Statistical Learning II

Lecture 12 - Kernel methods

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
11.12.2024

Feature maps

Idea: Introduce a feature map:

$$\boldsymbol{\varphi} : \mathbb{R}^d \to \mathbb{R}^p$$
$$\boldsymbol{x} \mapsto \boldsymbol{\varphi}(\boldsymbol{x})$$

And consider a linear predictor in feature space:

$$f_\theta(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(\boldsymbol{x}) \rangle$$

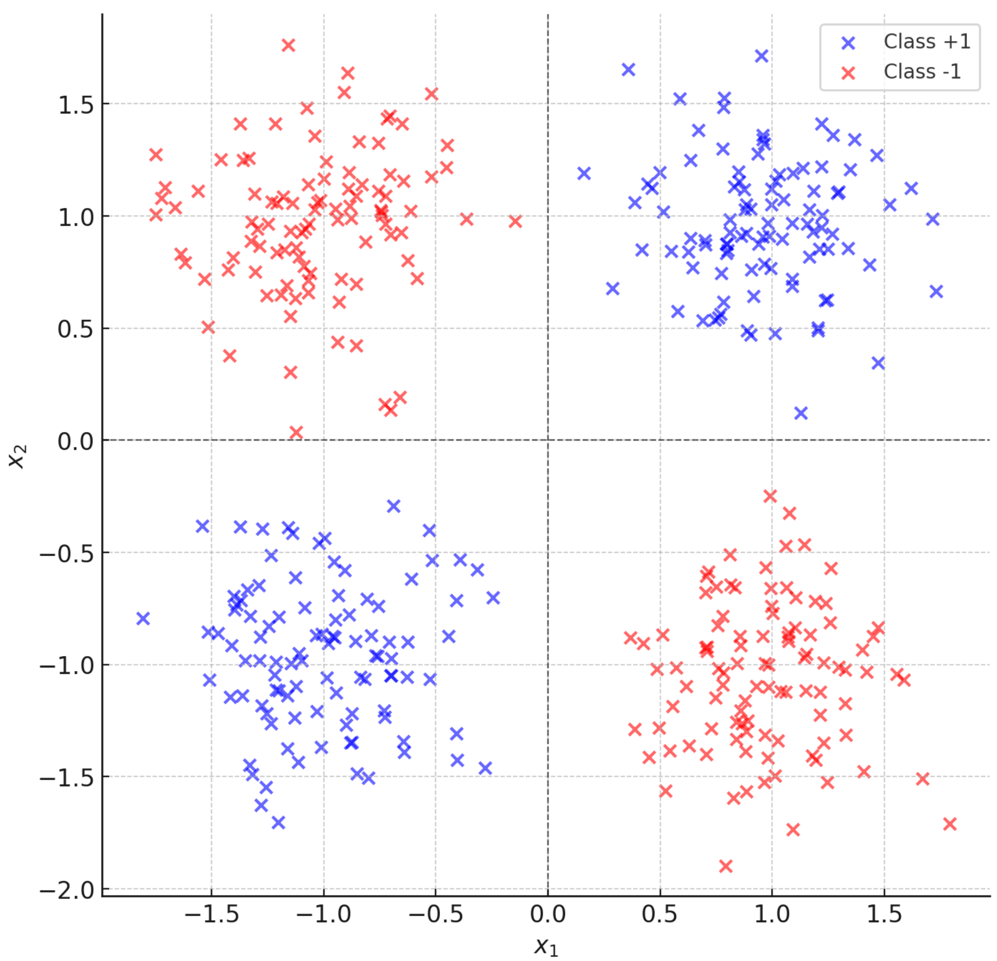
- Now we have $\boldsymbol{\theta} \in \mathbb{R}^p$.
- f_θ still a linear function of $\boldsymbol{\theta}$.
- Typically $p > d$.
- More generally, we can consider $\boldsymbol{\varphi} : \mathscr{X} \to \mathbb{R}^p$

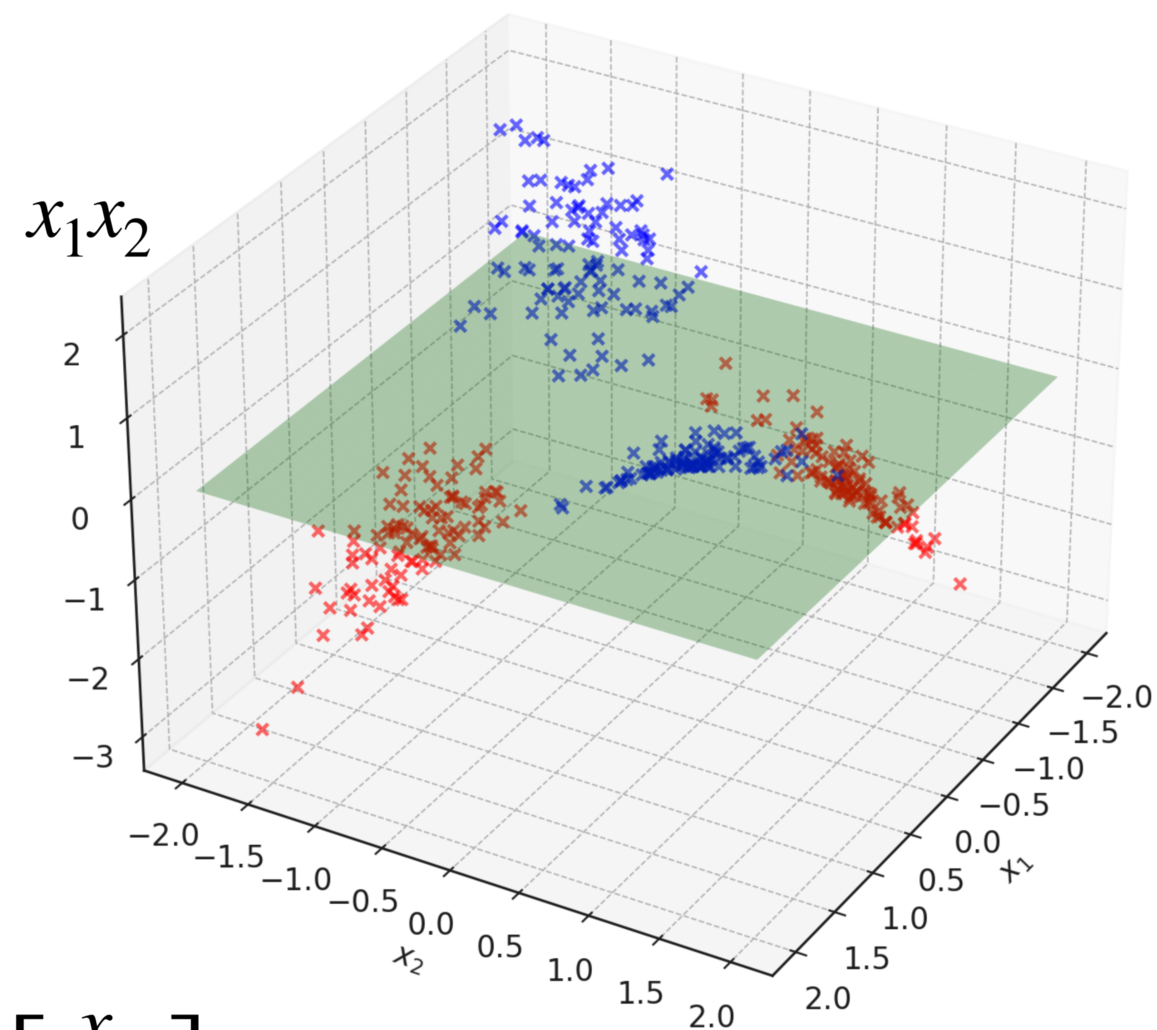
Example: $\mathscr{X}$ a collection of books.

Examples: XOR Gaussian mixture

$x \in \mathbb{R}^2 \quad (d = 2)$

$$p(\boldsymbol{x}) = \frac{1}{4} \sum_{k=1}^{4} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{I}_2)$$





$$\boldsymbol{\varphi}(\boldsymbol{x}) = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}$$

Ridge regression on feature space

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathbb{R} : i \in [n]\}$ denote training data and $\varphi : \mathscr{X} \to \mathbb{R}^p$ a feature map.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(x_i) \rangle)^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2$$

Ridge regression on feature space

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathbb{R} : i \in [n]\}$ denote training data and $\varphi : \mathscr{X} \to \mathbb{R}^p$ a feature map.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(x_i) \rangle)^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2$$

Defining the feature matrix and label vector:

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\varphi}(x_1) \\ \vdots \\ \boldsymbol{\varphi}(x_n) \end{bmatrix} \in \mathbb{R}^{n \times p} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$$

Ridge regression on feature space

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathbb{R} : i \in [n]\}$ denote training data and $\varphi : \mathscr{X} \to \mathbb{R}^p$ a feature map.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(x_i) \rangle)^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2$$

Defining the feature matrix and label vector:

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\varphi}(x_1) \\ \vdots \\ \boldsymbol{\varphi}(x_n) \end{bmatrix} \in \mathbb{R}^{n \times p} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$$

The above admits an explicit solution:

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y}) = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + n\lambda \boldsymbol{I}_p)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y}$$

Ridge regression on feature space

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathbb{R} : i \in [n]\}$ denote training data and $\varphi : \mathscr{X} \to \mathbb{R}^p$ a feature map.

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle \boldsymbol{\theta}, \boldsymbol{\varphi}(x_i) \rangle)^2 + \frac{\lambda}{2} ||\boldsymbol{\theta}||_2^2$$

Note we can equivalently write:

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y}) = \begin{cases} (\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + n\lambda \boldsymbol{I}_p)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y} \\ \boldsymbol{\Phi}^\top (\boldsymbol{\Phi} \boldsymbol{\Phi}^\top + n\lambda \boldsymbol{I}_n)^{-1} \boldsymbol{y} \end{cases}$$

Same result, but one might be cheaper than the other.

Kernels

Note that the solution:

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y}) = \boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + n\lambda \boldsymbol{I}_n)^{-1}\boldsymbol{y}$$

Actually lives in the $\mathrm{span}(\boldsymbol{\varphi}(x_1), \ldots, \boldsymbol{\varphi}(x_n))$.

Kernels

Note that the solution:

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y}) = \boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + n\lambda \boldsymbol{I}_n)^{-1}\boldsymbol{y}$$

Actually lives in the $\mathrm{span}(\boldsymbol{\varphi}(x_1), \ldots, \boldsymbol{\varphi}(x_n))$. This means we can also write:

$$\hat{\boldsymbol{\theta}}_\lambda = \boldsymbol{\Phi}^\top \hat{\boldsymbol{\alpha}}_\lambda \qquad \hat{\boldsymbol{\alpha}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y}) = (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + n\lambda \boldsymbol{I}_n)^{-1}\boldsymbol{y}$$

Kernels

Note that the solution:

$$\hat{\boldsymbol{\theta}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y}) = \boldsymbol{\Phi}^\top(\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + n\lambda \boldsymbol{I}_n)^{-1}\boldsymbol{y}$$

Actually lives in the $\mathrm{span}(\boldsymbol{\varphi}(x_1), \ldots, \boldsymbol{\varphi}(x_n))$. This means we can also write:

$$\hat{\boldsymbol{\theta}}_\lambda = \boldsymbol{\Phi}^\top \hat{\boldsymbol{\alpha}}_\lambda \qquad \hat{\boldsymbol{\alpha}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y}) = (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + n\lambda \boldsymbol{I}_n)^{-1}\boldsymbol{y}$$

And the predictor:

$$f_\theta(x) = \langle \hat{\boldsymbol{\theta}}_\lambda, \boldsymbol{\varphi}(x) \rangle = \langle \hat{\boldsymbol{\alpha}}_\lambda, \boldsymbol{\Phi}\boldsymbol{\varphi}(x) \rangle$$

Kernels

$$f_\theta(x) = \langle \hat{\boldsymbol{\theta}}_\lambda, \boldsymbol{\varphi}(x) \rangle = \langle \hat{\boldsymbol{\alpha}}_\lambda, \boldsymbol{\Phi}\boldsymbol{\varphi}(x) \rangle$$

$$\hat{\boldsymbol{\alpha}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y}) = (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + n\lambda \boldsymbol{I}_n)^{-1}\boldsymbol{y}$$

Note everything only depends on the scalar product of features

$$K(x, x') = \langle \boldsymbol{\varphi}(x), \boldsymbol{\varphi}(x') \rangle$$

This is also known as a *kernel*.

Kernels

$$f_\theta(x) = \langle \hat{\boldsymbol{\theta}}_\lambda, \boldsymbol{\varphi}(x) \rangle = \langle \hat{\boldsymbol{\alpha}}_\lambda, \boldsymbol{\Phi}\boldsymbol{\varphi}(x) \rangle$$

$$\hat{\boldsymbol{\alpha}}_\lambda(\boldsymbol{\Phi}, \boldsymbol{y}) = (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + n\lambda \boldsymbol{I}_n)^{-1}\boldsymbol{y}$$

Note everything only depends on the scalar product of features

$$K(x, x') = \langle \boldsymbol{\varphi}(x), \boldsymbol{\varphi}(x') \rangle$$

This is also known as a *kernel*.

⚠ This is true for any linear predictor, and goes under the name of "representer theorem"

Kernel methods

Hilbert space

As we have shown in the previous examples, it is easier to linearly separate a function in higher dimensions.

Key idea: Take the number of features to infinity ($p \to \infty$)

Hilbert space

As we have shown in the previous examples, it is easier to linearly separate a function in higher dimensions.

<u>Key idea:</u> Take the number of features to infinity ($p \to \infty$)

First, we need to make sense of $\mathbb{R}^\infty$...

Hilbert space

As we have shown in the previous examples, it is easier to linearly separate a function in higher dimensions.

Key idea: Take the number of features to infinity ($p \to \infty$)

First, we need to make sense of $\mathbb{R}^\infty$...

Definition (Hilbert space)

A Hilbert space $\mathscr{H}$ is a vector space (over $\mathbb{R}$ or $\mathbb{C}$) with an inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}} : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$ which is complete.

Hilbert space

As we have shown in the previous examples, it is easier to linearly separate a function in higher dimensions.

💡 <u>Key idea:</u> Take the number of features to infinity ($p \to \infty$)

🤔 First, we need to make sense of $\mathbb{R}^\infty$...

Definition (Hilbert space)

A Hilbert space $\mathscr{H}$ is a vector space (over $\mathbb{R}$ or $\mathbb{C}$) with an inner product $\langle \cdot, \cdot \rangle_{\mathscr{H}} : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$ which is complete.

Informally, an inner product is the minimum we need to do linear algebra in infinite dimensions

Hilbert space

Definition (Hilbert space)

A Hilbert space $\mathscr{H}$ is a vector space (over $\mathbb{R}$ or $\mathbb{C}$) with an inner product $\langle \cdot , \cdot \rangle_{\mathscr{H}} : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$ which is complete.

- Vector space (over $\mathbb{R}$): Let $a, b \in \mathbb{R}$ and $f, g \in \mathscr{H}$

$$af + bg \in \mathscr{H}$$ + usual properties of the sum

Hilbert space

Definition (Hilbert space)

A Hilbert space $\mathscr{H}$ is a vector space (over $\mathbb{R}$ or $\mathbb{C}$) with an inner product $\langle \cdot , \cdot \rangle_{\mathscr{H}} : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$ which is complete.

- Vector space (over $\mathbb{R}$): Let $a, b \in \mathbb{R}$ and $f, g \in \mathscr{H}$

$$af + bg \in \mathscr{H}$$ + usual properties of the sum

- Inner product: a function $\langle \cdot , \cdot \rangle_{\mathscr{H}} : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$ such that:
  - $\langle f, g \rangle_{\mathscr{H}} = \langle g, f \rangle_{\mathscr{H}}$
  - $||f||^2_{\mathscr{H}} = \langle f, f \rangle_{\mathscr{H}} \geq 0$ with equality iff $f = 0$
  - $\langle af + bg, h \rangle_{\mathscr{H}} = a\langle f, h \rangle_{\mathscr{H}} + b\langle g, h \rangle_{\mathscr{H}}$

Hilbert space

Definition (Hilbert space)

A Hilbert space $\mathscr{H}$ is a vector space (over $\mathbb{R}$ or $\mathbb{C}$) with an inner product $\langle \cdot , \cdot \rangle_{\mathscr{H}} : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$ which is complete.

- Vector space (over $\mathbb{R}$): Let $a, b \in \mathbb{R}$ and $f, g \in \mathscr{H}$

$$af + bg \in \mathscr{H}$$ + usual properties of the sum

- Inner product: a function $\langle \cdot , \cdot \rangle_{\mathscr{H}} : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$ such that:
  - $\langle f, g \rangle_{\mathscr{H}} = \langle g, f \rangle_{\mathscr{H}}$
  - $||f||^2_{\mathscr{H}} = \langle f, f \rangle_{\mathscr{H}} \geq 0$ with equality iff $f = 0$
  - $\langle af + bg, h \rangle_{\mathscr{H}} = a\langle f, h \rangle_{\mathscr{H}} + b\langle g, h \rangle_{\mathscr{H}}$

Inner product induces norm, but converse not always true.

Hilbert space

Definition (Hilbert space)

A Hilbert space $\mathscr{H}$ is a vector space (over $\mathbb{R}$ or $\mathbb{C}$) with an inner product $\langle \cdot , \cdot \rangle_{\mathscr{H}} : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$ which is complete.

- Vector space (over $\mathbb{R}$): Let $a, b \in \mathbb{R}$ and $f, g \in \mathscr{H}$

  $$af + bg \in \mathscr{H}$$ + usual properties of the sum

- Inner product: a function $\langle \cdot , \cdot \rangle_{\mathscr{H}} : \mathscr{H} \times \mathscr{H} \to \mathbb{R}$ such that:
  - $\langle f, g \rangle_{\mathscr{H}} = \langle g, f \rangle_{\mathscr{H}}$
  - $||f||^2_{\mathscr{H}} = \langle f, f \rangle_{\mathscr{H}} \geq 0$ with equality iff $f = 0$
  - $\langle af + bg, h \rangle_{\mathscr{H}} = a\langle f, h \rangle_{\mathscr{H}} + b\langle g, h \rangle_{\mathscr{H}}$

- Complete: Cauchy sequences $f_n \in \mathscr{H}$ converge $f_\infty \in \mathscr{H}$

Examples of Hilbert spaces

- $\mathscr{H} = \mathbb{R}^d$ with the usual Euclidean inner product:

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_2 = \sum_{i=1}^{d} v_i u_i$$

Examples of Hilbert spaces

- $\mathscr{H} = \mathbb{R}^d$ with the usual Euclidean inner product:

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_2 = \sum_{i=1}^{d} v_i u_i$$

- $\ell^2(\mathbb{R})$: sequences $\boldsymbol{u} = (u_1, u_2, \ldots)$ with

Such that: $$||\boldsymbol{u}||_{\ell^2}^2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle_{\ell^2} = \sum_{i=1}^{\infty} |u_i|^2 < \infty$$

Examples of Hilbert spaces

- $\mathscr{H} = \mathbb{R}^d$ with the usual Euclidean inner product:

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_2 = \sum_{i=1}^{d} v_i u_i$$

- $\ell^2(\mathbb{R})$: sequences $\boldsymbol{u} = (u_1, u_2, \ldots)$ with

Such that: $$||\boldsymbol{u}||_{\ell^2}^2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle_{\ell^2} = \sum_{i=1}^{\infty} |u_i|^2 < \infty$$

- $L^2(\mathbb{R})$: functions $f : \mathbb{R} \to \mathbb{R}$ with $$\langle f, g \rangle_{L^2(\mathbb{R})} = \int_{-\infty}^{\infty} f(x) g(x) \mathrm{dx}$$

Such that: $$||f||_{L^2(\mathbb{R})}^2 = \langle f, f \rangle_{L^2(\mathbb{R})} = \int_{-\infty}^{\infty} |f(x)|^2 \mathrm{dx} < \infty$$

Infinite dimensional features

This provides the right structure to define infinite dimensions features.

Infinite dimensional features

This provides the right structure to define infinite dimensions features.

Let $\mathscr{H}$ denote a Hilbert space with inner product $\langle \cdot , \cdot \rangle_{\mathscr{H}}$.

Idea: Given data $x \in \mathscr{X}$, define features:

$$\varphi : \mathscr{X} \to \mathscr{H}$$
$$x \mapsto \varphi(x)$$

and predictors: $f_\theta(x) = \langle \theta, \varphi(x) \rangle_{\mathscr{H}}$ with $\theta \in \mathscr{H}$.

Infinite dimensional features

This provides the right structure to define infinite dimensions features.

Let $\mathscr{H}$ denote a Hilbert space with inner product $\langle \cdot , \cdot \rangle_{\mathscr{H}}$.

Idea: Given data $x \in \mathscr{X}$, define features:

$$
\begin{aligned}
\varphi : \mathscr{X} &\to \mathscr{H} \\
x &\mapsto \varphi(x)
\end{aligned}
$$

and predictors: $f_\theta(x) = \langle \theta, \varphi(x) \rangle_{\mathscr{H}}$ with $\theta \in \mathscr{H}$.

Problems:

- In general $f \notin \mathscr{H}$.
- Class of functions $f : \mathscr{X} \to \mathbb{R}$ defined this way can be small.

Example

Let $\mathscr{H} \subset \mathbb{R}^2$ with standard Euclidean inner product.

Let $\mathscr{X} = \{x_1, x_2, x_3\}$ be a discrete data space. Define $\varphi : \mathscr{X} \to \mathscr{H}$

$$\varphi(x_1) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \varphi(x_2) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \varphi(x_3) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Example

Let $\mathscr{H} \subset \mathbb{R}^2$ with standard Euclidean inner product.

Let $\mathscr{X} = \{x_1, x_2, x_3\}$ be a discrete data space. Define $\varphi : \mathscr{X} \to \mathscr{H}$

$$\varphi(x_1) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \varphi(x_2) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \varphi(x_3) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

For any $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \in \mathscr{H}$, define the function: $f(x) = \langle \theta, \varphi(x) \rangle$

We have: $f(x_1) = \theta_1 \quad f(x_2) = \theta_2 \quad f(x_3) = \theta_1 + \theta_2$

Example

Let $\mathscr{H} \subset \mathbb{R}^2$ with standard Euclidean inner product.

Let $\mathscr{X} = \{x_1, x_2, x_3\}$ be a discrete data space. Define $\varphi : \mathscr{X} \to \mathscr{H}$

$$\varphi(x_1) = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \varphi(x_2) = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \qquad \varphi(x_3) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

For any $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \in \mathscr{H}$, define the function: $\quad f(x) = \langle \theta, \varphi(x) \rangle$

We have: $\quad f(x_1) = \theta_1 \quad f(x_2) = \theta_2 \quad f(x_3) = \theta_1 + \theta_2$

Only few functions on $\mathscr{X}$ can be expressed this way.

e.g. can't express $\quad f(x_1) = 1 \quad f(x_2) = 0 \quad f(x_3) = 2$

Reproducing property

To make the Hilbert space compatible with $\mathscr{X}$, we need the following reproducing property:

Definition (RKHS)

A Hilbert space $\mathscr{H}$ of functions over $\mathscr{X}$ is said to be a "Reproducing Kernel Hilbert Space" (RKHS) if there exists $\varphi \in \mathscr{H}$ such that:

$$\forall x \in \mathscr{X} \quad \forall f \in \mathscr{H} \qquad f(x) = \langle f, \varphi(x) \rangle_{\mathscr{H}}$$

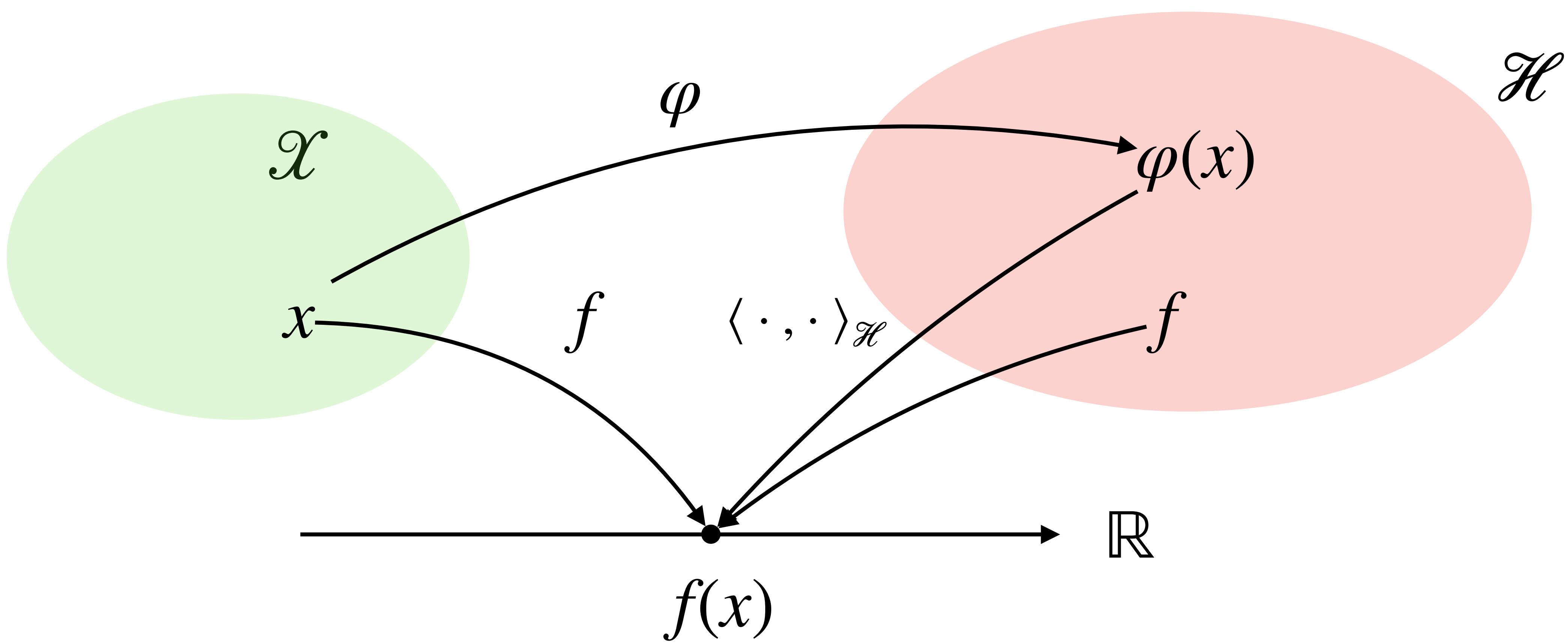
Reproducing property

To make the Hilbert space compatible with $\mathscr{X}$, we need the following reproducing property:

Definition (RKHS)

A Hilbert space $\mathscr{H}$ of functions over $\mathscr{X}$ is said to be a "Reproducing Kernel Hilbert Space" (RKHS) if there exists $\varphi \in \mathscr{H}$ such that:

$$\forall x \in \mathscr{X} \quad \forall f \in \mathscr{H} \qquad f(x) = \langle f, \varphi(x) \rangle_{\mathscr{H}}$$



Kernel ridge regression

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathbb{R} : i \in [n]\}$ denote training data. We now have everything we need to define ERM on a RKHS.

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

Kernel ridge regression

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathbb{R} : i \in [n]\}$ denote training data. We now have everything we need to define ERM on a RKHS.

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

By using the feature map $\varphi : \mathscr{X} \to \mathscr{H}$, this can be equivalently written:

$$\min_{\theta \in \mathscr{H}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle \theta, \varphi(x_i) \rangle_{\mathscr{H}})^2 + \frac{\lambda}{2} ||\theta||_{\mathscr{H}}^2$$

Kernel ridge regression

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathbb{R} : i \in [n]\}$ denote training data. We now have everything we need to define ERM on a RKHS.

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

By using the feature map $\varphi : \mathscr{X} \to \mathscr{H}$, this can be equivalently written:

$$\min_{\theta \in \mathscr{H}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle \theta, \varphi(x_i) \rangle_{\mathscr{H}})^2 + \frac{\lambda}{2} ||\theta||_{\mathscr{H}}^2$$

🤔 Closed-form in terms of "infinite dimensional" matrices "$\Phi \in \mathbb{R}^{n \times \infty}$"?

Kernel ridge regression

Let $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathbb{R} : i \in [n]\}$ denote training data. We now have everything we need to define ERM on a RKHS.

$$\min_{\theta \in \mathcal{H}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle \theta, \varphi(x_i) \rangle_{\mathcal{H}})^2 + \frac{\lambda}{2} ||\theta||_{\mathcal{H}}^2$$

As before, defining the kernel function and matrix

$$K(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}} \qquad \boldsymbol{K}_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathcal{H}}$$

Kernel ridge regression

Let $\mathscr{D} = \{(x_i, y_i) \in \mathscr{X} \times \mathbb{R} : i \in [n]\}$ denote training data. We now have everything we need to define ERM on a RKHS.

$$\min_{\theta \in \mathscr{H}} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \langle \theta, \varphi(x_i) \rangle_{\mathscr{H}})^2 + \frac{\lambda}{2} ||\theta||_{\mathscr{H}}^2$$

As before, defining the kernel function and matrix

$$K(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathscr{H}} \qquad \boldsymbol{K}_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle_{\mathscr{H}}$$

The solution can be written as:

$$\hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_{\lambda,i} K(x, x_i) \qquad \hat{\boldsymbol{\alpha}}_{\lambda}(\boldsymbol{\Phi}, \boldsymbol{y}) = (\boldsymbol{K} + n\lambda \boldsymbol{I}_n)^{-1} \boldsymbol{y}$$

Kernels

Note that in practice, to do ridge regression on $\mathscr{H}$ we don't even need to know what φ is. It suffices to have K.

Kernels

Note that in practice, to do ridge regression on $\mathcal{H}$ we don't even need to know what φ is. It suffices to have K.

Theorem (Aronszajn, 1950)

A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines a positive definite Kernel if and only if there exists a Hilbert space $\mathcal{H}$ and a map $\varphi : \mathcal{X} \to \mathcal{H}$ such that:

$$K(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}} \qquad \forall x, x' \in \mathcal{X}$$

Kernels

Note that in practice, to do ridge regression on $\mathscr{H}$ we don't even need to know what φ is. It suffices to have K.

Theorem (Aronszajn, 1950)

A function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ defines a positive definite Kernel if and only if there exists a Hilbert space $\mathscr{H}$ and a map $\varphi : \mathscr{X} \to \mathscr{H}$ such that:

$$K(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathscr{H}} \qquad \forall x, x' \in \mathscr{X}$$

In words: specifying $\mathscr{H}$ and φ is completely equivalent to specifying K,

Kernels

Note that in practice, to do ridge regression on $\mathscr{H}$ we don't even need to know what φ is. It suffices to have K.

Theorem (Aronszajn, 1950)

A function $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ defines a positive definite Kernel if and only if there exists a Hilbert space $\mathscr{H}$ and a map $\varphi : \mathscr{X} \to \mathscr{H}$ such that:

$$K(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathscr{H}} \qquad \forall x, x' \in \mathscr{X}$$

In words: specifying $\mathscr{H}$ and φ is completely equivalent to specifying K,

A kernel can correspond to several feature maps. e.g. $\mathscr{X} = \mathbb{R}$

$$\varphi(x) = x \qquad \varphi(x) = \frac{1}{\sqrt{2}} \begin{bmatrix} x \\ x \end{bmatrix} \qquad K(x, x') = xx'$$

Examples of Kernels

- Gaussian kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{1}{2\sigma^2}||\boldsymbol{x}-\boldsymbol{x}'||_2^2}$ (a.k.a. RBF)
- Laplace kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\lambda||\boldsymbol{x}-\boldsymbol{x}'||_2}$
- Polynomial kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + b)^k$

Examples of Kernels

- Gaussian kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{1}{2\sigma^2}||\boldsymbol{x}-\boldsymbol{x}'||_2^2}$ (a.k.a. RBF)
- Laplace kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\lambda||\boldsymbol{x}-\boldsymbol{x}'||_2}$
- Polynomial kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + b)^k$
- Translational invariant kernels $K(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\boldsymbol{x} - \boldsymbol{x}')$
- Rotationally invariant kernels $K(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle)$

Examples of Kernels

- Gaussian kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{1}{2\sigma^2}||\boldsymbol{x}-\boldsymbol{x}'||_2^2}$ (a.k.a. RBF)
- Laplace kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\lambda||\boldsymbol{x}-\boldsymbol{x}'||_2}$
- Polynomial kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + b)^k$
- Translational invariant kernels $K(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\boldsymbol{x} - \boldsymbol{x}')$
- Rotationally invariant kernels $K(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle)$

Or any other positive-definite function…

Examples of Kernels

- Gaussian kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\frac{1}{2\sigma^2}||\boldsymbol{x}-\boldsymbol{x}'||_2^2}$ (a.k.a. RBF)
- Laplace kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = e^{-\lambda||\boldsymbol{x}-\boldsymbol{x}'||_2}$
- Polynomial kernel: $K(\boldsymbol{x}, \boldsymbol{x}') = (\langle \boldsymbol{x}, \boldsymbol{x}' \rangle + b)^k$
- Translational invariant kernels $K(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\boldsymbol{x} - \boldsymbol{x}')$
- Rotationally invariant kernels $K(\boldsymbol{x}, \boldsymbol{x}') = \kappa(\langle \boldsymbol{x}, \boldsymbol{x}' \rangle)$

Or any other positive-definite function...

In general, finding φ associated to these is not obvious.

Examples of Kernels

$$y_i = \sin(x) + \varepsilon \qquad n = 100$$

$$\varepsilon \sim \mathcal{N}(0, 0.2^2) \qquad \lambda = 0.1$$