{\mathcal{H}} : \mathcal{H} \to [0, \infty)$ is said to be a NORM on $\mathcal{H}$ if

(i) $\|f\|_{\mathcal{H}} = 0 \iff f = 0$ (norm separates points)

(ii) $\|\lambda f\|_{\mathcal{H}} = |\lambda| \, \|f\|_{\mathcal{H}} \quad \forall \lambda \in \mathbb{R} \quad \forall f \in \mathcal{H}$

(iii) $\|f + g\|_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} + \|g\|_{\mathcal{H}} \quad \forall f, g \in \mathcal{H}$ (triangle ineq.)

→ In every normed vector space, one can define a metric induced by the norm: $d(f, g) = \|f - g\|_{\mathcal{H}}$.

Ex: • $(\mathbb{R}, |.|)$ $(\mathbb{C}, |.|)$

• $(\mathbb{R}^d, \|x\|_p)$, where $\|x\|_p^p = \sum_{i=1}^d |x_i|^p$

As $p \to \infty$, $\|x\|_\infty = \max_i |x_i|$

• $(\mathcal{C}[a,b], \|f\|_p)$, where $\|f\|_p^p = \int_a^b |f(x)|^p dx$, $p \geq 1$

Definition (Inner product). Let $\mathcal{H}$ be a vector space over $\mathbb{R}$.
A function $\langle \cdot, \cdot \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said to be an INNER PRODUC on $\mathcal{H}$ if

(i) $\langle \lambda_1 f_1 + \lambda_2 f_2, g \rangle_{\mathcal{H}} = \lambda_1 \langle f_1, g \rangle + \lambda_2 \langle f_2, g \rangle \quad \forall \lambda_1, \lambda_2 \in \mathbb{R}$

(ii) $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$

(iii) $\langle f, f \rangle_{\mathcal{H}} \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0 \iff f = 0$.

→ In every inner product vector space, one can define a norm induced by the inner product $\|f\|_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}^{1/2}$.

Ex: • $\mathcal{H} = \mathbb{R}^d$, $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$

• $\mathcal{H} = \mathcal{C}[a,b]$, $\langle f, g \rangle = \int_a^b f(x) g(x) \, dx$.

Remark: • Inner product is needed to study useful geometrical notions analogous to those of Euclidean space $\mathbb{R}^d$. For example, the angle θ between $f, g \in \mathcal{H} \setminus \{0\}$ is given by

$$\cos \theta = \frac{\langle f, g \rangle_{\mathcal{H}}}{\|f\|_{\mathcal{H}} \|g\|_{\mathcal{H}}}.$$

[Not possible is $\mathcal{H}$ is only equiped with a norm]

• Key relations • $|\langle f, g \rangle| \leq \|f\| . \|g\|$ (CS ineq.)

• $4 \langle f, g \rangle = \|f + g\|^2 - \|f - g\|^2$ .../...

Definition (Convergent sequence) A sequence $\{f_n\}$ of elements of a normed space $(\mathcal{H}, \|.\|_{\mathcal{H}})$ is said to converge to $f \in \mathcal{H}$ if
$\forall \varepsilon > 0 \quad \exists N \in \mathbb{N} \quad \forall n \geq N \quad \|f_n - f\|_{\mathcal{H}} < \varepsilon$.

Definition (Cauchy sequence) A sequence $\{f_n\}_{n\geq 1}$ of elements of a normed vector space $(\mathcal{H}, \|.\|_{\mathcal{H}})$ is said to be a CAUCHY SEQUENCE if
$\forall \varepsilon > 0 \quad \exists N \in \mathbb{N} \quad \forall n, m \geq N \quad \|f_n - f_m\|_{\mathcal{H}} < \varepsilon$

↘ Since $\|f_n - f_m\|_{\mathcal{H}} \leq \|f_n - f\|_{\mathcal{H}} + \|f - f_m\|_{\mathcal{H}}$
A convergent sequence ⇒ it is a Cauchy sequence

↘ However, the converse is not true : Cauchy ⇏ convergent
Ex: sequence in $\mathbb{Q}$ converging to $\sqrt{2} \notin \mathbb{Q}$

→ A COMPLETE space $\mathcal{H}$ is such that every Cauchy sequence $\{f_n\}_{n\geq 1}$ in $\mathcal{H}$ converges: it has a limit, and this limit is in $\mathcal{H}$.

Definition (Hilbert space)
A HILBERT SPACE is a complete inner product space

Ex: • Space $L_2(X) = \{ f : X \to \mathbb{R} \mid \int_X |f(x)|^2 dx < \infty \}$ is a Hilbert space with inner product

$$\langle f, g \rangle = \int_X f(x) g(x) dx \qquad (\text{e.g. } X = \mathbb{R})$$

• Space $\ell^2(\mathbb{N})$ of sequences $\{x_n\}_{n \in \mathbb{N}}$ of real numbers satisfying $\sum |x_n|^2 < \infty$ is a Hilbert space, endowed with the inner product $\langle \{x_n\}, \{y_n\} \rangle_{\ell^2(\mathbb{N})} = \sum_{n \in \mathbb{N}} x_n y_n$

• Space $\mathbb{R}^3$ with $\langle x, y \rangle = x^t y$.

I.2. LINEAR OPERATORS.

Let $\mathcal{H}$ and $\mathcal{G}$ be two normed vector spaces over $\mathbb{R}$.

Definition (Linear Operator) A function $A : \mathcal{H} \to \mathcal{G}$ is said to be a LINEAR OPERATOR if

$$A(\lambda_1 f_1 + \lambda_2 f_2) = \lambda_1 A(f_1) + \lambda_2 A(f_2)$$

$\forall \lambda_1, \lambda_2 \in \mathbb{R} \quad \forall f_1, f_2 \in \mathcal{H}$.

Remark: Operators with $\mathcal{G} = \mathbb{R}$ are called FUNCTIONALS

Ex: Let $\mathcal{H}$ = inner product space and $g \in \mathcal{H}$.

• $A_g : \mathcal{H} \to \mathbb{R}$

$\quad f \mapsto A_g(f) = \langle f, g \rangle_{\mathcal{H}}$

A_g is a linear functional (obvious?)

Definition (Continuity) A function $A : \mathcal{H} \to \mathcal{G}$ is said to be continuous at $f_0 \in \mathcal{H}$ if

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \quad \|f - f_0\|_{\mathcal{H}} < \delta \Rightarrow \|Af - Af_0\|_{\mathcal{G}} < \varepsilon$$

A is continuous on $\mathcal{H}$ if it is continuous at every point of $\mathcal{H}$

Remark: Continuity means that a convergent sequence in $\mathcal{H}$ is mapped to a convergent sequence in $\mathcal{G}$.
A stronger form of continuity is that of LIPSCHITZ CONTINUITY:
$\exists C > 0 \quad \forall f_1, f_2 \in \mathcal{H} \quad \|Af_1 - Af_2\|_{\mathcal{G}} \leq C \|f_1 - f_2\|_{\mathcal{H}}$.

Ex: A_g previously defined is lipschitz continuous:

$A_g : \mathcal{H} \to \mathbb{R}$

$$|A_g(f_1) - A_g(f_2)| = |\langle f_1 - f_2, g \rangle_{\mathcal{H}}| \underset{\text{CS}}{\leq} \|g\|_{\mathcal{H}} \|f_1 - f_2\|_{\mathcal{H}}$$

Definition (Operator norm) The operator norm of a linear operator $A : \mathcal{H} \to \mathcal{G}$ is defined as

$$\|A\| = \sup_{\substack{f \in \mathcal{H} \\ \neq 0}} \frac{\|Af\|_{\mathcal{G}}}{\|f\|_{\mathcal{H}}}$$

If $\|A\| < \infty$, A is called a BOUNDED LINEAR OPERATOR

$\|A\|$ is the smallest number λ such that $\|Af\|_{\mathcal{G}} \leq \lambda \|f\|_{\mathcal{H}}$ holds $\forall f \in \mathcal{H}$.

Interpretation: A maps the closed unit ball in $\mathcal{H}$, into a subset of the closed ball in $\mathcal{G}$ centered at $0 \in \mathcal{G}$, with radius $\|A\|$.

Theorem. Let $(\mathcal{H}, \|.\|_{\mathcal{H}})$ and $(\mathcal{G}, \|.\|_{\mathcal{G}})$ be two normed linear spaces. If L is a linear operator, then the following three conditions are equivalent:

(i) L is a bounded operator

(ii) L is continuous on $\mathcal{H}$

(iii) L is continuous at one point of $\mathcal{H}$.

Proof: (i) $\Rightarrow$ (ii) Suppose $\exists \lambda < \infty$ s.t. $\forall f \in \mathcal{H}$,

$\| Lf \|_g \leqslant \lambda \| f \|_{\mathcal{H}}$.

Let $\varepsilon > 0$

Put $\delta = \varepsilon / \lambda$

let $f_0 \in \mathcal{H}$ such that $\| f - f_0 \|_{\mathcal{H}} < \varepsilon / \lambda$.

Then
$$\begin{aligned} \| Lf - Lf_0 \|_g &= \| L(f - f_0) \|_g \\ &\leqslant \lambda \| f - f_0 \|_{\mathcal{H}} \\ &< \lambda \, \varepsilon / \lambda \\ &= \varepsilon \end{aligned}$$
Boundedness

$\Rightarrow$ Continuity at f_0 (f_0 arbitrary)

(ii) $\Rightarrow$ (iii) Obvious

(iii) $\Rightarrow$ (i) Assume that L is continuous at one point $f_0 \in \mathcal{H}$.

Then $\exists \delta > 0 \quad \forall \| \Delta \|_{\mathcal{H}} \leqslant \delta \Rightarrow \| L\Delta \|_g = \| L(f_0 + \Delta) - Lf_0 \|_g \leqslant 1$

Now, $\forall f \in \mathcal{H}$, $f \neq 0$, $\left\| \dfrac{\delta}{2\|f\|_{\mathcal{H}}} f \right\|_{\mathcal{H}} < \delta$

Apply L

$$\left\| L\left(\frac{\delta f}{2\|f\|_{\mathcal{H}}} \right) \right\|_g \leqslant 1$$

Linearity of L

$$\| Lf \|_g \leqslant \frac{2}{\delta} \| f \|_{\mathcal{H}}$$

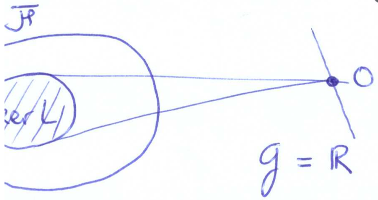
$\Rightarrow$ Boundedness

Remark: Closed versus Complete

→ $M \subseteq \mathcal{H}$ is CLOSED (in $\mathcal{H}$) if it contains limits of all sequences in M that converge in $\mathcal{H}$.

→ M is COMPLETE (with no reference to a larger space) if all Cauchy sequences in M converge in M.

→ The KERNEL of L is $\ker L = \{ f \in \mathcal{H} \mid Lf = 0 \}$

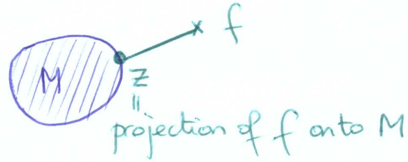


The kernel of a continuous functional necessarily is closed, as the preimage of the closed set $\{0\}$.
[See remark at the top of page 6]

Theorem (Projections in Hilbert Spaces)

Let $\mathcal{H}$ be a Hilbert Space and M a closed subspace.
Then $\forall f \in \mathcal{H}$, we have the decomposition $f = z + z^{\perp}$

$z \in M$, $z^{\perp} \in M^{\perp}$: $\forall m \in M \ \langle z^{\perp}, m \rangle = 0$



We have seen page 5 that $A_g := \langle \cdot, g \rangle_{\mathcal{H}}$ is a linear functional.

Theorem (RIESZ REPRESENTATION THEOREM)

In a Hilbert space $\mathcal{H}$, for every continuous linear functional $L : \mathcal{H} \to \mathbb{R}$, there exists a unique $g \in \mathcal{H}$, such that
$$Lf = \langle f, g \rangle_{\mathcal{H}}$$

Proof:
- If $L = 0$, then $g = 0$ will do
- If $L \neq 0$, then take $y \in (\ker L)^{\perp}$, $y \neq 0$

Since $L \neq 0$, we have $\ker L \subsetneq \mathcal{H}$ $\Rightarrow (\ker L)^{\perp} \neq \{0\}$

Rk: Since L is a continuous linear functional, $\ker L$ is closed & thus the theorem of projections in Hilbert Spaces ensure the existence of a $y \neq 0$ in $(\ker L)^{\perp} \neq \{0\}$

Put $z = \dfrac{y}{Ly}$ ← Note that $Ly \neq 0$ since $y \in (\ker L)^{\perp}$

We have that $Lz = 1$.

$\forall f \in \mathcal{F}$, $\quad f - zLf \in \ker L$ since

$$L(f - zLf) \underset{\substack{\uparrow \\ \text{linearity}}}{=} Lf - LzLf = 0$$

$$\Rightarrow \langle \underbrace{f - zLf}_{\in \ker L}, \underbrace{z}_{\in (\ker L)^\perp} \rangle_{\mathcal{F}} = 0 \Rightarrow Lf = \langle f, \frac{z}{\langle z, z\rangle} \rangle_{\mathcal{F}}$$

This is our g!

<u>Uniqueness</u>: Suppose $\exists f_1, f_2 \in \mathcal{F}$ s.t $\forall f \in \mathcal{F}$,

$\langle f, g_1 \rangle = \langle f, g_2 \rangle \quad \Rightarrow \quad \langle f, g_1 - g_2 \rangle = 0$

Take $f = g_1 - g_2$ and the result follows. ∎

Remark: Orthonormal basis

An orthonormal set $\{u_j\}$ is such that $\langle u_j, u_k \rangle_{\mathcal{F}} = \begin{cases} 1 & \text{if } j=k \\ 0 & \text{o/w} \end{cases}$

If it is also a basis, then denoting $\hat{f}_j = \langle f, u_j \rangle_{\mathcal{F}}$, we have

$$f = \sum \hat{f}_j u_j \Rightarrow \langle f, g \rangle_{\mathcal{F}} = \sum \hat{f}_j \hat{g}_j = \langle \{\hat{f}_j\}, \{\hat{g}_j\} \rangle_{\ell^2(\mathbb{N})}$$

<u>Definition</u>. Two Hilbert spaces $\mathcal{H}$ and $\mathcal{F}$ are said to be ISOMETRICALLY ISOMORPHIC if there is a linear bijective map $U : \mathcal{H} \to \mathcal{F}$ which preserves the inner product $\langle h_1, h_2 \rangle_{\mathcal{H}} = \langle Uh_1, Uh_2 \rangle_{\mathcal{F}}$

Although $\mathcal{H}$ and $\mathcal{F}$ may have elements of a different nature, (functions vs sequences), they still have the same geometric structure.

Theorem Every Hilbert space has an orthonormal basis. Thus, all Hilbert spaces are isometrically isomorphic to $\ell^2(\mathbb{N})$.

we need a separable Hilbert space [Contains a dense subset. Ex: $\mathbb{R}$

II. REPRODUCING KERNEL HILBERT SPACES (RKHS)

II.1. Evaluation Functional View of RKHS.

Let $X \subseteq \mathbb{R}^d$, and $\mathcal{H}$ = Hilbert Space of functions $X \to \mathbb{R}$.

For a fixed $x \in X$, the map $\delta_x : \mathcal{H} \to \mathbb{R}$, $f \mapsto f(x)$ is called the <u>evaluation functional</u> at x.

↳ Evaluation functionals are always linear since $\forall f, g \in \mathcal{H}$, $\forall \lambda, \nu \in \mathbb{R}$,

$$\delta_x(\lambda f + \nu g) = (\lambda f + \nu g)(x) = \lambda f(x) + \nu g(x) = \lambda \delta_x f + \nu \delta_x g.$$

↳ Evaluation functionals are not always continuous.

<u>Definition</u>:
A Hilbert Space $\mathcal{H}$ of functions $f : X \to \mathbb{R}$ defined on a non-empty set X is said to be a Reproducing Kernel Hilbert Space (RKHS) if the evaluation functional δ_x is continuous $\forall x \in X$.

↳ <u>Consequence</u>: if two functions converge in the RKHS norm, they converge pointwise at any point: $\forall x \in X$,

$$|f_n(x) - f(x)| = |\delta_x f_n - \delta_x f| = |\delta_x(f_n - f)| \le \|\delta_x\| \|f_n - f\|$$

The norm $\|\delta_x\|$ is bounded, since δ_x is a continuous linear operator on $\mathcal{H}$.

<u>Next</u>: We discuss 3 distinct topics: Reproducing Kernel, Kernel, Positive Definite Function, and then show that they are equivalent

I.2. KEPRODUCING KERNEL

Definition. (Reproducing kernel)

Let $\mathcal{H}$ be a Hilbert Space of functions $f: X \to \mathbb{R}$ defined on a non-empty set X.

A function $K: X \times X \to \mathbb{R}$ is called a REPRODUCING KERNEL of $\mathcal{H}$ if it satisfies

(i) $\forall x \in X \quad K_x = K(\cdot, x) \in \mathcal{H}$

(ii) $\forall x \in X \quad \forall f \in \mathcal{H} \quad \langle f, K(\cdot, x) \rangle_{\mathcal{H}} = f(x)$

Quite restrictive: does such a function exist at all?

The reproducing property.

In particular, $\forall x, y \in X$, $K_y = K(\cdot, y) \in \mathcal{H} \Rightarrow$

$$K(x,y) = \langle K(\cdot, y), K(\cdot, x) \rangle_{\mathcal{H}}$$
$$= \langle K(\cdot, x), K(\cdot, y) \rangle_{\mathcal{H}}$$

Remark: If it exists, the reproducing kernel is unique. Indeed, assume that $\mathcal{H}$ has two reproducing kernels K_1 and K_2. Then
$\langle f, K_1(\cdot, x) - K_2(\cdot, x) \rangle_{\mathcal{H}} = f(x) - f(x) = 0 \quad \forall f \in \mathcal{H} \quad \forall x \in X.$
In particular, taking $f(\cdot) = K_1(\cdot, x) - K_2(\cdot, x)$ gives uniqueness.

What about existence?

abstract definition

less abstract: we start characterizing the elements of an RKHS

Theorem: $\mathcal{H}$ is a RKHS $\Leftrightarrow$ $\mathcal{H}$ has a reproducing kernel.

Proof $\boxed{\Leftarrow}$ Suppose that $\mathcal{H}$ has a reproducing kernel. Then

$$|\delta_x f| = |f(x)| = |\langle f, K(\cdot, x) \rangle_{\mathcal{H}}|$$
$$\leq \| K(\cdot, x) \|_{\mathcal{H}} \| f \|_{\mathcal{H}}$$
$$= \langle K(\cdot, x), K(\cdot, x) \rangle_{\mathcal{H}}^{1/2} \| f \|_{\mathcal{H}}$$
$$= K(x, x)^{1/2} \| f \|_{\mathcal{H}}$$

$\Rightarrow \delta_x : \mathcal{H} \to \mathbb{R}$ is a bounded linear operator, hence a continuous one.

$\boxed{\Rightarrow}$ Suppose that $\delta_x : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional.
The Riesz representation theorem ensures the existence of an element $f_{\delta_x} \in \mathcal{H}$ such that

$$\delta_x f = \langle f, f_{\delta_x} \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}.$$

Define $K(y, x) = f_{\delta_x}(y) \quad \forall x, y \in X$.

Then $\to K(\cdot, x) = f_{\delta_x} \in \mathcal{H}$

$\to \langle f, K(\cdot, x) \rangle_{\mathcal{H}} = \delta_x f = f(x)$

Thus K is the reproducing kernel ∎

I.3. INNER PRODUCT BETWEEN FEATURES.

Definition (Kernel) A function $K: X \times X \to \mathbb{R}$ is a KERNEL on X if (i) $\exists$ a Hilbert Space $\mathcal{H}$

(ii) A mapping $\Phi: X \to \mathcal{H}$

such that $\forall x, y \in X \quad K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$

We drop the reproducing property, as $\mathcal{H}$ may not be an RKHS (not even necessarily a function space)

Φ known as a FEATURE MAP

$\mathcal{H}$ known as a FEATURE SPACE

Corollary: Every Reproducing Kernel is a Kernel.

here, a function space

Take $\Phi: y \mapsto K(\cdot, y) \in \mathcal{H}$

Reproducing kernel

Then $\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} = \langle K(\cdot, x), K(\cdot, y) \rangle_{\mathcal{H}} = K(x, y)$

× Examples

- $X = \mathbb{R}^2$

$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$

$$
\begin{aligned}
K(x,y) &= \langle x, y \rangle^2 \\
&= (x_1 y_1)^2 + (x_2 y_2)^2 + 2 x_1 x_2 y_1 y_2 \\
&= (x_1^2 \;\; x_2^2 \;\; \sqrt{2} x_1 x_2)(y_1^2 \;\; y_2^2 \;\; \sqrt{2} y_1 y_2)^t \\
&= \Phi(x)^t \, \Phi(y),
\end{aligned}
$$

where we defined $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$, $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \end{pmatrix}$

monomials of order 2

Take $\mathcal{H} = \mathbb{R}^3$ = Feature space.

Note that the feature map & the feature space are not unique, since we can as well define

$K(x,y) = \bar{\Phi}(x)\bar{\Phi}(y)$, with $\bar{\Phi} : \mathbb{R}^2 \to \mathbb{R}^4$, $\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \\ x_1 x_2 \end{pmatrix}$

$\bar{\mathcal{H}} = \mathbb{R}^4$

$\mathcal{H}$ & $\bar{\mathcal{H}}$ are not RKHS (they are not spaces of functions)

- $X = \mathbb{R}^d$

$x = (x_1 \; \dots \; x_d)^t$, $y = (y_1 \; \dots \; y_d)^t$

$$
\begin{aligned}
K(x,y) &= \langle x, y \rangle^m \\
&= \left(\sum x_i y_i \right)^m \\
&= \sum_{j_1 + \dots + j_d = m} \frac{m!}{j_1! \dots j_d!} (x_1 y_1)^{j_1} \times \dots \times (x_d y_d)^{j_d}.
\end{aligned}
$$

$$
K(x,y) = \sum_{j_1 + \dots + j_d = m} \underbrace{\sqrt{\frac{m!}{j_1! \dots j_d!}} \, x_1^{j_1} \dots x_d^{j_d}}_{= \phi_j(x)} \; \underbrace{\sqrt{\frac{m!}{j_1! \dots j_d!}} \, y_1^{j_1} \dots y_d^{j_d}}_{= \phi_j(y)}
$$

$j = (j_1, \dots, j_d)$

$$
\begin{aligned}
&= \sum_{j_1 + \dots + j_d = m} \phi_j(x) \, \phi_j(y) \\
&= (\phi_{m,0,\dots,0}(x), \; \phi_{0,m,0,\dots,0}(x), \; \dots) \cdot (\phi_{m,0,\dots,0}(y), \; \phi_{0,m,0,\dots,0}(y), \dots)^t
\end{aligned}
$$

$=: \Phi(x)$ $\quad \Phi(y)^t$

$\Rightarrow$ We extracted a feature map $\Phi : \mathbb{R}^d \to \mathbb{R}^{\binom{d+m-1}{m}}$ and a feature space $\mathcal{H} = \mathbb{R}^{\binom{d+m-1}{m}}$, so K is a kernel indeed.

Note that elements of Φ contains all monomials of order m.

- $X = \mathbb{R}^2$

$$
\begin{aligned}
K(x,y) &= (1 + \langle x, y \rangle)^2 \\
&= (1 + x_1 y_1 + x_2 y_2)^2 \\
&= 1 + (x_1 y_1)^2 + (x_2 y_2)^2 + 2 x_1 y_1 + 2 x_2 y_2 + 2 x_1 x_2 y_1 y_2 \\
&= (1 \;\; \sqrt{2} x_1 \;\; \sqrt{2} x_2 \;\; x_1^2 \;\; x_2^2 \;\; \sqrt{2} x_1 x_2) \cdot (1 \;\; \sqrt{2} y_1 \;\; \sqrt{2} y_2 \;\; y_1^2 \;\; y_2^2 \;\; \sqrt{2} y_1 y_2)^t
\end{aligned}
$$

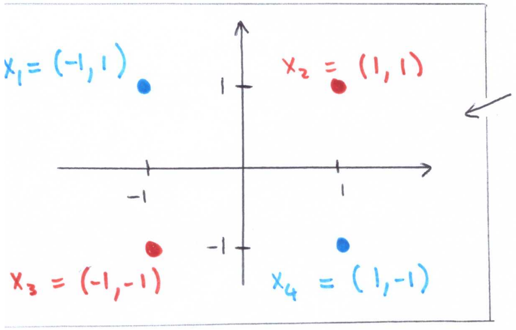
$\Phi(x)$

constant — original features — 2nd order polynomial — product

The feature space is $\mathcal{H} = \mathbb{R}^6$.
Elements of Φ are all monomials up to & including order 2.

- Generalizing the previous example, we see that the kernel $K(x,y) = (1 + \langle x, y \rangle)^m$, $x, y \in \mathbb{R}^d$, is associated with a feature map containing all monomials of order $\leq m$.
  ↪ Φ is finite dimensional.
  aka "polynomial" kernel

× Application: Binary classification.
Consider a sample of size 4 : $\mathcal{L}_4 = \{(x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4)\}$
where $x_i \in \mathbb{R}^2$ $(=X)$, and $y_i \in \{-1, 1\}$, and such that



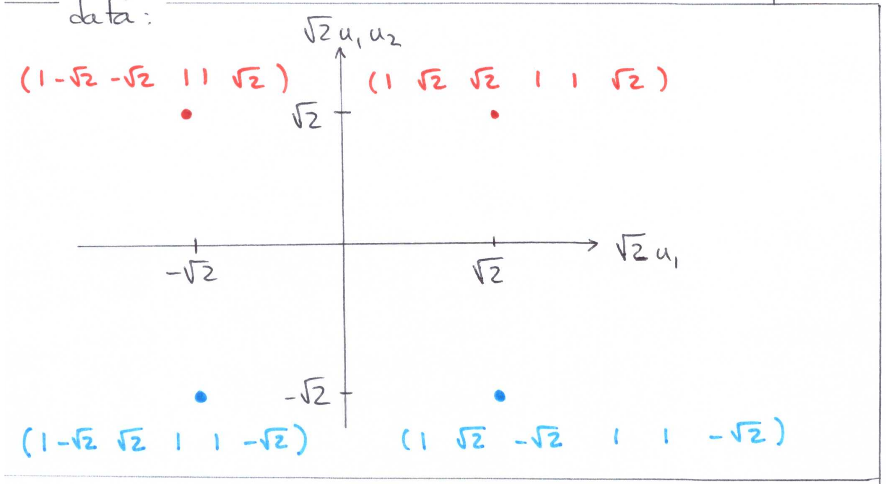
Training sample is not linearly separable in the original feature space X.

Consider the kernel
$K(x,y) = (1 + \langle x, y \rangle)^2$.

The data is projected onto a larger space via a feature map. Take $\Phi(u) = (1 \;\; \sqrt{2}u_1 \;\; \sqrt{2}u_2 \;\; u_1^2 \;\; u_2^2 \;\; \sqrt{2}u_1u_2)$ and $\mathcal{H} = \mathbb{R}^6$. Then

$$
\begin{aligned}
x_1 = (-1,1) &\hookrightarrow (1 \;\; -\sqrt{2} \;\; \sqrt{2} \;\; 1 \;\; 1 \;\; -\sqrt{2})\\
x_2 = (1,1) &\hookrightarrow (1 \;\; \sqrt{2} \;\; \sqrt{2} \;\; 1 \;\; 1 \;\; \sqrt{2})\\
x_3 = (-1,-1) &\hookrightarrow (1 \;\; -\sqrt{2} \;\; -\sqrt{2} \;\; 1 \;\; 1 \;\; \sqrt{2})\\
x_4 = (1,-1) &\hookrightarrow (1 \;\; \sqrt{2} \;\; -\sqrt{2} \;\; 1 \;\; 1 \;\; -\sqrt{2})
\end{aligned}
$$

Plot the second and last component of the transformed data:



⇒ The transformed data is linearly separable in the enlarged feature space.
Note that the features are "automatically" generated, and not manually selected. They appear implicitely as soon as a kernel function is selected → It will be useful later in a machine learning context.

- $X = \mathbb{R}^d$
  Let $\sigma^2 > 0$ and put $K(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$.
  "Gaussian kernel"

We show that K is indeed a kernel, by extracting a feature map, and a feature space.
$\|x-y\|^2 = \|x\|^2 + \|y\|^2 - 2\langle x, y \rangle$
$\Rightarrow K(x,y) = e^{-\|x\|^2/2\sigma^2} \, e^{-\|y\|^2/2\sigma^2} \, e^{\langle x, y \rangle/\sigma^2}$

$$K(x,y) = e^{-\|x\|^2/2\sigma^2} e^{-\|y\|^2/2\sigma^2} \sum_{k\geq 0} \frac{1}{\sigma^{2k} k!} \langle x, y\rangle^k$$

with
$$\langle x,y\rangle^k = \sum_{j_1+\cdots+j_d=k} \frac{k!}{j_1!\cdots j_d!} (x_1 y_1)^{j_1} \times\cdots\times (x_d y_d)^{j_d}$$

$$= e^{-\|x\|^2/2\sigma^2} e^{-\|y\|^2/2\sigma^2} \sum_{j_1+\cdots+j_d\geq 0} \frac{1}{\sigma^{2(j_1+\cdots+j_d)} j_1! \times\cdots\times j_d!}$$

$$\times\, x_1^{j_1}\cdots x_d^{j_d}\, y_1^{j_1}\cdots y_d^{j_d}$$

$\Rightarrow$ Put $\phi_j(x) = \sqrt{\dfrac{1}{\sigma^{2(j_1+\cdots+j_d)} j_1!\cdots j_d!}}\; x_1^{j_1}\cdots x_d^{j_d}$

$j=(j_1,\ldots,j_d)$; positive entries.

We have the representation $K(x,y) = \Phi(x)^t\,\Phi(y)$, with

$$\Phi(x) := e^{-\|x\|^2/2\sigma^2}\left(\phi_{0,\ldots,0}(x),\ \phi_{1,0,\ldots,0}(x),\ \ldots\right)$$

Infinite dimensional

$\Rightarrow$ The feature space is infinite dimensional as well.

(compare with polynomial kernels)

• Remark:

$$\underset{\text{kernel}}{K(x,y)} = \frac{K'(x,y)}{\sqrt{K'(x,x)\,K'(y,y)}}, \text{ with } K'(x,y) := e^{\frac{\langle x,y\rangle}{\sigma^2}}$$

kernel — also a kernel

II.4. Positive definite functions.

A symmetric function $h: X\times X\to\mathbb{R}$ is positive definite if for any $n\geq 1$, for any $\lambda_1,\ldots,\lambda_n\in\mathbb{R}$ & for any $x_1,\ldots,x_n\in X$, holds

$$\boxed{\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j\, h(x_i,x_j) \geq 0}$$

In other words, the GRAM MATRIX $H := [h(x_i,x_j)]_{i,j=1,\ldots,n}$ is symmetric positive semi-definite: $\lambda^t H\lambda \geq 0$; $\lambda = (\lambda_1,\ldots,\lambda_n)\in\mathbb{R}^n$.

× Consequence: A Kernel is a positive definite function.

Indeed, consider a kernel $K: X\times X\to\mathbb{R}$, associated with a Hilbert Space $\mathcal{H}$, and the feature map $\Phi: X\to\mathcal{H}$.

Then
$$\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j K(x_i,x_j) \underset{\text{definition of a kernel}}{=} \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j\langle\Phi(x_i),\Phi(x_j)\rangle_{\mathcal{H}}$$

$$\underset{\text{bilinearity}}{=} \left\langle \sum_{i=1}^{n}\lambda_i\Phi(x_i),\ \sum_{j=1}^{n}\lambda_j\Phi(x_j)\right\rangle_{\mathcal{H}}$$

$$= \left\|\sum_{i=1}^{n}\lambda_i\Phi(x_i)\right\|_{\mathcal{H}}^2 \geq 0$$

× Summary:

$\mathcal{H}$ is an RKHS $\Leftrightarrow$ $\mathcal{H}$ has a unique reproducing kernel

& Reproducing kernel $\Rightarrow$ Kernel $\Rightarrow$ Positive Definite function

$\Rightarrow$ Given an RKHS $\mathcal{H}$, $\mathcal{H}$ defines a unique reproducing kernel, which is a positive definite function. (18)

It turns out that the converse is also true, as the next theorem shows:

> **Theorem (Moore – Aronszajn)**
>
> Let $K: X \times X \to \mathbb{R}$ be a positive definite function.
>
> Then there exists a unique RKHS $\mathcal{H}$ (space of functions $X \to \mathbb{R}$) with reproducing kernel K.

The RKHS & the reproducing kernel are unique. The feature map is not.

To prove the Moore – Aronszajn theorem, we proceed in several steps:

(i) First, we construct a pre-Hilbert space $\mathcal{H}_0$.
$\mathcal{H}_0$ is not a Hilbert space, but is equiped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$.
$\hookrightarrow$ The structure of $\mathcal{H}_0$ informs us about what the elements in the final RKHS look like.

(ii) Add limit points: completion of $\mathcal{H}_0 \to \mathcal{H}$
Define a new object $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, constructed from $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$, and show that it is an inner product on $\mathcal{H}$.

(iii) Show that $\mathcal{H}$ is complete + evaluation functionals are continuous on $\mathcal{H}$.

Step (a) Let K = Positive Definite, Symmetric function on $X \times X$

Put $\mathcal{H}_0 := \left\{ f: X \to \mathbb{R} \mid f(\cdot) = \sum_{i=1}^{n} \lambda_i K(\cdot, x_i), \; n \geq 1, \; \lambda_i \in \mathbb{R}, \; x_i \in X \right\}$

Let $f, g \in \mathcal{H}_0$; $f(\cdot) = \sum_i \lambda_i K(\cdot, x_i)$ (19)
$g(\cdot) = \sum_j \gamma_j K(\cdot, y_j)$,

for some $\lambda_i, x_i, \gamma_j, y_j$.

Put $\boxed{\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i,j} \lambda_i \gamma_j K(x_i, y_j)}$

Q: Does $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ defines an inner product on $\mathcal{H}_0$?

First of all, note that the definition of $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ does not depend on the representation of f and g. Indeed, consider

$$f(\cdot) = \sum_i \lambda_i K(\cdot, x_i) = \sum_i \lambda_i' K(\cdot, x_i')$$
$$g(\cdot) = \sum_j \gamma_j K(\cdot, y_j) = \sum_j \gamma_j' K(\cdot, y_j').$$

Then

$$\begin{aligned}
\langle f, g \rangle_{\mathcal{H}_0} &= \sum_{i,j} \lambda_i \gamma_j K(x_i, y_j) \\
&= \sum_i \lambda_i \underbrace{\sum_j \gamma_j K(x_i, y_j)}_{= g(x_i)} \\
&= \sum_i \lambda_i g(x_i) \\
&= \sum_i \lambda_i \sum_j \gamma_j' K(x_i, y_j') \qquad \text{second representation of } g \\
&= \sum_j \gamma_j' \underbrace{\sum_i \lambda_i K(x_i, y_j')}_{= f(y_j')} \\
&= \sum_j \gamma_j' f(y_j') \\
&= \sum_j \gamma_j' \sum_i \lambda_i' K(x_i', y_j') \\
&= \sum_{i,j} \lambda_i' \gamma_j' K(x_i', y_j'), \text{ as required. } \blacksquare
\end{aligned}$$

→ Symmetry of $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ follows from the symmetry of K.

→ Bilinearity is a direct consequence of the definition of $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$.

→ $\forall f \in \mathcal{H}_0$, $\langle f, f \rangle_{\mathcal{H}_0} = \sum_{i,j} \lambda_i \lambda_j K(x_i, x_j) \geq 0$,

since K is a kernel, hence positive definite.

Note in addition that $\forall n \geq 1$, $\forall \lambda_1, .., \lambda_n$,

$\forall f_1, .., f_n \in \mathcal{H}_0$, $\lambda^t F \lambda = \langle \underbrace{\sum_i \lambda_i f_i}_{\in \mathcal{H}_0}, \underbrace{\sum_j \lambda_j f_j}_{\in \mathcal{H}_0} \rangle \geq 0$

$\lambda := (\lambda_1, .., \lambda_n)$

$F := [\langle f_i, f_j \rangle_{\mathcal{H}_0}]$

= Gram matrix

$\Rightarrow$ Matrix F is positive semi definite.

• Consequence: K satisfies the Cauchy-Schwartz ineq:

Indeed, taking $n = 2$, the determinant of F must be non negative:

$$\left| \begin{pmatrix} \langle f_1, f_1 \rangle & \langle f_1, f_2 \rangle \\ \langle f_2, f_1 \rangle & \langle f_2, f_2 \rangle \end{pmatrix} \right| \geq 0$$

$$\langle f_1, f_1 \rangle \langle f_2, f_2 \rangle - \langle f_1, f_2 \rangle^2 \geq 0$$

$$\Rightarrow \boxed{\langle f_1, f_2 \rangle^2 \leq \langle f_1, f_1 \rangle \langle f_2, f_2 \rangle} \quad (*)$$

& similarly for K using $\begin{pmatrix} K(x,x) & K(x,y) \\ K(y,x) & K(x,x) \end{pmatrix}$,

we get $K(x,y)^2 \leq K(x,x) K(y,y)$

→ It remains to show that $\langle f, f \rangle_{\mathcal{H}_0} = 0 \Rightarrow f = 0$

Let $f \in \mathcal{H}_0$.

Then $f(x) = \sum_i \lambda_i K(x, x_i)$ for some λ_i, x_i

$$\Rightarrow \boxed{f(x)} \underset{\substack{\uparrow \\ \text{by definition of } \langle \cdot, \cdot \rangle_{\mathcal{H}_0}}}{=} \underline{\langle f, K(\cdot, x) \rangle_{\mathcal{H}_0}}$$

The reproducibility property!

(& in particular, taking $f(x) = K(x,y)$,

$f(x) = K(x, y) \underset{\substack{\uparrow \\ \text{reproducible kernel}}}{=} \langle K(\cdot, y), K(\cdot, x) \rangle$)

Next,

$|f(x)|^2 = | \langle f, K(\cdot, x) \rangle_{\mathcal{H}_0} |^2$

$\leq \langle f, f \rangle_{\mathcal{H}_0} \langle K(\cdot, x), K(\cdot, x) \rangle_{\mathcal{H}_0}$

from (*) page 20 $= K(x, x) \langle f, f \rangle_{\mathcal{H}_0}$

And indeed, $\langle f, f \rangle_{\mathcal{H}_0} = 0 \Rightarrow f(x) = 0 \ \forall x$.

Summary: Given K = symmetric, positive definite function, we may define the space of functions

$\mathcal{H}_0 := \{ f : X \to \mathbb{R} \mid f(\cdot) = \sum_{i=1}^n \lambda_i K(\cdot, x_i) \}$,

endowed with the inner product

$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i,j} \lambda_i \gamma_j K(x_i, y_j)$, $\quad f(\cdot) = \sum \lambda_i K(\cdot, x_i)$, $g(\cdot) = \sum \gamma_j K(\cdot, y_j)$

• **Remark**: Evaluation functionals are continuous on $\mathcal{H}_0$.

Take $f, g \in \mathcal{H}_0$.

Then $\forall x \in X$,
$$f(x) = \langle f, K(\cdot, x)\rangle = \delta_x f$$
$$g(x) = \langle g, K(\cdot, x)\rangle = \delta_x g.$$

$$\Rightarrow |\delta_x f - \delta_x g| = |\langle f-g, K(\cdot,x)\rangle_{\mathcal{H}_0}|$$
$$\leq \sqrt{K(x,x)}\, \|f-g\|_{\mathcal{H}_0} \quad (*)$$

$\Rightarrow$ The functional δ_x is continuous on $\mathcal{H}_0$ for all $x \in X$. ∎

↳ **Next**: Complete $\mathcal{H}_0$ with all its limit points:

Let $\{f_n\}$ be a Cauchy sequence in $(\mathcal{H}_0, \langle\cdot,\cdot\rangle_{\mathcal{H}_0})$.

Since
$$|f_n(x) - f_m(x)| \leq \underbrace{\|f_n - f_m\|_{\mathcal{H}_0}}_{\downarrow\, 0} \sqrt{K(x,x)},$$

we conclude that the sequence $\{f_n(x)\}$ is a Cauchy sequence in $\mathbb{R}$, and therefore that it converges.

$\Rightarrow$ Add in $\mathcal{H}_0$ the functions that are pointwise limits of Cauchy Sequences in $\mathcal{H}_0$.

— Call this enlarged space $\mathcal{H}$ —

It "remains" to define an inner product on $\mathcal{H}$, show that $\mathcal{H}$ is complete for this inner product, and that the evaluation functional is continuous on $\mathcal{H}$. Before we complete the proof, we illustrate what elements of $\mathcal{H}_0$ look like for some specific choices of K.

• **Examples / Digression**

(i) • $X = \mathbb{R}^d$

$$K(x,y) = \langle x, y\rangle = x_1 y_1 + \dots + x_d y_d$$
$$= (x_1 \dots x_d)(y_1 \dots y_d)^t$$
$$= \Phi(x)\, \Phi(y)$$

$$\begin{cases} \Phi(x) = (x_1, \dots, x_d) = \text{feature map} = \text{Identity} \\ \mathcal{H} = \mathbb{R}^d = \text{feature space} \end{cases}$$

↑ not the RKHS, nor the pre-RKHS.

The pre-Hilbert space $\mathcal{H}_0$ associated with K has elements of the form $f(\cdot) = \sum_{i=1}^{n} \lambda_i K(\cdot, x_i)$,

for $n \geq 1$, & some $\lambda_1, \dots, \lambda_n \in \mathbb{R}$
$x_1, \dots, x_n \in \mathbb{R}^d$

Thus
$$f(x) = \lambda_1 K(x, x_1) + \dots + \lambda_n K(x, x_n)$$
$$= \lambda_1 x^t x_1 + \dots + \lambda_n x^t x_n$$
$$= x^t \underbrace{(\lambda_1 x_1 + \dots + \lambda_n x_n)}_{\in \mathbb{R}^d}$$

Denote this vector γ

$$= x^t \gamma$$
$$= \underline{\text{linear function of } x}.$$

The pre-Hilbert space $\mathcal{H}_0$ contains linear functions of x (and, of course, the after completion, the RKHS contains as well linear functions of x).

$\Rightarrow$ There is no substantial gain in using the kernel $K(x,y) = \langle x, y\rangle$.

(ii) • $X = \mathbb{R}^2$

$$K(x, y) = \langle x, y \rangle^2 = (x_1 y_1)^2 + (x_2 y_2)^2 + 2\, x_1 x_2 y_1 y_2 .$$

Let $f \in \mathcal{H}_0$. Then

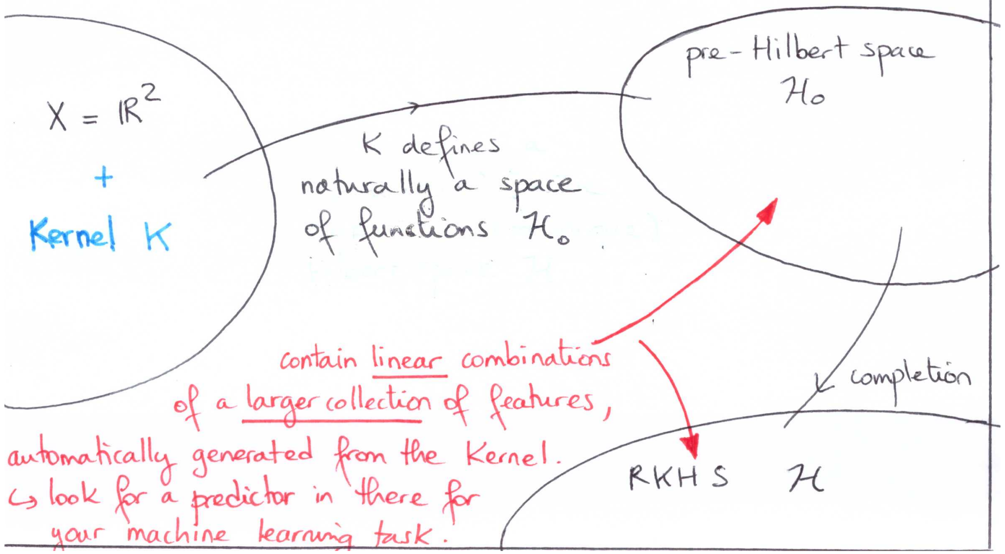
$$f(\cdot) = \sum_{i=1}^{n} \lambda_i K(\cdot, x_i) \qquad \begin{array}{l} n \geq 1 \\ \lambda_1, \dots, \lambda_n \in \mathbb{R} \\ x_1, \dots, x_n \in X = \mathbb{R}^2 \end{array}$$

$$\begin{aligned} f(x) &= \lambda_1 \langle x, x_1 \rangle^2 + \dots + \lambda_n \langle x, x_n \rangle^2 \\ &= x_1^2 \left[\lambda_1 x_{11}^2 + \dots + \lambda_n x_{n1}^2 \right] \\ &\quad + x_2^2 \left[\lambda_1 x_{12}^2 + \dots + \lambda_n x_{n2}^2 \right] \\ &\quad + 2 x_1 x_2 \left[\lambda_1 x_{11} x_{12} + \dots + \lambda_n x_{n1} x_{n2} \right] \end{aligned}$$

$$f(x) = a\, x_1^2 + b\, x_2^2 + c\, x_1 x_2$$

= linear combination of monomials of order 2.

(= elements of the feature map $\Phi(x) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \end{pmatrix}$)



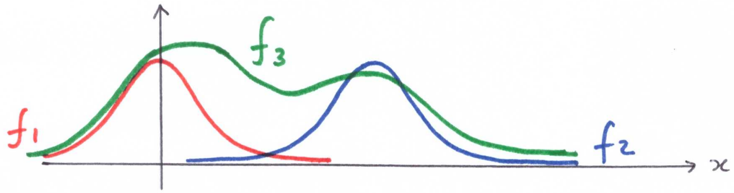
(iii) • $X = \mathbb{R}$

• $K(x, y) = \exp\left(-\frac{1}{2\sigma^2} (x - y)^2 \right)$ = Gaussian kernel

↳ $f_1(x) := K(x, 0) = e^{-x^2/2\sigma^2} \in \mathcal{H}$

$f_2(x) := K(x, 1) = e^{-(x-1)^2/2\sigma^2} \in \mathcal{H}$

$f_3(x) := K(x, 0) + \frac{1}{2} K(x, 1) \in \mathcal{H}$



■

<u>Step (b)</u> = Completion

↳ See Appendix

III. CONSTRUCTING KERNELS

Standard operations can be used to define new kernels.
Given kernels K_1 and K_2, the following functions K are also kernels:

(i) $K(x,y) = c\,K_1(x,y)$, $c>0$

Since K_1 is a kernel, there exists a feature map Φ and a Hilbert Space $\mathcal{H}$ s.t. $K_1(x,y) = \langle \Phi(x), \Phi(y)\rangle_{\mathcal{H}}$.

$\Rightarrow c\,K_1(x,y) = \langle \sqrt{c}\,\Phi(x), \sqrt{c}\,\Phi(y)\rangle_{\mathcal{H}}$

$\Rightarrow c\,K_1$ is also a kernel.

(ii) $K(x,y) = f(x)\,K_1(x,y)\,f(y)$, f = any function.

Similarly, $f(x)\,K_1(x,y)\,f(y) = \langle f(x)\Phi(x), f(y)\Phi(y)\rangle_{\mathcal{H}}$

(iii) $K(x,y) = K_1(x,y) + K_2(x,y)$

Consider the Gram matrices $\underline{\underline{K_1}}$ & $\underline{\underline{K_2}}$ associated with the kernels K_1 & K_2.
$\underline{\underline{K_1}}$ & $\underline{\underline{K_2}}$ are positive semi-definite $\Rightarrow \lambda^t \underline{\underline{K_1}}\lambda \geq 0$, $\lambda^t \underline{\underline{K_2}}\lambda \geq 0$

Thus $\lambda^t \underline{\underline{K}}\lambda = \lambda^t \underline{\underline{K_1}}\lambda + \lambda^t \underline{\underline{K_2}}\lambda \geq 0$

Gram matrix associated with K.

(iv) $K(x,y) = K_1(x,y)\,K_2(x,y)$ [HADAMARD product]

We consider the Gram Matrices of K, K_1, K_2, denoted $\underline{\underline{K}} = (K_{ij})$, $\underline{\underline{K_1}} = (K_{ij}^{(1)})$, $\underline{\underline{K_2}} = (K_{ij}^{(2)})$, respectively.
Then $K_{ij} = K_{ij}^{(1)} K_{ij}^{(2)}$.
We show that $\underline{\underline{K}}$ is positive semi-definite. To do so, we show that $\underline{\underline{K}}$ is the covariance matrix of some random vector.

Let $U \sim \mathcal{N}(0, \underline{\underline{K_1}})$, $U = (U_1,\dots,U_n)^t$
$V \sim \mathcal{N}(0, \underline{\underline{K_2}})$, $V = (V_1,\dots,V_n)^t$ } indpt

and put $W := (U_1V_1,\dots,U_nV_n) = (W_1,\dots,W_n)$

Then • $\mathbb{E}W = 0$ since $\mathbb{E}(U_iV_i) \underset{\text{independence}}{=} \mathbb{E}U_i\,\mathbb{E}V_i = 0$

• $\underline{\underline{\Sigma_W}} = \mathbb{E}(WW^t)$,

where
$$(\underline{\underline{\Sigma_W}})_{ij} = \mathbb{E}(W_iW_j) = \mathbb{E}(U_iV_iU_jV_j) = \mathbb{E}(U_iU_j)\,\mathbb{E}(V_iV_j) = K_{ij}^{(1)}K_{ij}^{(2)}$$

$\Rightarrow \underline{\underline{\Sigma_W}} = \underline{\underline{K}}$.

(v) $K(x,y) = q(K_1(x,y))$, q = polynomial with ≥ 0 coef

From (iv) we see that any power of K_1 is a kernel [take $K_1 = K_2 \Rightarrow K_1^2(x,y)$ is a kernel, etc].
Combined with (i) + (iii) gives the result.

(vi) $K(x,y) = \exp(K_1(x,y))$

$$K(x,y) = \sum_{k\geq 0} \frac{1}{k!}(K_1(x,y))^k$$

= polynomial with positive coefficients.

(In fact, one can show that if $\sum_{n\geq 0} a_n x^n$ is a power series with radius of convergence ρ, and $a_n \geq 0$, and if K is a kernel taking values in $(-\rho, \rho)$, then $\sum_{n\geq 0} a_n K^n$ is also a kernel.)

IV _ MERCER REPRESENTATION

Let • X = compact metric space

• K = continuous and symmetric function $X \times X \to \mathbb{R}$.

Then K admits a uniformly convergent expansion of the form

$$\boxed{K(x,y) = \sum_{j \geq 1} \lambda_j \varphi_j(x) \varphi_j(y)} \quad , \text{ with } \lambda_j > 0$$

iff. $\forall$ square integrable function $f \in \ell_2(X)$, the following condition holds

$$\boxed{\iint_{X \times X} K(x,y) f(x) f(y)\, dx\, dy \geqslant 0}$$

aka MERCER'S CONDITION

* Remarks (i) Let $K_n(x,y) := \sum_{j=1}^{n} \lambda_j \varphi_j(x) \varphi_j(y)$

Uniform convergence of K_n towards K means that

$$\sup_{(x,y) \in X \times X} | K_n(x,y) - K(x,y) | \to 0 \quad \text{as } n \to +\infty.$$

(ii) Consider the integral operator $T_K : \ell_2(X) \to \ell_2(X)$:

$$(T_K f)(x) = \int K(x,y) f(y)\, dy.$$

↳ Note that if $\iint |K(x,y)|^2 dx\, dy < \infty$, then indeed $T_K f \in \ell_2(X)$:

$$\int |(T_K f)(x)|^2 dx = \int \left(\int K(x,y) f(y) dy \right)^2 dx$$
$$\leqslant \int \left(\int |K(x,y)|^2 dy \right) \left(\int |f(y)|^2 dy \right) dx$$
$$= \|f\|_2 \iint |K(x,y)|^2 dx\, dy.$$

Then one can show that the functions φ_j appearing in the expansion of K correspond to the eigenfunctions associated with the operator T_K, and that $\{\varphi_j\}$ are orthonormal in $\ell_2(X)$: $\int \varphi_i(x) \varphi_j(x)\, dx = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{o/w} \end{cases}$

Coefficients $\lambda_j > 0$ are the associated eigenvalues.

More generally, we can consider a measure space (X, μ), so that $\int \varphi_i(x) \varphi_j(y) \mu(dx) = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{o/w} \end{cases}$.
Here, we take μ = lebesgue measure.

(iii) The sequence $\{ \sqrt{\lambda_j}\, \varphi_j(x) \}_{j \geq 1}$ is square integrable (the space of square integrable sequences is denoted $\ell_2(\mathbb{N})$),

since

$$\sum_{j \geq 1} \left(\sqrt{\lambda_j}\, \varphi_j(x) \right)^2 = \sum_{j \geq 1} \lambda_j \varphi_j^2(x) = K(x,x) < +\infty$$

We can easily extract a feature map and a feature space from Mercer's representation:

$$\Phi(x) = \left(\dots \sqrt{\lambda_j}\, \varphi_j(x) \dots \right)^t \Rightarrow K(x,y) = \langle \Phi(x), \Phi(y) \rangle_{\ell_2(\mathbb{N})}$$

where $\langle \cdot , \cdot \rangle_{\ell_2(\mathbb{N})}$ denotes the inner product in the Hilbert space $\ell_2(\mathbb{N})$: $\langle x, y \rangle = \sum_{n \geq 1} x_n y_n$, where $x = \{x_n\}_{n \geq 1}$, $y = \{y_n\}_{n \geq 1} \in \ell_2(\mathbb{N})$.

(iv) The series $\sum_{j \geq 1} c_j \varphi_j(x)$ converges absolutely $\forall x \in X$ whenever the sequence $\{ c_j / \sqrt{\lambda_j} \}$ is square integrable:

$$\sum_{j \geq 1} | c_j \varphi_j(x) | \leqslant \left(\sum \left(\frac{c_j}{\sqrt{\lambda_j}} \right)^2 \right)^{1/2} \left(\sum \left(\sqrt{\lambda_j}\, \varphi_j(x) \right)^2 \right)^{1/2}$$
$$= \| \{ c_j / \sqrt{\lambda_j} \} \|_{\ell_2(\mathbb{N})} \sqrt{K(x,x)}.$$

• **Theorem**: Let • X = compact metric space
• $K : X \times X \to \mathbb{R}$ a continuous kernel.

Define

$$\mathcal{H} := \left\{ f \mid f(x) = \sum_{j \geq 1} c_j \varphi_j(x) \; ; \; \left\{ \frac{c_j}{\sqrt{\lambda_j}} \right\} \in \ell_2(\mathbb{N}) \right\},$$

with inner product

$$\left\langle \sum_{j \geq 1} c_j \varphi_j \, , \, \sum_{j \geq 1} d_j \varphi_j \right\rangle_{\mathcal{H}} := \sum_{j \geq 1} \frac{c_j d_j}{\lambda_j}$$

Then $\mathcal{H}$ is the RKHS associated with K.

proof: We do not prove that $\langle \cdot , \cdot \rangle_{\mathcal{H}}$ defines an inner product, and that $\mathcal{H}$ is an Hilbert space.

We show that K is a reproducing kernel.

↘ $K(\cdot, x) \in \mathcal{H}$ since $K(\cdot, x) = \sum_{j \geq 1} \underbrace{\lambda_j \varphi_j(x)}_{=: c_j} \varphi_j(\cdot)$

and $$\sum_{j \geq 1} \frac{c_j^2}{\lambda_j} = \sum_{j \geq 1} \frac{\lambda_j^2 \varphi_j^2(x)}{\lambda_j} = \sum_{j \geq 1} \lambda_j \varphi_j^2(x) = K(x,x) < +\infty$$

↘ $$\langle f, K(\cdot, x) \rangle_{\mathcal{H}} = \left\langle \sum_{j \geq 1} c_j \varphi_j(\cdot), \sum_{k \geq 1} \lambda_k \varphi_k(x) \varphi_k(\cdot) \right\rangle_{\mathcal{H}}$$
$$= \sum_{j \geq 1} \frac{c_j \cancel{\lambda_j} \varphi_j(x)}{\cancel{\lambda_j}}$$
$$= \sum_{j \geq 1} c_j \varphi_j(x) = f(x)$$

$\mathcal{H}$ is a Hilbert space of functions with reproducing kernel K, so it must be equal to the RKHS $\mathcal{H}$ by uniqueness of RKHS.

× **Remark**: The inner product $\langle \cdot , \cdot \rangle_{\mathcal{H}}$ defined previously coincides with the inner product $\langle \cdot , \cdot \rangle_{\mathcal{H}_0}$ defined on the pre-Hilbert space $\mathcal{H}_0$ (page 19):

Let $f, g \in \mathcal{H}_0$: $f(\cdot) = \sum \alpha_i K(\cdot, x_i) = \sum_{i,j} \alpha_i \lambda_j \varphi_j(x_i) \varphi_j(\cdot)$

$g(\cdot) = \sum \gamma_k K(\cdot, y_k) = \sum_{k,j} \gamma_k \lambda_j \varphi_j(y_k) \varphi_j(\cdot)$

Then

$$f(\cdot) = \sum_j \overbrace{\left(\lambda_j \sum_i \alpha_i \varphi_j(x_i) \right)}^{=: c_j} \varphi_j(\cdot)$$

$$g(\cdot) = \sum_j \underbrace{\left(\lambda_j \sum_k \gamma_k \varphi_j(y_k) \right)}_{=: d_k} \varphi_j(\cdot)$$

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{j \geq 1} \frac{c_j d_j}{\lambda_j}$$
$$= \sum_{j \geq 1} \frac{1}{\lambda_j} \left(\lambda_j \sum_i \alpha_i \varphi_j(x_i) \right) \left(\lambda_j \sum_k \gamma_k \varphi_j(y_k) \right)$$
$$= \sum_{j \geq 1} \lambda_j \left(\sum_i \alpha_i \varphi_j(x_i) \right) \left(\sum_k \gamma_k \varphi_j(y_k) \right)$$
$$= \sum_{i,k} \alpha_i \gamma_k \underbrace{\left(\sum_j \lambda_j \varphi_j(x_i) \varphi_j(y_k) \right)}_{= K(x_i, y_k)}$$
$$= \sum_{i,k} \alpha_i \gamma_k K(x_i, y_k)$$
$$= \langle f, g \rangle_{\mathcal{H}_0}.$$

■

V - APPLICATIONS IN MACHINE LEARNING.

In this section, we prove the representer theorem, which states that when looking for a function in an RKHS (possibly of ∞ dimension) to optimize some penalized cost function, it is sufficient to look for a solution in a finite dimensional su space of the RKHS.

Let $\mathcal{L}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ be our learning sample, with $X_i \in \mathcal{X}$, $Y_i \in \mathbb{R}$.

Let . $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be a loss function.

. K = a kernel on $\mathcal{X}$, with associated RKHS $\mathcal{H}$.

Consider the minimization of the penalized criterion

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) + \lambda \| f \|_{\mathcal{H}}}$$

goodness of fit term.

$f \in \mathcal{H}$

$\lambda > 0$ (tuning parameter)

penalty term

↳ Why considering $\| f \|_{\mathcal{H}}$ as a penalty term?

(i) First, note that $f \in \mathcal{H}$ implies that $\| f \|_{\mathcal{H}} < +\infty$. This in turn means that f cannot fluctuate too much.
Consider the Gaussian kernel $K(x,y) = \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right)$, $\mathcal{X}$ = compact $\subset \mathbb{R}$.

Mercer representation: $f(x) = \sum_{j \geq 1} a_j \sqrt{\lambda_j}\, \varphi_j(x)$, with $\{a_j\} \in \ell_2(\mathbb{N})$, and $\| f \|_{\mathcal{H}} = \sum_{j\geq 1} a_j^2$.

⇒ The coefficients a_j must be decaying fast enough to 0 with the index j: we are penalizing functions φ_j with a large index more.

For a Gaussian kernel, it is possible to show that

$$\lambda_j = \sqrt{\frac{2a}{A}}\, B^j$$

$$\varphi_j(x) = \exp\{-(c-a)x^2\}\, H_j(x\sqrt{2c}),$$

where

$a^{-1} = 4\sigma^2$ $\qquad$ $A = a+b+c$

$b^{-1} = 2\sigma^2$ $\qquad$ $B = b/A < 1$

$c = \sqrt{a^2 + 2ab}$

&

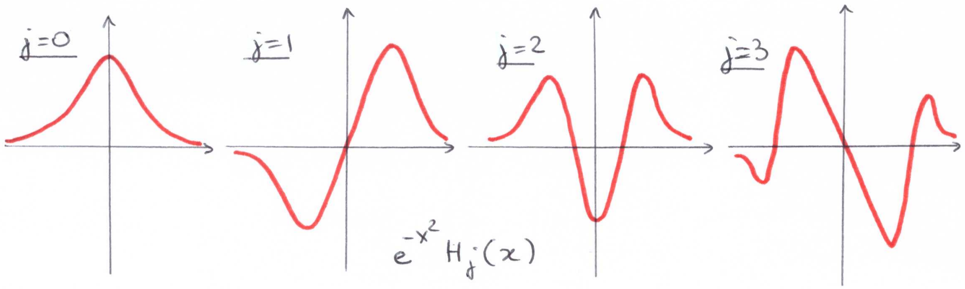
H_j = j-th order Hermite polynomial

$$H_j(x) = (-1)^j e^{x^2} \frac{d^j}{dx^j} e^{-x^2} \longrightarrow$$

. $H_0(x) = 1$

. $H_1(x) = 2x$

. $H_2(x) = 2(2x^2 - 1)$



⇒ Functions φ_j with a large index are more wiggly!

⇒ Functions in the RKHS cannot be too wild → what we want.

(ii) Second, recall that $\|f\|_{\mathcal{H}}$ close to 0 implies that $\forall x \in X$, $f(x)$ is also nearly 0.
(since the evaluation functional is continuous)

(iii) Third, for $f \in \mathcal{H}$, $\forall x, y \in X$,

$$|f(x) - f(y)| = |\langle f, K(\cdot, x)\rangle - \langle f, K(\cdot, y)\rangle|$$
$$= |\langle f, K(\cdot, x) - K(\cdot, y)\rangle|$$
$$\leq \|f\|_{\mathcal{H}} \, \|K(\cdot, x) - K(\cdot, y)\|_{\mathcal{H}},$$

where

$$\|K(\cdot, x) - K(\cdot, y)\|_{\mathcal{H}}$$
$$= \sqrt{\langle K(\cdot,x) - K(\cdot,y), K(\cdot,x) - K(\cdot,y)\rangle}$$
$$= \sqrt{K(x,x) + K(y,y) - 2K(x,y)}$$
$$=: d_K(x,y)$$

This quantity is ≥ 0 since it corresponds to
$$\begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} K(x,x) & K(x,y) \\ K(x,y) & K(y,y) \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$
& K is positive definite

Check: d_K is a distance:
(i) $d_K(x,y) \geq 0$
(ii) Symmetry
(iii) $d_K(x,y) = 0 \Leftrightarrow x = y$
(iv) $d_K(x,z) \leq d_K(x,y) + d_K(y,z)$.

Thus $|f(x) - f(y)| \leq \|f\|_{\mathcal{H}} \, d_K(x,y)$
↳ f is Lipschitz with respect to d_K, with Lipschitz constant $\|f\|_{\mathcal{H}}$
⇒ $\|f\|_{\mathcal{H}}$ controls the smoothness of f.

Theorem (Representer Theorem)

Let. K = kernel $X \times X \to \mathbb{R}$ with associated RKHS $\mathcal{H}$
- $G = \mathbb{R}_+ \to \mathbb{R}$ a strictly increasing function.
- $\mathcal{L}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ = learning sample.

Then the solutions to the optimization problem

$$\boxed{\underset{f \in \mathcal{H}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) + G(\|f\|_{\mathcal{H}})}$$

all have the form $f^*(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$

proof: Let $f \in \mathcal{H}$.
Consider the finite dimensional subspace spanned by the $K(\cdot, x_i)$, $i = 1, \dots, n$.
↳ A closed subspace of $\mathcal{H}$; denote it M

The theorem of projections in Hilbert spaces yields the decomposition
$f = f_1 + f_2$, where f_1 = orthogonal projection of f onto M
f_2 = the orthogonal complement ($\in M^\perp$)

Then

$$f(x_i) = \langle f, K(\cdot, x_i)\rangle_{\mathcal{H}}$$
$$= \langle f_1 + f_2, K(\cdot, x_i)\rangle_{\mathcal{H}}$$
Since $f_2 \in M^\perp$
$$= \langle f_1, K(\cdot, x_i)\rangle_{\mathcal{H}}$$
Since $f_1 \in M$
$$= \left\langle \sum_{j=1}^{n} \alpha_j K(\cdot, x_j), K(\cdot, x_i)\right\rangle_{\mathcal{H}}$$
$$= \sum_{j=1}^{n} \alpha_j \langle K(\cdot, x_j), K(\cdot, x_i)\rangle_{\mathcal{H}}$$
$$= \sum_{j=1}^{n} \alpha_j K(x_i, x_j)$$

$\Rightarrow$ The goodness of fit term in the expression of the penalized criterion does not depend on f_2.

In addition, $G(\|f\|_{\mathcal{H}}) = G\left(\sqrt{\|f_1\|_{\mathcal{H}}^2 + \|f_2\|_{\mathcal{H}}^2}\right)$

$$\geqslant G(\|f_1\|_{\mathcal{H}}),$$

where equality is obtained if $f_2 = 0$.

$\Rightarrow$ Choosing $f_2 = 0$ does not affect the empirical risk, but strictly decreases the penalty term $\Rightarrow$ Any minimizer must have $f_2 = 0$. ∎

VI.1. Kernel Ridge Regression.

Consider the minimization of a penalized criterion with square loss, in an RKHS $\mathcal{H}$:

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

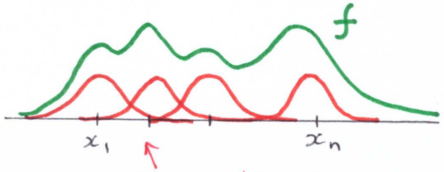
The representer theorem ensures that the minimizer has the form $f(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$.

$\hookrightarrow$ The goal is to compute the α_is.

Put $\alpha = \begin{pmatrix} \alpha_1 \\ | \\ \alpha_n \end{pmatrix} \in \mathbb{R}^n$

$$K_x = \begin{pmatrix} K(x, x_1) \\ \vdots \\ K(x, x_n) \end{pmatrix} \in \mathbb{R}^n$$

$\Rightarrow f(x) = \alpha^t K_x$



Gaussian Kernel.

Moreover,

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} = \begin{pmatrix} K_{x_1}^t \\ \vdots \\ K_{x_n}^t \end{pmatrix} \alpha = \begin{pmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & & \\ K(x_n, x_1) & \cdots & K(x_n, x_n) \end{pmatrix} \alpha$$

$$= \underline{\underline{K}}\, \alpha$$

Gram Matrix.

We obtain

- $\sum_{i=1}^{n} (y_i - f(x_i))^2 = (y - \underline{\underline{K}}\alpha)^t (y - \underline{\underline{K}}\alpha)$ $\quad y = \begin{pmatrix} y_1 \\ | \\ y_n \end{pmatrix}$

- $\|f\|_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}} = \sum_{i,j} \alpha_i \alpha_j \langle K(\cdot, x_i), K(\cdot, x_j) \rangle$

$$= \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)$$

$$= \alpha^t \underline{\underline{K}}\, \alpha$$

$\Rightarrow$ Solve / Compute $\hat{\alpha} := \underset{\alpha \in \mathbb{R}^n}{\operatorname{argmin}} (y - \underline{\underline{K}}\alpha)^t (y - \underline{\underline{K}}\alpha) + \lambda \alpha^t \underline{\underline{K}} \alpha$

$$= (\underline{\underline{K}} + \lambda I_n)^{-1} y.$$

Prediction is $\hat{y} = \underline{\underline{K}}\hat{\alpha} = \underline{\underline{K}}(\underline{\underline{K}} + \lambda I_n)^{-1} y$

x Remark: Special case $K(x, y) = \langle x, y \rangle$

We get

$$\underset{n\times n}{\underline{\underline{K}}} = \begin{pmatrix} X_1^t X_1 & \cdots & X_1^t X_n \\ \vdots & & \vdots \\ X_1^t X_n & \cdots & X_n^t X_n \end{pmatrix} = \underset{n\times d}{\begin{pmatrix} — X_1^t — \\ \\ — X_n^t — \end{pmatrix}} \underset{d\times n}{\begin{pmatrix} | & & | \\ X_1 & \cdots & X_n \\ | & & | \end{pmatrix}}$$

$$= X X^t$$

where $\underset{n\times d}{X}$ = matrix of observations.

Thus $\hat{\alpha} = (K + \lambda I_n)^{-1} y = (XX^t + \lambda I_n)^{-1} y$, (38)

and $\hat{y} = K\hat{\alpha}$

$= XX^t (XX^t + \lambda I_n)^{-1} y$.

We use the following identity:

$$\underset{I_d}{P}\ \underset{X^t}{B^t}(\underset{X}{B}\ \underset{I_d}{P}\ \underset{X^t}{B^t} + \underset{\lambda I_n}{R})^{-1} = (\underset{I_d}{P^{-1}} + \underset{X^t}{B^t}\underset{\frac{1}{\lambda}I_n}{R^{-1}}\underset{X}{B})^{-1}\underset{X^t}{B^t}\underset{\frac{1}{\lambda}I_n}{R^{-1}}$$

$$\Rightarrow X^t (XX^t + \lambda I_n)^{-1} = \frac{1}{\lambda}\left(I_d + \frac{1}{\lambda}X^t X\right)^{-1} X^t$$
$$= (X^t X + \lambda I_d)^{-1} X^t,$$

so that

$$\hat{y} = XX^t (XX^t + \lambda I_n)^{-1} y$$
$$= X \underbrace{(X^t X + \lambda I_d)^{-1} X^t}_{= H_\lambda} y = X\hat{\beta}_\lambda, \text{ where}$$

$$\hat{\beta}_\lambda = (X^t X + \lambda I_d)^{-1} X^t y$$

$=$ ridge solution $\rightarrow$ cf SL: RIDGE REG & LASSO

I.2. Kernel SVM.

Empirical Risk Minimization formulation of the SVM objective

is $\min\limits_{\beta_0, \beta} \frac{1}{n}\sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \lambda \|\beta\|^2$

hinge loss $\qquad f(x) = \beta_0 + \beta^t x$.

$\rightarrow$ cf SL: SUPPORT VECTOR MACHINE

Let K = kernel with associated RKHS $\mathcal{H}$. (39)

Consider the following optimization problem:

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

The representer theorem ensures that the minimizer of $(*)$ is of the form $\sum_{j=1}^{n} \beta_j K(x, x_j)$.

$\hookrightarrow \|f\|^2_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}} = \beta^t \underline{\underline{K}} \beta$ (same as before).

Optimization problem $(*)$ can be re-expressed in its equivalent form

$$(**) \quad \begin{array}{|ll|} \text{minimize}_{\beta, \xi} & \frac{1}{2}\beta^t K \beta + C\sum_{i=1}^{n} \xi_i \\ \text{s.t.} & y_i\left(\sum_{j=1}^{n} \beta_j K(x_i, x_j)\right) \geq 1 - \xi_i \\ & \xi_i \geq 0 \\ \text{PRIMAL PROBLEM} & \\ \hline \end{array}$$

Analysis of the solution to the primal problem using KKT conditions make the equivalence formal.

The Lagrangian is

$$\mathcal{L}(\beta, \xi, \lambda, \nu) = \frac{1}{2}\beta^t K \beta + C\sum \xi_i - \sum \lambda_i (y_i f(x_i) + \xi_i - 1) - \sum \nu_i \xi_i.$$

- **KKT conditions:**

① Primal Constraints $\quad y_i f(x_i) - 1 + \xi_i \geq 0$, $\quad \xi_i \geq 0$

② Dual Constraints $\quad \lambda_i \geq 0$, $\quad \nu_i \geq 0$

③ Complementary Slackness $\lambda_i\left(y_i f(x_i) - 1 + \xi_i\right) = 0$ (40)

$v_i \xi_i = 0$

④ Gradient of $\mathcal{L}$ w.r.t. β, ξ vanishes

[4.1] • $\dfrac{\partial \mathcal{L}}{\partial \xi_i} = C - v_i - \lambda_i = 0$

[4.2] • $\dfrac{\partial \mathcal{L}}{\partial \beta_k} = \sum_{j=1}^{n} \beta_j K(x_j, x_k) - \sum_{i=1}^{n} y_i \lambda_i K(x_i, x_k) = 0$

$$\frac{\partial}{\partial \beta_k} \frac{1}{2} \beta^t K \beta = \frac{\partial}{\partial \beta_k} \frac{1}{2} \sum_{i,j} \beta_i \beta_j K(x_i, x_j)$$

$$= \frac{1}{2}\left\{ 2\beta_k K(x_k, x_k) + 2 \sum_{j \neq k} \beta_j K(x_k, x_j) \right\}$$

$$= \sum_{j=1}^{n} \beta_j K(x_j, x_k)$$

$$\frac{\partial}{\partial \beta_k} y_i \lambda_i f(x_i) = \frac{\partial}{\partial \beta_k} y_i \lambda_i \sum_{j=1}^{n} \beta_j K(x_i, x_j)$$

$$= y_i \lambda_i K(x_i, x_k)$$

• Dual problem.

Expression [4.1] allows us to express the coefficients β as a function of $\lambda \mapsto \beta = \beta(\lambda)$.

$$\mathcal{L}(\beta(\lambda), \xi, \lambda, v) = \frac{1}{2} \beta(\lambda)^t K \beta(\lambda) + \cancel{C \sum} \xi_i - \sum y_i f(x_i) f(x_i) - \sum \cancel{\xi_i \lambda_i} + \sum \lambda_i - \cancel{\sum_i \xi_i} v_i$$

since [4.1] $v_i = C - \lambda_i$

(41)

$$\Rightarrow \mathcal{L}(\beta(\lambda), \lambda) = \sum \lambda_i + \underbrace{\frac{1}{2}\beta(\lambda)^t K \beta(\lambda)}_{\mathrm{I}} - \underbrace{\sum y_i \lambda_i f(x_i)}_{\mathrm{II}}$$

$$\mathrm{I} = \frac{1}{2} \sum_{j,k} \beta_j(\lambda) \beta_k(\lambda) K(x_j, x_k)$$

$$= \frac{1}{2} \sum_{j,k} \beta_k(\lambda)\, y_j \lambda_j K(x_j, x_k)$$

$$= \frac{1}{2} \sum_{j,k} \beta_k(\lambda)\, y_j \lambda_j K(x_j, x_k)$$

$$= \frac{1}{2} \sum_{j,k} y_j y_k \lambda_j \lambda_k K(x_j, x_k)$$

$$\mathrm{II} = \sum_i y_i \lambda_i \sum_j \beta_j K(x_i, x_j) \underset{[4.2]}{=} \sum_{i,j} y_i y_j \lambda_i \lambda_j K(x_i, x_j)$$

× The Lagrange dual function is

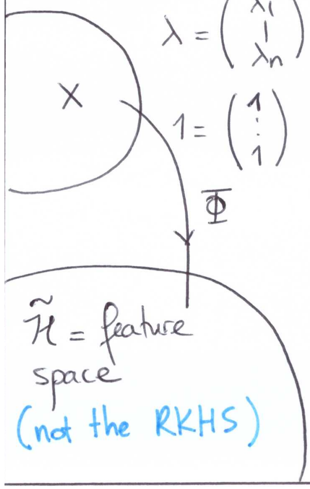
$$\ell(\lambda) = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j} y_i y_j \lambda_i \lambda_j K(x_i, x_j).$$

Put $\underset{n \times n}{H} = (H_{ij})$; $H_{ij} := y_i y_j K(x_i, x_j)$

$= y_i y_j \langle \Phi(x_i), \Phi(x_j) \rangle_{\tilde{\mathcal{H}}}$

$\lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix}$

$1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$



Inner product in the feature space $\tilde{\mathcal{H}}$.

↳ Can be efficiently computed by simply evaluating the bivariate function K at x_i and x_j.

& we do not need to explicitely construct the feature map Φ : all is done automatically.

The dual problem is

$$\boxed{\begin{array}{ll} \text{maximize} & 1^t\lambda - \frac{1}{2}\lambda^t H \lambda \\ \text{s.t.} & 0 \preccurlyeq \lambda \preccurlyeq C \end{array}}$$

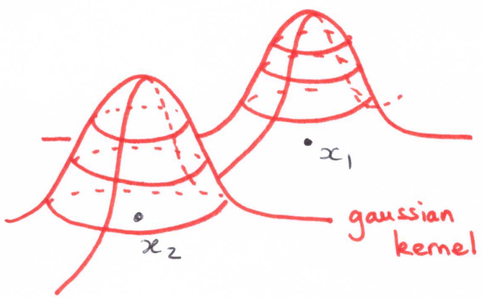
DUAL PROBLEM

Denote the optimal point λ^*.

↳ The (kernel) soft classifier is $\sum_{i=1}^{n} y_i \lambda_i^* K(x, x_i)$,

and the associated hard classifier $\text{sign}\left(\sum_{i=1}^{n} y_i \lambda_i^* K(x, x_i)\right)$.

Compare with the SVM original classifier, associated with $K(x, y) = \langle x, y \rangle = x^t y$: $\text{sign}\left(\sum_i y_i \lambda_i^* x_i^t x\right)$.



The substitution $\langle x, y \rangle \rightarrow K(x, y)$ is known in the litterature as the <u>kernel trick</u>: « Any algorithm that process finite-dimensional vectors that can be expressed only in terms of pairwise inner-products can be applied to potentially ∞-dimensional vectors in the feature space of a positive definite kernel by replacing each inner product evaluation by a kernel evaluation ».