Kernel methods

Eder Arley León Gómez

Universidad Nacional de Colombia

May 4, 2023

Schedule

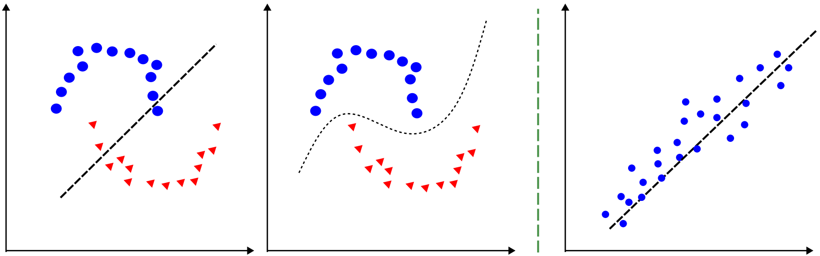
Introduction

Motivation

$$\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^{N}, \text{ where } \mathbf{x}_n \in \mathbf{X} \text{ and } \mathbf{y}_n \in \mathbf{y}$$

A linear model finds a hyperplane $\mathbf{W}$ such that the output is:

$$f(\mathbf{X}) = \mathbf{W}^\top \mathbf{X} \tag{1}$$



Mathematical review

Basic definitions

Norm

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

1. $||\mathbf{u}||_{\mathcal{H}} = 0$ if and only if $\mathbf{u} = 0$
2. $||\lambda \mathbf{u}||_{\mathcal{H}} = |\alpha| ||\mathbf{u}||_{\mathcal{H}}, \ \forall \ \lambda \in R, \ \forall \ \mathbf{u} \in \mathbb{R}$
3. $||\mathbf{u} + \mathbf{v}||_{\mathcal{H}} \leq ||\mathbf{u}||_{\mathcal{H}} + ||\mathbf{v}||_{\mathcal{H}}, \ \forall \ \mathbf{u}, \mathbf{v} \in \mathcal{H}$

Cauchy sequence

A sequence $\{\mathbf{x}_n\}_{n=1}^{\infty}$ of real number is said to be Cauchy sequence if for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that if $m, n \geq N$ $||\mathbf{x}_m - \mathbf{x}_n||_{\mathcal{H}} < \epsilon$

Banach space

Let $\mathcal{H}$ be a vector space equipped with a norm $||.||$. We say that $\mathcal{H}$ is Bananch space with respect to $||.||$ if every Cauchy sequence in $\mathcal{H}$ converged to a vector $\mathbf{u} \in \mathcal{H}$

Inner product space

Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle .,. \rangle_{\mathcal{H}} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said to be an inner product on $\mathcal{H}$ if:

1. Bilinearity

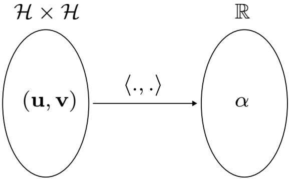
$$\langle \lambda_1 \mathbf{u} + \lambda_2 \mathbf{v}, \mathbf{w} \rangle_{\mathcal{H}} = \lambda_1 \langle \mathbf{u}, \mathbf{w} \rangle_{\mathcal{H}} + \lambda_2 \langle \mathbf{v}, \mathbf{w} \rangle_{\mathcal{H}}$$

2. Symmetry

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\mathcal{H}} = \langle \mathbf{v}, \mathbf{u} \rangle_{\mathcal{H}}$$

3. Positive semi-definiteness

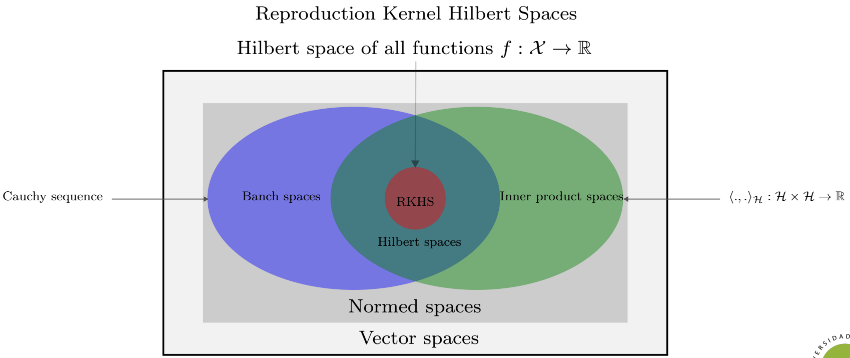
$\langle \mathbf{u}, \mathbf{u} \rangle_{\mathcal{H}} \geq 0$ and $\langle \mathbf{u}, \mathbf{u} \rangle_{\mathcal{H}} = 0$ if and only if $\mathbf{u} = 0$



Hilbert spaces

Hilbert space

Hilbert space is a complete inner product space, i.e., it is a Banach space with an inner product.

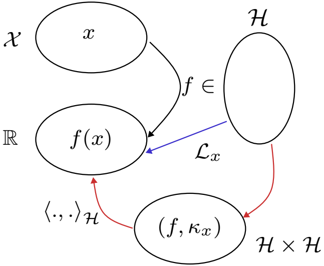


Reproduction Kernel Hilbert Spaces

Definition of RKHS I

Let $\mathcal{X}$ be a set and $\mathcal{H}$ a Hilbert space of all functions $f : \mathcal{X} \to \mathbb{R}$. For each element $x \in \mathcal{X}$, the evaluation functional is a linear functional that evaluates each $f \in H$ at the point x, written:

$$\mathcal{L}_x : \mathcal{H} \to \mathbb{R}, \text{ where } \mathcal{L}_x(f) = f(x) \text{ for all } f \in \mathcal{H}$$



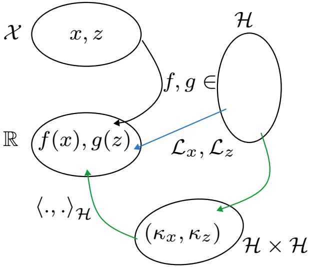
We say that $\mathcal{H}$ is a RKHS if, for all $x \in \mathcal{X}$, $\mathcal{L}_x$ is continuous at every $f \in \mathcal{H}$.

Corollario: Reproducing property

$$\mathcal{L}_x(f) = f(x) = \langle f, \kappa_x \rangle_{\mathcal{H}}$$

Definition of RKHS II

$\mathcal{L}_x(f) = f(x) = \langle f, \kappa_x \rangle_{\mathcal{H}}$ $\qquad$ $\mathcal{L}_z(g) = g(z) = \langle g, \kappa_z \rangle_{\mathcal{H}}$



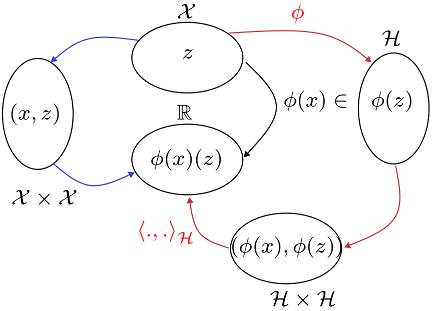
Kernel reproduction

Let $\mathcal{H}$ be a Hilbert space of $\mathcal{R}$-valued functions defined on a non-empty $\mathcal{X}$. A funtions $\kappa : \mathcal{H} \times \mathcal{H} \to \mathcal{R}$ is called Reproduction Kernel of $\mathcal{H}$ if it satisfies:

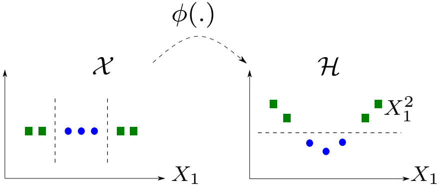
1. $\forall x \in \mathcal{X}$, $\kappa(., x) \in \mathcal{H}$
2. $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, \kappa(., x) \rangle_{\mathcal{H}} = f(x)$ (the reproducing property)

Kernel trick

$$\kappa(x, y) = \langle \phi(x), \phi(z) \rangle_{\mathcal{H}}$$



How to make comparisons?



Idea

$$\kappa(\mathbf{x}, \mathbf{c}) = (1 + \mathbf{x}^\top \mathbf{c})\ ; \mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2]\ , \mathbf{x} = [\mathbf{c}_1, \mathbf{c}_2]$$

$$\kappa(\mathbf{x}, \mathbf{c}) = \left(1 + \begin{pmatrix}\mathbf{x}_1 & \mathbf{x}_2\end{pmatrix}\begin{pmatrix}\mathbf{c}_1 \\ \mathbf{c}_2\end{pmatrix}\right)^2 = (1 + \mathbf{c}_1\mathbf{x}_1 + \mathbf{c}_2\mathbf{x}_2)^2$$

$$\kappa(\mathbf{x}, \mathbf{c}) = \mathbf{c}_1^2\mathbf{x}_1^2 + \mathbf{c}_2^2\mathbf{x}_2^2 + 2\mathbf{c}_1\mathbf{x}_1 + 2\mathbf{c}_2\mathbf{x}_2 + 2\mathbf{c}_1\mathbf{c}_2\mathbf{x}_1\mathbf{x}_2 + 1$$

So,

$$\phi(\mathbf{x}) = [1, \mathbf{x}_1^2, \sqrt{2}\mathbf{x}_1\mathbf{x}_2, \mathbf{x}_2^2, \sqrt{2}\mathbf{x}_1, \sqrt{2}\mathbf{x}_2]^\top$$

$$\phi(\mathbf{c}) = [1, \mathbf{c}_1^2, \sqrt{2}\mathbf{c}_1\mathbf{c}_2, \mathbf{c}_2^2, \sqrt{2}\mathbf{c}_1, \sqrt{2}\mathbf{c}_2]^\top$$

Kernel methods problem

Kernel trick

$$\kappa(\mathbf{x}, \mathbf{y}) =< \phi(\mathbf{x}), \phi(\mathbf{y}) > \quad (2)$$

Main idea

- Define a comparison function: $\mathbf{K} : \mathcal{X} \times \mathcal{X}$
- Represent a set of n data points $S = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ by the $n \times n$ matrix:

$$[\mathbf{K}]_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j) \quad (3)$$

Remarks

- K is always an $n \times n$ matrix, whatever the nature of data.
- Poor scalability with respect to the dataset size (n2 to compute and store K)

Random Fourier Features

Theorem I

This theorem expresses the power density in terms of the autocorrelation function

$$R(t) = \int_{-\infty}^{\infty} \overline{f(\tau)} f(t+\tau) d\tau \qquad\qquad f(\tau) = \int_{-\infty}^{\infty} f(t) e^{-jwt} dt$$

Wiener-khintchine

$$R(\tau) = \mathcal{F}\left[|f(w)|^2\right](t)$$

The autocorrelation is simply given by the Fourier transform of the absolute square of $f(t)$

Theorem II

Kosambi-Karhunen loeve

A stochastic process $X(t,w)$ defined in some interval T and in some probability space w, characterized by its mean, $u(t)$, and its covariance, $K(s,t) = (X(s) - u(t))(X(s)u(t))$ can be expressed through the expansion:

$$X(t,w) = u(t) + \sum_{j=1}^{\infty} \sqrt{\lambda}\phi_j(t)z_j(w) \tag{4}$$

Where λ_j and ϕ are Mercer eigenmodes (orthogonal functions) for K, and z_j are uncorrelated and of unit variance

Main idea

Basic idea

Let's go back to a lower-dimensional representation, using random Fourier features

Approximate the inner product $\kappa(\mathbf{x}, \mathbf{y}) =< \phi(\mathbf{x}), \phi(\mathbf{y}) >_v$ with a random mapping $z : R^D \to R^R$ where $D << R$

$$\kappa(\mathbf{x}, \mathbf{y}) =< \phi(\mathbf{x}), \phi(\mathbf{y}) >_v \approx z(\mathbf{x})^\top z(\mathbf{y}) \tag{5}$$

Where $z(.)$ is a good projection of ϕ

Translation invariant kernel

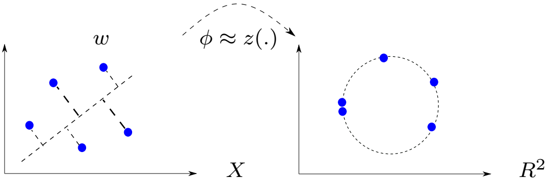
A Kernel $\kappa : \mathcal{H} \times \mathcal{H} \to \mathcal{R}$ is called translation invariant, if it only depends on the difference between its argument:

$$\forall_{\mathbf{x},\mathbf{y}} \in Z,\ \kappa(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x} - \mathbf{y})$$

For some $\varphi : \mathcal{H} \to \mathcal{R}$. Such a function φ is called positive definite if the corresponding kernel φ is positive definite.

Random Fourier Features

"Each component of the feature map $z(\boldsymbol{x})$ projects $\boldsymbol{x}$ onto a random direction w drawn from the Fourier transform"



$z(\mathbf{x}) = \cos(w^\top \mathbf{x} + b)$; $w \in R^R$, $w \in R$

Bochner theorem

A continuous function $\varphi : \mathcal{H} \to \mathcal{R}$ is positive definite if and only if it is the Fourier transform of a symmetric and positive finite measure $\mu \in M(R^d)$.

$$\varphi(\mathbf{x} - \mathbf{y}) = \underbrace{\int p(w) exp\left(jw\Delta\right) dw}_{} = E\left[\xi_w(\mathbf{x})\xi_w(\mathbf{y})^*\right] \tag{6}$$

Math review

$$\kappa(\mathbf{x},\mathbf{y}) = \kappa(\mathbf{x}-\mathbf{y}) = \int p(w)e^{jw(\mathbf{x}-\mathbf{y})}dw = E_w\left[\xi_w(\mathbf{x})\xi_w(\mathbf{y})^*\right] \tag{7}$$

where $\xi_w(\mathbf{x}) = e^{jw^\top \mathbf{x}}$

$$\kappa(\mathbf{x},\mathbf{y}) = \kappa(\mathbf{x}-\mathbf{y}) = E_w\left[e^{jw^\top \mathbf{x}}e^{-jw^\top \mathbf{y}}\right] = \int p(w)\left[\cos(w^\top(\mathbf{x}-\mathbf{y})) + j\sin(w^\top(\mathbf{x}-\mathbf{y}))\right]$$

$$\kappa(\mathbf{x},\mathbf{y}) = \kappa(\mathbf{x}-\mathbf{y}) = \int p(w)\cos(w^\top(\mathbf{x}-\mathbf{y})) + \underbrace{j\int p(w)\sin(w^\top(\mathbf{x}-\mathbf{y}))}_{0}$$

$$\hat{\kappa}(\mathbf{x},\mathbf{y}) = \hat{\kappa}(\mathbf{x}-\mathbf{y}) = E_w\left[z_w(\mathbf{x})z_w(\mathbf{y})\right]$$

$$z(\mathbf{x}) = \sqrt{\frac{2}{D}}\left[\cos(w_1^\top\mathbf{x}+b_1), \cos(w_2^\top\mathbf{x}+b_2), ..., \cos(w_D^\top\mathbf{x}+b_D)\right]$$

Dino Sejdinovic, Arthur Gretton
What is an RKHS?

Reproducing Kernel Hilbert Spaces Machine Learning
https://ngilshie.github.io/jekyll/update/2018/02/01/RKHS.html.

Rahimi, Ali and Recht, Benjamin
Random Features for Large-Scale Kernel Machines
Advances in neural information processing systems (2007).

Random Fourier Features
https://gregorygundersen.com/blog/2019/12/23/random-fourier-features/.

Thank you!