Lecture 9: On PDS Kernels

While RKHS methods, via the application of rep thms, are widely applicable in ML. The quality of the resulting algorithms is highly contingent upon the choice of the Kernel, including its hyper parameters.

In this lecture we will focus our attention on PDS Kernels, their properties, & recipes for constructing them.

9.1 Kernel Calculus

Given a family of PDS Kernels $K_i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ it is natural to ask what operations among these kernels will define another PDS Kernel.

Such results are crucial in enabling the design of PDS Kernels for complex ML tasks

• (Linear Combinations) Given PDS Kernels K_1, K_2, K_3 & $\alpha > 0$ it holds that

αK_1 & $K_2 + K_3$ are PDS Kernels.

• (Mappings) Let $\mathcal{X}$ & $\mathcal{Y}$ be sets & let $A : \mathcal{X} \to \mathcal{Y}$ & suppose $K : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a PDS kernel. Then $K(A(x), A(x'))$ is a PDS Kernel on $\mathcal{X}$!

• (Tensor Products) Given PDS Kernels K_1 on $\mathcal{X}_1$ & K_2 on $\mathcal{X}_2$, then $K_1 \otimes K_2$

i.e, $K_1 \otimes K_2((x_1, x_2), (x_1', x_2')) = K_1(x_1, x_1') K_2(x_2, x_2')$

is a PDS kernel on $\mathcal{X}_1 \times \mathcal{X}_2$.

Thm (PDS Kernel Closure Properties)
PDS Kernels are closed under sum, product, tensor product, point wise limit & composition with power series.

Proof (i) sums: Consider K, K' & kernel matrices $\Theta = K(X,X)$ & $\Theta' = K'(X,X)$ we have

$$\underline{c}^T(\Theta + \Theta')\underline{c} = \underline{c}^T\Theta\underline{c} + \underline{c}^T\Theta'\underline{c} \geq 0.$$

The sum is also clearly sym.

(ii) Products: since Θ is PDS, we can write

$\Theta = MM^T$ (e.g. Cholesky factorization)

Now let $\tilde{\Theta}$ be the matrix associated with

$\tilde{K}(x,x') = K(x,x')K'(x,x')$ ie, $\tilde{\Theta}_{ij} = \Theta_{ij}\,\Theta'_{ij}$

Then

$$\sum_{i,j} c_i c_j \tilde{\Theta}_{ij} = \sum_{ij} c_i c_j \Theta_{ij}\,\Theta'_{ij}$$

$$= \sum_{ijk} c_i c_j M_{ik} M_{jk} \Theta'_{ij}$$

$$= \sum_{ijk} c_i M_{ik}\, c_j M_{jk}\, \Theta'_{ij}$$

$$= \sum_k \underline{z}_k^T \Theta' \underline{z}_k \geq 0, \quad \underline{z}_k := \begin{bmatrix} c_1 M_{1k} \\ c_2 M_{2k} \\ \vdots \\ c_n M_{nk} \end{bmatrix}$$

$\mathcal{X}_1 \times \mathcal{X}_2$

(iii) Tensor product: Given $(x_1, x_2), (x_1', x_2') \in$ & PDS kernels K_1 on $\mathcal{X}_1$ & K_2 on $\mathcal{X}_2$ we can directly verify that

$$((x_1, x_2), (x_1', x_2')) \longmapsto K_1(x_1, x_1')$$

$$((x_1, x_2), (x_1', x_2')) \longmapsto K_2(x_2, x_2')$$

are PDS kernels on $\mathcal{X}_1 \times \mathcal{X}_2$ & $K_1 \otimes K_2$ is simply the product of these kernels by (ii).

(iv) Pointwise limit: Let $K_n \xrightarrow{n \to \infty} K$ pointwise & K_n are PDS. Let Θ be the kernel matrix of K & Θ_n those of K_n. Then $\Theta = \lim_{n \to \infty} \Theta_n \succeq 0$

(v) Power series: Let $f : x \longmapsto \sum_{j=1}^{\infty} a_j x^j$ with $a_j \geq 0$ & with radius of convergence $r > 0$. Suppose K is a PDS kernel & $|K(x,x')| \leq r$ $\forall x, x' \in \mathcal{X}$. Then for any $j \in \mathbb{N}$ we have that $a_j K^j$ is PDS by (ii). Furthermore, $\sum_{j=1}^{n} a_j K^j$ is PDS by (i). Finally, $f \circ K$ is PDS by (iv).

Using the above calculus we can easily design or verify PDS Kernels.

e.g. We showed directly that $K(\underline{x},\underline{x}') = \underline{x}^T\underline{x}'$ is PDS. Then using the Binomial formula we can check that $K(\underline{x},\underline{x}') = (\underline{x}^T\underline{x}' + c)^\alpha$, $\alpha \in \mathbb{N}$ & $c \in \mathbb{R}$ is PDS.

e.g. By the Taylor expansion of the exponential function we can verify that the exponential kernel $K(\underline{x},\underline{x}') = \exp(\underline{x}^T\underline{x}')$ is PDS.

9.2 Mercer's Theorem

PDS Kernels have alot of useful properties, some of which we will use extensively in the next lecture. Among these & one of the most useful ones, is Mercer's theorem which allows us to think of Kernels like extensions of Matrices.

some quick setup. Given a Banach space $\mathcal{X}$ & a finite Borel measure μ on $\mathcal{X}$ we define

$$L^2(\mathcal{X},\mu) := \left\{ f : \mathcal{X} \to \mathbb{R} \,\middle|\, \int_{\mathcal{X}} |f(x)|^2 d\mu(x) < +\infty \right\}$$

This is a generalization of the usual L^2 space to more general topological spaces. The resulting space is a Hilbert space very much as before

$$\langle f, g \rangle_{L^2(\mathcal{X},\mu)} = \int_{\mathcal{X}} f(x) g(x) \, d\mu(x)$$

(See Bogachev "Measure Theory Vol I", Ch4).

Thm (Mercer's)

Let $\mathcal{X}$ be a Banach space & μ a finite Borel measure supported on $\mathcal{X}$ (ie, $\mu(\mathcal{X}) = 1$) Suppose K is a continuous PDS kernel on $\mathcal{X}$. Then there exists an orthonormal set $\{\psi_j\}_{j=1}^{\infty}$ in $L^2(\mathcal{X},\mu)$ consisting of the eigenfunctions of the integral operator

$$T_K f := \int_{\mathcal{X}} K(\cdot, x) f(x) \, d\mu(x)$$

such that the corresponding seq. of eigenvalus of T_K, $\{\lambda_j\}_{j=1}^{\infty}$ are non-neg. Furthermore, the eigenfunctions ψ_j that correspond to the non-zero λ_j can be taken to be continuous & it holds that

$$K(x,x') = \sum_{j=1}^{\infty} \lambda_j \psi_j(x) \psi_j(x') \quad \forall x,x'$$

The most useful way of thinking about Mercer's thm is in terms of symm. pos. def. matrices & their eigen. decomp.

If $K \in \mathbb{R}^n \times \mathbb{R}^n$ is sym. pos. def. then we can write
$$K = \Psi \Lambda \Psi^T = \sum_{j=1}^{n} \lambda_j \underline{\psi}_j \underline{\psi}_j^T$$
where Ψ is orthonormal & $\lambda_j \geq 0$.