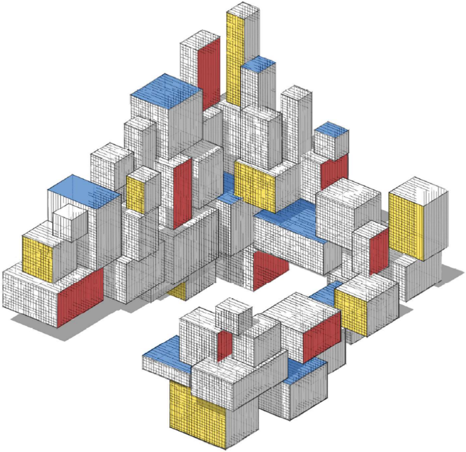
Constructing Kernels



Purpose

In this lecture we discuss:

- How to construct kernels from
  - characteristic functions
  - orthogonal features
  - other kernels
- Mercer's theorem

Kernels from Characteristic Functions

One way to construct reproducing kernels on $\mathcal{X} = \mathbb{R}^p$ makes use of the properties of *characteristic functions*.

Theorem: Reproducing Kernel from a Characteristic Function

Let $\boldsymbol{X} \sim \mu$ be an $\mathbb{R}^p$-valued random vector that is symmetric about the origin (that is, $\boldsymbol{X}$ and $-\boldsymbol{X}$ are identically distributed), and let ψ be its characteristic function:

$$\psi(\boldsymbol{t}) = \mathbb{E}\, \mathrm{e}^{\mathrm{i}\boldsymbol{t}^\top \boldsymbol{X}} = \int \mathrm{e}^{\mathrm{i}\boldsymbol{t}^\top \boldsymbol{x}}\, \mu(\mathrm{d}\boldsymbol{x}) \text{ for } \boldsymbol{t} \in \mathbb{R}^p.$$

Then $\kappa(\boldsymbol{x}, \boldsymbol{x}') := \psi(\boldsymbol{x} - \boldsymbol{x}')$ is a valid reproducing kernel on $\mathbb{R}^p$.

Gaussian Kernel

The multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $b^2\,\mathbf{I}_p$ is clearly symmetric around the origin. Its characteristic function is

$$\psi(\boldsymbol{t}) = \exp\left(-\frac{1}{2}b^2\,\|\boldsymbol{t}\|^2\right), \quad \boldsymbol{t} \in \mathbb{R}^p.$$

Taking $b^2 = 1/\sigma^2$, this gives the popular Gaussian kernel on $\mathbb{R}^p$:

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{1}{2}\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\sigma^2}\right). \tag{1}$$

The parameter σ is sometimes called the bandwidth.

Note that in the machine learning literature, the Gaussian kernel is sometimes referred to as "the" radial basis function (rbf) kernel.

Sinc Kernel

From the proof of the Moore–Aronszajn theorem, we see that the RKHS $\mathcal{G}$ determined by the Gaussian kernel κ is the space of pointwise limits of functions of the form

$$g(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \exp\left(-\frac{1}{2}\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{\sigma^2}\right).$$

We can think of each point $\boldsymbol{x}_i$ having a feature $\kappa_{\boldsymbol{x}_i}$ that is a scaled multivariate Gaussian pdf centered at $\boldsymbol{x}_i$.

The characteristic function of a $\mathcal{U}[-1,1]$ random variable (which is symmetric around 0) is

$$\psi(t) = \mathrm{sinc}(t) := \sin(t)/t,$$

so $\kappa(x, x') = \mathrm{sinc}(x - x')$ is a valid kernel.

Universal Approximation Property

One of the reasons why the Gaussian kernel (1) is popular is that it enjoys the universal approximation property: the space of functions spanned by the Gaussian kernel is dense in the space of continuous functions with support $\mathcal{X} \subset \mathbb{R}^p$.

However, note that *every* function g in the RKHS $\mathcal{G}$ associated with a Gaussian kernel κ is infinitely differentiable.

Moreover, a Gaussian RKHS does not contain non-zero constant functions.

Matérn Kernel

If it is known that g is differentiable only to a certain order, one may prefer the Matérn kernel with parameters $\nu, \sigma > 0$:

$$\kappa_\nu(\boldsymbol{x}, \boldsymbol{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\, \|\boldsymbol{x} - \boldsymbol{x}'\|/\sigma\right)^\nu K_\nu \left(\sqrt{2\nu}\, \|\boldsymbol{x} - \boldsymbol{x}'\|/\sigma\right), \qquad (2)$$

which gives functions that are differentiable to order $\lfloor \nu \rfloor$. Here, K_ν denotes the modified Bessel function of the second kind.

Can be defined through the characteristic function of the (radially symmetric) multivariate Student's t distribution with s degrees of freedom:

$$\psi(\boldsymbol{t}) = \frac{2^{1-s}}{\Gamma(s)} \|\boldsymbol{t}\|^{s-p/2} K_{p/2-s}(\|\boldsymbol{t}\|).$$

Reproducing Kernels Using Orthonormal Features

Suppose $\mathcal{X} \subseteq \mathbb{R}^p$ and let $L^2(\mathcal{X})$ be the set of square-integrable functions on $\mathcal{X}$.

Let $\{\xi_1, \xi_2, \ldots\}$ be an orthonormal basis of $L^2(\mathcal{X})$ and let $c_1, c_2, \ldots$ be a sequence of positive numbers.

We view each $c_i \xi_i =: \phi_i$ as a feature function and define

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') := \sum_{i \geqslant 1} \phi_i(\boldsymbol{x})\, \phi_i(\boldsymbol{x}') = \sum_{i \geqslant 1} \lambda_i\, \xi_i(\boldsymbol{x})\, \xi_i(\boldsymbol{x}'), \tag{3}$$

where $\lambda_i = c_i^2, i = 1, 2, \ldots$. This is well-defined as long as $\sum_{i \geqslant 1} \lambda_i < \infty$, which we assume from now on.

Linear Space $\mathcal{H}$

Let $\mathcal{H}$ be the linear space of functions of the form $f = \sum_{i \geqslant 1} \alpha_i \xi_i$, where $\sum_{i \geqslant 1} \alpha_i^2 / \lambda_i < \infty$.

$\mathcal{H}$ is a linear subspace of $L^2(\mathcal{X})$, as every function $f \in L^2(\mathcal{X})$ can be represented as $f = \sum_{i \geqslant 1} \langle f, \xi_i \rangle \xi_i$.

On $\mathcal{H}$ define the inner product

$$\langle f, g \rangle_{\mathcal{H}} := \sum_{i \geqslant 1} \frac{\langle f, \xi_i \rangle \langle g, \xi_i \rangle}{\lambda_i}.$$

With this inner product, the squared norm of $f = \sum_{i \geqslant 1} \alpha_i \, \xi_i$ is

$$\|f\|_{\mathcal{H}}^2 = \sum_{i \geqslant 1} \alpha_i^2 / \lambda_i < \infty.$$

RKHS $\mathcal{H}$

$\mathcal{H}$ is actually an RKHS with kernel κ, because:

- $$\kappa_{\boldsymbol{x}} = \sum_{i \geqslant 1} \lambda_i \, \xi_i(\boldsymbol{x}) \, \xi_i \qquad \in \mathcal{H},$$

  as $\sum_i \lambda_i < \infty$ by assumption, and so κ is finite.
- The reproducing property holds. Namely, let $f = \sum_{i \geqslant 1} \alpha_i \, \xi_i$. Then,

$$\langle \kappa_{\boldsymbol{x}}, f \rangle_{\mathcal{H}} = \sum_{i \geqslant 1} \frac{\langle \kappa_{\boldsymbol{x}}, \xi_i \rangle \langle f, \xi_i \rangle}{\lambda_i} = \sum_{i \geqslant 1} \frac{\lambda_i \, \xi_i(\boldsymbol{x}) \, \alpha_i}{\lambda_i} = \sum_{i \geqslant 1} \alpha_i \xi_i(\boldsymbol{x}) = f(\boldsymbol{x}).$$

Kernels via Orthogonal Features

Thus, kernels can be constructed via (3).

In fact, (under mild conditions) any given reproducing kernel κ can be written in the form (3). This result is known as Mercer's theorem.

The main idea is that a reproducing kernel κ can be thought of as a generalization of a positive semidefinite matrix $\mathbf{K}$, which can be written as $\mathbf{K} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{V}$ is a matrix of orthonormal eigenvectors $[\boldsymbol{v}_\ell]$ and $\mathbf{D}$ the diagonal matrix of the (positive) eigenvalues $[\lambda_\ell]$; that is,

$$\mathbf{K}(i, j) = \sum_{\ell \geqslant 1} \lambda_\ell \, v_\ell(i) \, v_\ell(j).$$

Theorem: Mercer

Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel for a compact set $\mathcal{X} \subset \mathbb{R}^p$. Then (under mild conditions) there exists a countable sequence of non-negative numbers $\{\lambda_\ell\}$ decreasing to zero and functions $\{\xi_\ell\}$ orthonormal in $L^2(\mathcal{X})$ such that

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = \sum_{\ell \geqslant 1} \lambda_\ell \, \xi_\ell(\boldsymbol{x}) \, \xi_\ell(\boldsymbol{x}'), \qquad \text{for all } \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}, \tag{4}$$

where (4) converges absolutely and uniformly on $\mathcal{X} \times \mathcal{X}$.

Further, if $\lambda_\ell > 0$, then (λ_ℓ, ξ_ℓ) is an (eigenvalue, eigenfunction) pair for the integral operator $K : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ defined by $[Kf](\boldsymbol{x}) := \int_{\mathcal{X}} \kappa(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}$ for $\boldsymbol{x} \in \mathcal{X}$.

In (4), $\boldsymbol{x}, \boldsymbol{x}'$ play the role of i, j, and ξ_ℓ plays the role of $\boldsymbol{v}_\ell$.

Example

Suppose $\mathcal{X} = [-1, 1]$ and the kernel is $\kappa(x, x') = 1 + xx'$ which corresponds to the RKHS $\mathcal{G}$ of affine functions from $\mathcal{X} \to \mathbb{R}$.

To find the (eigenvalue, eigenfunction) pairs for the integral operator appearing in Mercer's theorem, we need to find numbers $\{\lambda_\ell\}$ and orthonormal functions $\{\xi_\ell(x)\}$ that solve

$$\int_{-1}^{1} (1 + xx')\, \xi_\ell(x')\, \mathrm{d}x' = \lambda_\ell\, \xi_\ell(x)\,, \quad \text{for all } x \in [-1, 1].$$

Consider first a constant function $\xi_1(x) = c$. Then, for all $x \in [-1, 1]$, we have that $2c = \lambda_1 c$, and the normalization condition requires that $\int_{-1}^{1} c^2\, \mathrm{d}x = 1$. Together, these give $\lambda_1 = 2$ and $c = \pm 1/\sqrt{2}$.

Example (cont.)

Next, consider an affine function $\xi_2(x) = a + bx$. Orthogonality requires that

$$\int_{-1}^{1} c(a + bx)\,\mathrm{d}x = 0,$$

which implies $a = 0$ (since $c \neq 0$). Moreover, the normalization condition then requires

$$\int_{-1}^{1} b^2 x^2\,\mathrm{d}x = 1,$$

or, equivalently, $2b^2/3 = 1$, implying $b = \pm\sqrt{3/2}$. Finally, the integral equation reads

$$\int_{-1}^{1} (1 + xx')\,bx'\,\mathrm{d}x' = \lambda_2\,bx \iff \frac{2bx}{3} = \lambda_2 bx,$$

implying that $\lambda_2 = 2/3$.

Example (cont.)

We take the positive solutions (i.e., $c > 0$ and $b > 0$), and note that

$$\lambda_1 \, \xi_1(x) \, \xi_1(x') + \lambda_2 \, \xi_2(x) \, \xi_2(x') = 2 \frac{1}{\sqrt{2}} \frac{1}{\sqrt{2}} + \frac{2}{3} \frac{\sqrt{3}}{\sqrt{2}} x \frac{\sqrt{3}}{\sqrt{2}} x' = 1 + x x' = \kappa(x, x'),$$

and so we have found the decomposition appearing in (4).

Observe that ξ_1 and ξ_2 are orthonormal versions of the first two Legendre polynomials.

The corresponding feature map can be explicitly identified as $\phi_1(x) = \sqrt{\lambda_1} \, \xi_1(x) = 1$ and $\phi_2(x) = \sqrt{\lambda_2} \, \xi_2(x) = x$.

Theorem: Rules for Constructing Kernels from Kernels

1. If $\kappa : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a reproducing kernel and $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^p$ is a function, then $\kappa(\boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}'))$ is a reproducing kernel from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$.
2. If $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel and $f : \mathcal{X} \to \mathbb{R}_+$ is a function, then $f(\boldsymbol{x})\kappa(\boldsymbol{x}, \boldsymbol{x}')f(\boldsymbol{x}')$ is also a reproducing kernel from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$.
3. If κ_1 and κ_2 are reproducing kernels from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then so is their sum $\kappa_1 + \kappa_2$.
4. If κ_1 and κ_2 are reproducing kernels from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then so is their product $\kappa_1\kappa_2$.
5. If κ_1 and κ_2 are reproducing kernels from $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ respectively, then $\kappa_+((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}')) := \kappa_1(\boldsymbol{x}, \boldsymbol{x}') + \kappa_2(\boldsymbol{y}, \boldsymbol{y}')$ and $\kappa_\times((\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}')) := \kappa_1(\boldsymbol{x}, \boldsymbol{x}')\kappa_2(\boldsymbol{y}, \boldsymbol{y}')$ are reproducing kernels from $(\mathcal{X} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$.

Proof

For Rules 1, 2, and 3 it is easy to verify that the resulting function is finite, symmetric, and positive semidefinite, and so is a valid reproducing kernel, by Moore–Aronszajn. In particular, it holds true for $\boldsymbol{y}_i = \boldsymbol{\phi}(\boldsymbol{x}_i)$, $i = 1, \ldots, n$. Rule 4 is easy to show for kernels κ_1, κ_2 that admit a representation of the form (3), since

$$\begin{aligned}
\kappa_1(\boldsymbol{x}, \boldsymbol{x}')\,\kappa_2(\boldsymbol{x}, \boldsymbol{x}') &= \left(\sum_{i \geqslant 1} \phi_i^{(1)}(\boldsymbol{x})\,\phi_i^{(1)}(\boldsymbol{x}')\right)\left(\sum_{j \geqslant 1} \phi_j^{(2)}(\boldsymbol{x})\,\phi_j^{(2)}(\boldsymbol{x}')\right) \\
&= \sum_{i,j \geqslant 1} \phi_i^{(1)}(\boldsymbol{x})\,\phi_j^{(2)}(\boldsymbol{x})\,\phi_i^{(1)}(\boldsymbol{x}')\,\phi_j^{(2)}(\boldsymbol{x}') \\
&= \sum_{k \geqslant 1} \phi_k(\boldsymbol{x})\,\phi_k(\boldsymbol{x}') =: \kappa(\boldsymbol{x}, \boldsymbol{x}'),
\end{aligned}$$

showing that $\kappa = \kappa_1 \kappa_2$ also admits a representation of the form (3). The proof of 5 is left as an exercise.

Example: Polynomial Kernel

Consider $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^2$ with

$$\kappa(\boldsymbol{x}, \boldsymbol{x}') = (1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle)^2,$$

where $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle = \boldsymbol{x}^\top \boldsymbol{x}'$. This is an example of a polynomial kernel. By rules 3 and 4, we find that, since $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ and the constant function 1 are kernels, so are $1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$ and $(1 + \langle \boldsymbol{x}, \boldsymbol{x}' \rangle)^2$. By writing

$$\begin{aligned} \kappa(\boldsymbol{x}, \boldsymbol{x}') &= (1 + x_1 x_1' + x_2 x_2')^2 \\ &= 1 + 2x_1 x_1' + 2x_2 x_2' + 2x_1 x_2 x_1' x_2' + (x_1 x_1')^2 + (x_2 x_2')^2, \end{aligned}$$

we see that $\kappa(\boldsymbol{x}, \boldsymbol{x}')$ can be written as the inner product in $\mathbb{R}^6$ of the two feature vectors $\boldsymbol{\phi}(\boldsymbol{x})$ and $\boldsymbol{\phi}(\boldsymbol{x}')$, where the feature map $\boldsymbol{\phi} : \mathbb{R}^2 \to \mathbb{R}^6$ can be explicitly identified as

$$\boldsymbol{\phi}(\boldsymbol{x}) = [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1 x_2, x_1^2, x_2^2]^\top.$$

Thus, the RKHS determined by κ can be explicitly identified with the space of functions $\boldsymbol{x} \mapsto \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^6$.