Positive-definite kernel

From Wikipedia, the free encyclopedia

In operator theory, a branch of mathematics, a positive definite kernel is a generalization of a positive definite function or a positive-definite matrix. It was first introduced by James Mercer in the early 20th century, in the context of solving integral operator equations. Since then positive definite functions and their various analogues and generalizations have arisen in diverse parts of mathematics. They occur naturally in Fourier analysis, probability theory, operator theory, complex function-theory, moment problems, integral equations, boundary-value problems for partial differential equations, machine learning, embedding problem, information theory, and other areas.

This article will discuss some of the historical and current developments of the theory of positive definite kernels, starting with the general idea and properties before considering practical applications.

Contents

Definition

Let $\mathcal{X}$ be a nonempty set, sometimes referred to as the index set. A symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a positive definite (p.d.) kernel on $\mathcal{X}$ if

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) \geq 0 \qquad (1.1)$$

holds for any $n \in \mathbb{N}, x_1, \ldots, x_n \in \mathcal{X}, c_1, \ldots, c_n \in \mathbb{R}$.

In mathematical literature, kernels are usually complex valued functions, but in this article we assume real-valued functions, which is the common practice in machine learning and other applications of p.d. kernels.

Some general properties

- For a family of p.d. kernels $(K_i)_{i\in\mathbb{N}}, \quad K_i : \mathcal{X} \times \mathcal{X} \to R$
  - The sum $\sum_{i=1}^{n} \lambda_i K_i$ is p.d., given $\lambda_1, \ldots, \lambda_n \geq 0$
  - The product $K_1^{a_1} \ldots K_n^{a_n}$ is p.d., given $a_1, \ldots, a_n \in \mathbb{N}$
  - The limit $K = \lim_{n\to\infty} K_n$ is p.d. if the limit exists.
- If $(\mathcal{X}_i)_{i=1}^n$ is a sequence of sets, and $(K_i)_{i=1}^n, \quad K_i : \mathcal{X}_i \times \mathcal{X}_i \to R$ a sequence of p.d. kernels, then both

  $K((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \prod_{i=1}^{n} K_i(x_i, y_i)$ and

  $K((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \sum_{i=1}^{n} K_i(x_i, y_i)$

  are p.d. kernels on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$.
- Let $\mathcal{X}_0 \subset \mathcal{X}$. Then the restriction K_0 of K to $\mathcal{X}_0 \times \mathcal{X}_0$ is also a p.d. kernel.

Examples of p.d. kernels

- Common examples of p.d. kernels defined on Euclidean space $\mathbb{R}^d$ include:
  - Linear kernel: $K(x, y) = x^T y, x, y \in \mathbb{R}^d$.
  - Polynomial kernel: $K(x, y) = (x^T y + r)^n, x, y \in \mathbb{R}^d, r > 0$.
  - Gaussian kernel (RBF Kernel): $K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}, x, y \in \mathbb{R}^d, \sigma > 0$.
  - Laplacian kernel: $K(x, y) = e^{-\alpha\|x-y\|}, x, y \in \mathbb{R}^d, \alpha > 0$.
  - Abel kernel: $K(x, y) = e^{-\alpha|x-y|}, x, y \in \mathbb{R}, \alpha > 0$.
  - kernel generating Sobolev spaces $W_2^k(\mathbb{R}^d)$:

    $K(x, y) = \|x - y\|_2^{k-\frac{d}{2}} B_{k-\frac{d}{2}}(\|x - y\|_2)$, where B_ν is the Bessel function of third kind.
  - kernel generating Paley-Wiener space:
    $K(x, y) = \mathrm{sinc}(\alpha(x - y)), x, y \in \mathbb{R}, \alpha > 0$.

- If H is a Hilbert space, then its corresponding inner product $(\cdot,\cdot)_H : H \times H \to \mathbb{R}$ is a p.d. kernel. Indeed, we have

$$\sum_{i,j=1}^{n} c_i c_j (x_i, x_j)_H = \left(\sum_{i=1}^{n} c_i x_i, \sum_{j=1}^{n} c_j x_j \right)_H = \left\| \sum_{i=1}^{n} c_i x_i \right\|_H^2 \geq 0$$

- Kernels defined on $\mathbb{R}_+^d$ and histograms: Histograms are frequently encountered in applications of machine learning to real-life problems. Most observations are usually available under the form of nonnegative vectors of counts, which, if normalized, yield histograms of frequencies. It has been shown [1] that the following family of squared metrics, respectively Jensen divergence, the χ-square, Total Variation, and two variations of the Hellinger distance:

$$\psi_{JD} = H\left(\frac{\theta + \theta'}{2}\right) - \frac{H(\theta) + H(\theta')}{2},$$
$$\psi_{\chi^2} = \sum_i \frac{(\theta_i - \theta_i')^2}{\theta_i + \theta_i'}, \quad \psi_{TV} = \sum_i |\theta_i - \theta_i'|,$$
$$\psi_{H_1} = \sum_i |\sqrt{\theta_i} - \sqrt{\theta_i'}|, \psi_{H_2} = \sum_i |\sqrt{\theta_i} - \sqrt{\theta_i'}|^2,$$

can be used to define p.d. kernels using the following formula

$$K(\theta, \theta') = e^{-\alpha \psi(\theta, \theta')}, \alpha > 0.$$

History

PD kernels, as defined in (1.1), have arisen first in 1909 in a paper on integral equations by James Mercer.[2] Several other authors made use of this concept in the following two decades, but none of them explicitly used kernels $K(x, y) = f(x - y)$, i.e. p.d. functions (indeed M. Mathias and S. Bochner seem not to have been aware of the study of p.d. kernels). Mercer's work arose from Hilbert's paper of 1904 [3] on Fredholm integral equations of the second kind:

$$f(s) = \Phi(s) - \lambda \int_a^b K(s,t)\phi(t)dt. \qquad (1.2)$$

In particular, Hilbert had shown that

$$\int_a^b \int_a^b K(s,t)x(s)x(t)dsdt = \sum \frac{1}{\lambda_n} \left[\int_a^b \psi_n(s)x(s)ds \right]^2, \qquad (1.3)$$

where K is a continuous real symmetric kernel, x is continuous, $\{\psi_n\}$ is a complete system of orthonormal eigenfunctions, and λ_n's are the corresponding eigenvalues of (1.3). Hilbert defined a "definite" kernel as one for which the double integral $J(x) = \int_a^b \int_a^b K(s,t)x(s)x(t)dsdt$ satisfies $J(x) > 0$ except for $x(t) = 0$. The original object of Mercer's paper was to characterize the kernels which are definite in the sense of Hilbert, but Mercer soon found that the class of such functions was too restrictive to characterize in terms of determinants. He therefore defined a continuous real symmetric kernel $K(s,t)$ to be of positive type (i.e. positive definite) if $J(x) \geq 0$ for all real continuous functions x on $[a,b]$, and he proved that (1.2) is a necessary and sufficient condition for a kernel to be of positive type. Mercer then proved that for any continuous p.d. kernel the expansion

$$K(s,t) = \sum \frac{\psi_n(s)\psi_n(t)}{\lambda_n}$$

holds absolutely and uniformly.

At about the same time W. H. Young.,[4] motivated by a different question in the theory of integral equations, showed that for continuous kernels condition (1.1) is equivalent to $J(x) \geq 0$ for all $x \in L^1[a,b]$.

E.H. Moore [5][6] initiated the study of a very general kind of p.d. kernel. If E is an abstract set, he calls functions $K(x,y)$ defined on $E \times E$ "positive Hermitian matrices" if they satisfy (1.1) for all $x_i \in E$. Moore was interested in generalization of integral equations and showed that to each such K there is a Hilbert space H of functions such that, for each $f \in H, f(y) = (f, K(.,y))_H$. This property is called the reproducing property of the kernel and turns out to have importance in the solution of boundary-value problems for elliptic partial differential equations, and is the main reason for the success of kernel methods in machine learning. More details on this will be presented in the following section.

Another line of development in which p.d. kernels played a large role was the theory of harmonics on homogeneous spaces as begun by E. Cartan in 1929, and continued by H. Weyl and S. Ito. The most comprehensive theory of p.d. kernels in homogeneous spaces is that of M. Krein[7] which includes as special cases the work on p.d. functions and irreducible unitary representations of locally compact groups.

In probability theory p.d. kernels arise as covariance kernels of stochastic processes.[8]

Connection with reproducing kernel Hilbert spaces and feature maps

Positive definite kernels provide a framework that encompasses some basic Hilbert space constructions. In the following we present a tight relationship between positive definite

kernels and two mathematical objects, namely reproducing Hilbert spaces and feature maps.

Let X be a set, H a Hilbert space of functions $f : X \to R$, and $(\cdot,\cdot)_H : H \times H \to R$ the corresponding inner product on H. For any $x \in X$ the evaluation functional $e_x : H \to R$ is defined by $f \mapsto e_x(f) = f(x)$. We first define a reproducing kernel Hilbert space (RKHS):

> Definition: Space H is called a reproducing kernel Hilbert space if the evaluation functionals are continuous.

Every RKHS has a special function associated to it, namely the reproducing kernel:

> Definition: Reproducing kernel is a function $K : X \times X \to R$ such that
>
> 1) $K_x(\cdot) \in H, \forall x \in X$, and
> 2) $(f, K_x) = f(x)$, for all $f \in H$ and $x \in X$.
>
> The latter property is called the reproducing property.

The following result shows equivalence between RKHS and reproducing kernels:

> Theorem: Every reproducing kernel K induces a unique RKHS, and every RKHS has a unique reproducing kernel.

Now the connection between p.d. kernels and RKHS is given by the following theorem

> Theorem: Every reproducing kernel is positive definite, and every p.d. kernel defines a unique RKHS, of which it is the unique reproducing kernel.

Thus given a positive definite kernel K, it is possible to build an associated RKHS with K as a reproducing kernel.

As stated earlier, p.d. kernels can be constructed from inner products. This fact can be used to connect p.d. kernels with another interesting object that arises in machine learning applications, namely the feature map. Let F be a Hilbert space, and $(\cdot,\cdot)_F$ the corresponding inner product. Any map $\Phi : X \to F$ is called a feature map. In this case we call F the feature space. It is easy to see [9] that every feature map defines a unique p.d. kernel by

$$K(x, y) = (\Phi(x), \Phi(y))_F.$$

Indeed, positive definiteness of K follows from the p.d. property of the inner product. On the other hand, every p.d. kernel, and its corresponding RKHS, have many associated

feature maps. For example: Let $F = H$, and $\Phi(x) = K_x$ for all $x \in X$. Then $(\Phi(x), \Phi(y))_F = (K_x, K_y)_H = K(x, y)$, by the reproducing property. This suggests a new look at p.d. kernels as inner products in appropriate Hilbert spaces, or in other words p.d. kernels can be viewed as similarity maps which quantify effectively how similar two points x and y are through the value $K(x, y)$. Moreover, through the equivalence of p.d. kernels and its corresponding RKHS, every feature map can be used to construct a RKHS.

Kernels and distances

Kernel methods, which are very popular machine learning applications of p.d. kernels, are often compared to distance based methods such as nearest neighbors. In this section we discuss parallels between their two respective ingredients, namely kernels K and distances d.

Here by a distance function between each pair of elements of some set X, we mean a metric defined on that set, i.e. any nonnegative-valued function d on $\mathcal{X} \times \mathcal{X}$ which satisfies

- $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$,
- $d(x, y) = d(y, x)$,
- $d(x, z) \leq d(x, y) + d(y, z)$.

The link between distances and p.d. kernels is given by a particular kind of kernel, called negative definite kernel, and defined as follows

> Definition: A symmetric function $\psi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a negative definite (n.d.) kernel on $\mathcal{X}$ if
>
> $$\sum_{i,j=1}^{n} c_i c_j \psi(x_i, x_j) \leq 0 \qquad (1.4)$$
>
> holds for any $n \in \mathbb{N}, x_1, \ldots, x_n \in \mathcal{X}$, and $c_1, \ldots, c_n \in \mathbb{R}$ such that $\sum_{i=1}^{n} c_i = 0$.

The parallel between n.d. kernels and distances is in the following: whenever a n.d. kernel vanishes on the set $\{(x, x) : x \in \mathcal{X}\}$, and is zero only on this set, then its square root is a distance for $\mathcal{X}$.[10] At the same time each distance does not correspond necessarily to a n.d. kernel. This is only true for Hilbertian distances, where distance d is called Hilbertian if one can embed the metric space $(\mathcal{X}, d)$ isometrically into some Hilbert space.

On the other hand, n.d. kernels can be identified with a subfamily of p.d. kernels known as infinitely divisible kernels. A nonnegative-valued kernel K is said to be infinitely divisible if for every $n \in \mathbb{N}$ there exists a positive definite kernel K_n such that $K = (K_n)^n$.

Some applications

Kernels in machine learning

Positive definite kernels, through their equivalence with reproducing kernel Hilbert spaces, are particularly important in the field of statistical learning theory because of the celebrated representer theorem which states that every function in an RKHS can be written as a linear combination of the kernel function evaluated at the training points. This is a practically useful result as it effectively simplifies the empirical risk minimization problem from an infinite dimensional to a finite dimensional optimization problem.

Kernels in probabilistic models

There are several different ways in which kernels arise in probability theory.

- Nondeterministic recovery problems: Assume that we want to find the response $f(x)$ of an unknown model function f at a new point x of a set $\mathcal{X}$, provided that we have a sample of input-response pairs $(x_i, f_i) = (x_i, f(x_i))$ given by observation or experiment. The response f_i at x_i is not a fixed function of x_i but rather a realization of a real-valued random variable $Z(x_i)$. The goal is to get information about the function $E[Z(x_i)]$ which replaces f in the deterministic setting. For two elements $x, y \in \mathcal{X}$ the random variables $Z(x)$ and $Z(y)$ will not be uncorrelated, because if x is too close to y the random experiments described by $Z(x)$ and $Z(y)$ will often show similar behaviour. This is described by a covariance kernel $K(x, y) = E[Z(x) \cdot Z(y)]$. Such a kernel exists and is positive definite under weak additional assumptions. Now a good estimate for $Z(x)$ can be obtained by using kernel interpolation with the covariance kernel, ignoring the probabilistic background completely.

Assume now that a noise variable $\epsilon(x)$, with zero mean and variance σ^2, is added to x, such that the noise is independent for different x and independent of Z there, then the problem of finding a good estimate for f is identical to the above one, but with a modified kernel given by $K(x, y) = E[Z(x) \cdot Z(y)] + \sigma^2 \delta_{xy}$.

- Density estimation by kernels: The problem is to recover the density f of a multivariate distribution over a domain $\mathcal{X}$, from a large sample $x_1, \ldots, x_n \in \mathcal{X}$ including repetitions. Where sampling points lie dense, the true density function must take large values. A simple density estimate is possible by counting the number of samples in each cell of a grid, and plotting the resulting histogram, which yields a piecewise constant density estimate. A better estimate can be obtained by using a nonnegative translation invariant kernel K, with total integral equal to one, and define

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

as a smooth estimate.

Numerical solution of partial differential equations

One of the greatest application areas of so-called meshfree methods is in the numerical solution of PDEs. Some of the popular meshfree methods are closely related to positive definite kernels (such as meshless local Petrov Galerkin (MLPG), Reproducing kernel particle method (RKPM) and smoothed-particle hydrodynamics (SPH)). These methods use radial basis kernel for collocation[11]

Stinespring dilation theorem

Other applications

In the literature on computer experiments [12] and other engineering experiments one increasingly encounters models based on p.d. kernels, RBFs or kriging. One such topic is response surface modeling. Other types of applications that boil down to data fitting are rapid prototyping and computer graphics. Here one often uses implicit surface models to approximate or interpolate point cloud data.

Applications of p.d. kernels in various other branches of mathematics are in multivariate integration, multivariate optimization, and in numerical analysis and scientific computing, where one studies fast, accurate and adaptive algorithms ideally implemented in high-performance computing environments.[13]

See also

- Integral equation
- Integral transform
- Positive definite function on a group
- Reproducing kernel Hilbert space
- Kernel method

References

1. Hein, M. and Bousquet, O. (2005). "Hilbertian metrics and positive definite kernels on probability measures". In Ghahramani, Z. and Cowell, R., editors, Proceedings of AISTATS 2005.
2. Mercer, J. (1909). “Functions of positive and negative type and their connection with the theory of integral equations”. Philosophical Transactions of the Royal Society of London, Series A 209, pp. 415-446.
3. Hilbert, D. (1904). "Grundzuge einer allgemeinen Theorie der linearen Integralgleichungen I", Gott. Nachrichten, math.-phys. K1 (1904), pp. 49-91.

4. Young, W. H. (1909). "A note on a class of symmetric functions and on a theorem required in the theory of integral equations", Philos. Trans. Roy.Soc. London, Ser. A, 209, pp. 415-446.
5. Moore, E.H. (1916). "On properly positive Hermitian matrices", Bull. Amer. Math. Soc. 23, 59, pp. 66-67.
6. Moore, E.H. (1935). "General Analysis, Part I", Memoirs Amer. Philos. Soc. 1, Philadelphia.
7. Krein. M (1949/1950). "Hermitian-positive kernels on homogeneous spaces I and II" (in Russian), Ukrain. Mat. Z. 1(1949), pp. 64-98, and 2(1950), pp. 10-59. English translation: Amer. Math. Soc. Translations Ser. 2, 34 (1963), pp. 69-164.
8. Loeve, M. (1960). "Probability theory", 2nd ed., Van Nostrand, Princeton, N.J.
9. Rosasco, L. and Poggio, T. (2015). "A Regularization Tour of Machine Learning - MIT 9.520 Lecture Notes" Manuscript.
10. Berg, C., Christensen, J. P. R., and Ressel, P. (1984). "Harmonic Analysis on Semigroups". Number 100 in Graduate Texts in Mathematics, Springer Verlag.
11. Schabak, R. and Wendland, H. (2006). "Kernel Techniques: From Machine Learning to Meshless Methods", Cambridge University Press, Acta Numerica (2006), pp. 1-97.
12. Haaland, B. and Qian, P. Z. G. (2010). "Accurate emulators for large-scale computer experiments", Ann. Stat.
13. Gumerov, N. A. and Duraiswami, R. (2007). "Fast radial basis function interpolation via preconditioned Krylov iteration". SIAM J. Scient. Computing 29/5, pp. 1876-1899.

Categories: Operator theory | Hilbert space
