Lecture 18: FoSML

- Given a RKHS $\mathcal{H}$ associated to a kernel K, we have $h(x) = \langle h, K(x,\cdot)\rangle$.

This commonly takes the form of an integral operator:

$$h(x) = \langle h, K(x,\cdot)\rangle_{L^2}$$

$$= \int_{\mathbb{R}} h(y) K(x,y)\, dy.$$

- Given $\mathcal{H}, K$, we say $\Phi: \mathcal{X} \to \mathcal{H}$ is a feature mapping if

$$K(x,x') = \langle \Phi(x), \Phi(x')\rangle_{\mathcal{H}}, \quad \forall x, x'.$$

In this case, we consider $\mathcal{H}$ as a feature space:

$$\begin{array}{ccc} \mathcal{H} & & \langle \Phi(x), \Phi(x')\rangle_{\mathcal{H}} \\ \uparrow \Phi & & \| \\ \mathcal{X} & & K(x,x') \end{array}$$

Note that $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ induces a norm $\|h\|_{\mathcal{H}} = |\langle h, h\rangle|^{1/2}$.

- To any kernel K, we may normalize K as

$$\tilde{K}(x,x') = \begin{cases} 0 \text{ if } K(x,x) = 0 \text{ or } K(x',x') = 0 \\ \dfrac{K(x,x')}{\sqrt{K(x,x)\cdot K(x',x')}}, \text{ else} \end{cases}$$

- This has the benefit of forcing $\tilde{K}(x,x) = 1 \;\; \forall x$ s.t. $K(x,x) \neq 0$.

ex: Let $K(x,x') = \exp(\langle x, x'\rangle_{\mathbb{R}^p}/\sigma^2)$. Then if we normalize, we acquire the Gaussian kernel:

$$\frac{K(x,x')}{\sqrt{K(x,x)\cdot K(x',x')}} = \frac{\exp(\langle x,x'\rangle/\sigma^2)}{\sqrt{\exp(\langle x,x\rangle/\sigma^2)\exp(\langle x',x'\rangle/\sigma^2)}}$$

$$= \frac{\exp(\langle x,x'\rangle/\sigma^2)}{\sqrt{\exp((\|x\|^2+\|x'\|^2)/\sigma^2)}}$$

$$= \frac{\exp(\langle x,x'\rangle/\sigma^2)}{\exp\left(\frac{\|x\|^2+\|x'\|^2}{2\sigma^2}\right)}$$

$$= \exp\left(\frac{1}{2\sigma^2}\left(2\langle x,x'\rangle - \|x\|^2 - \|x'\|^2\right)\right)$$

$$= \exp\left(\frac{1}{2\sigma^2}\left(-\|x-x'\|_2^2\right)\right).$$

- Normalized PDS kernels inherit the PDS property:

Lemma: Let K be a PDS kernel. Then its normalization $\tilde{K}$ is also PDS.

Proof: Let $\{x_i\}_{i=1}^m \subseteq X$ and $c \in \mathbb{R}^m$ be arbitrary. It suffices to show

$$\text{(★)} \quad \sum_{i,j=1}^m c_i c_j K(x_i,x_j) \geq 0.$$

By Cauchy-Schwartz, if $K(x_i,x_i)=0$, then $K(x_i,x_j)=0 \ \forall j \in \{1,\dots,m\}$. So, WLOG, $K(x_i,x_i)>0 \ \forall i \in \{1,\dots,m\}$, since otherwise these terms sum to 0 in (★). We now compute:

$$\sum_{i,j=1}^m \frac{c_i c_j K(x_i,x_j)}{\sqrt{K(x_i,x_i)K(x_j,x_j)}}$$

$$= \sum_{i,j=1}^{m} \frac{c_i c_j \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}}{\|\Phi(x_i)\|_{\mathcal{H}} \|\Phi(x_j)\|_{\mathcal{H}}}$$

$$= \left\| \sum_{i=1}^{m} \frac{c_i \Phi(x_i)}{\|\Phi(x_i)\|_{\mathcal{H}}} \right\|_{\mathcal{H}}^2$$

≥ 0,

where the map $\Phi: \mathcal{X} \to \mathcal{H}$ s.t. $\langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}} = k(x_i, x_j)$ exists by our theorem on RKHS. ∎

- Given a sample $\{x_i\}_{i=1}^m \subset \mathcal{X}$, one can consider an empirical kernel map

$$\Phi: \mathcal{X} \to \mathbb{R}^m$$
$$x \mapsto \{k(x, x_i)\}_{i=1}^m$$

- Letting K be the Gram matrix associated to k and $\{x_i\}_{i=1}^m$, i.e. $K_{ij} = k(x_i, x_j)$, we may note that $\Phi(x_i) = Ke_i$, where $e_i \in \mathbb{R}^n$ is the i^{th} canonical basis vector. In particular,

$$\langle \Phi(x_i), \Phi(x_j) \rangle = e_i^T K^2 e_j.$$

- Thus, the kernel matrix K' associated to Φ is K^2.
- Motivated by this, let $(K^\dagger)^{1/2}$ be the square root of the pseudo-inverse of K.

Then setting $\Psi(x) = (K^\dagger)^{1/2} \Phi(x)$, we have that

$$\langle \Psi(x_i), \Psi(x_j) \rangle = \left((K^\dagger)^{1/2} K e_i\right)^T \left((K^\dagger)^{\frac{1}{2}} K e_j\right)$$
$$= e_i K^T \left((K^\dagger)^{1/2}\right)^T (K^\dagger)^{\frac{1}{2}} K e_j$$
$$= e_i K^T (K^\dagger)^{\frac{1}{2}} (K^\dagger)^{\frac{1}{2}} K e_j$$

$$= e_i K^T \underbrace{K^\dagger K}_{\text{"}I} e_j$$

$$= e_i K^T e_j$$

$$= e_i K e_j ,$$

so that Ψ is associated to K.

- Similarly $\Omega : \mathcal{X} \to \mathbb{R}^m$
$x \mapsto K^\dagger \Phi(x)$

is ~~was~~ associated to the identity, assuming K is invertible, otherwise to $KK^\dagger$.

- We remark the the property of PDS is preserved (closed) under a range of algebraic operations, summarized as follows.

Theorem (Closure properties of PDS kernels): PDS kernels are closed under sums, products, tensor product, pointwise limit, and composition with $x \mapsto \sum_{n=0}^{\infty} a_n x^n$ for any coefficients $a_n \geq 0$ and $n \in \mathbb{Z}_{\geq 0}$.

Proof: Let K, K' be two $m \times m$ Gram matrices generated from an arbitrary set of m points and ~~PDS~~ PDS kernels k, k'. Since k, k' are PDS, K, K' are SPSD. We aim to show the same for the various ways of modifying K, K'. Let $c \in \mathbb{R}^m$ be arbitrary.

Sum: $c^T[K+K']c = c^T Kc + c^T K' c \geq 0$ by K, K' SPSD.

Product: Let K be SPSD. By the SVD, $\exists M$ s.t. $K = MM^T$. Then

$$\sum_{i,j=1}^{m} c_i c_j (K_{ij} K'_{ij})$$

$$= \sum_{i,j=1}^{m} c_i c_j \left(\sum_{k=1}^{m} M_{ik} M_{kj} \right) K'_{ij}$$

$$= \sum_{k=1}^{m} Z_k^T K' Z \geq 0, \quad Z_k = (c_1 M_{1k}, \dots, c_m M_{mk})^T.$$

Tensor Product: Let $K \otimes K'(x_1, x_2, x_1', x_2') = K(x_1, x_2) \cdot K'(x_1', x_2')$.

Then we can realize this as the product of the PSD kernels

the Gram matrix associated to

$\tilde{K}_1(x_1, x_2, x_1', x_2') = K(x_1, x_2)$

$\tilde{K}_2(x_1, x_2, x_1', x_2') = K'(x_1', x_2')$, so the result follows from our result on pointwise products.

Limits: Obvious from continuity of matrix multiplication w.r.t. matrix limit.

Power series composition: Follows from results on pointwise products, sums, and limits. ∎

Remark: For any PDS kernel K, $\exp(K)$ is PDS by the result on power series.