Lecture 19: Properties of kernel mean embeddings & MMD

19.1 Properties of kernel mean embeddings

Last lecture we introduced the idea of kernel mean embeddings of prob. meas

$\mu \in \mathbb{P}(\mathcal{X})$ $\qquad \mu_K := \int K(x,\cdot)\,\mu(dx)$

& in turn the MMD: $\qquad = \mathbb{E}_{x\sim\mu} K(x,\cdot)$

$$MMD(\mu,\mu') := \|\mu_K - \mu'_K\|_K$$

In this lecture we will take a closer look at the properties of kernel mean embeddings & the MMD. Starting with the following obs.

Observation In HW3 you showed for any $\phi \in \mathcal{H}_K^*$ that $\phi(f) = \langle f, K\phi\rangle \ \forall f \in \mathcal{H}_K$ where $K\phi := x \mapsto \phi(K(x,\cdot))$. Taking $\phi(f) := \int f(x)\mu(dx)$ we infer that $K\phi = \mu_K$!

Thm: If $\int \sqrt{K(x,x)}\,\mu(dx) < +\infty$ then $\mu_K \in \mathcal{H}_K$

& $\int f(x)\mu(dx) \equiv \mathbb{E}_{x\sim\mu} f(x) = \langle f, \mu_K \rangle_K \;\forall f \in \mathcal{H}_K$

Proof: Based on above observation we only need to check that $f \longmapsto \int f(x)\mu(dx)$ is a bdd lin. op. on $\mathcal{H}$.

$$\left|\int f(x)\,\mu(dx)\right| \leq \int |f(x)|\,\mu(dx) \quad \text{(Jensen's ineq.)}$$
$$= \int |\langle f, K(x,\cdot)\rangle_K|\,\mu(dx)$$
$$\leq \int \sqrt{K(x,x)}\,\|f\|_K\,\mu(dx) \quad \text{(C-S ineq.)}$$

Kernel mean embeddings are deeply connected to well-known objects in probability theory.

eg: (moment generating func.) Consider the kernel $K(\underline{x},\underline{x}') = \exp(\underline{x}^T\underline{x}')$ then we have

$$\mu_K(\underline{x}') = \mathbb{E}_{\underline{x}\sim\mu} \exp(\underline{x}^T\underline{x}')$$

which is the multi-variate moment generating function (when it exists).

eg: (Characteristic function) Consider the Fourier kernel $K(\underline{x},\underline{x}') = \exp(i\underline{x}^T\underline{x}')$ then

$$\mu_K(\underline{x}') = \int \exp(i\underline{x}^T\underline{x}')\,\mu(d\underline{x})$$

which is precisely the Characteristic func. or Fourier transform of μ!

More generally, by Mercer's thm

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

& so,

$$\mu_K = \mathbb{E}_{\underline{x}\sim\mu} K(\underline{x},\cdot) = \mathbb{E}_{\underline{x}\sim\mu} \sum_{j=1}^{\infty} \lambda_j \psi_j(\underline{x})\psi_j(\cdot)$$

$$= \sum_{j=1}^{\infty} \left(\lambda_j\, \mathbb{E}_{\underline{x}\sim\mu} \psi_j(\underline{x})\right) \psi_j(\cdot)$$

Since ψ_j are orth. normal in $L^2(\mathcal{X},\nu)$ then

$$\|\mu_K\|_K^2 = \sum_{j=1}^{\infty} \left(\lambda_j\, \mathbb{E}_{\underline{x}\sim\mu} \psi_j(\underline{x})\right)^2$$

so, the RKHS norm of μ_K is a weighted sum of generalized moments of μ!

19.2 Properties of MMD & its empirical Approx.

Recall that earlier we mentioned that MMD is not necessarily a metric!

Let $\mathbb{P}_K(\mathcal{X}) = \{\mu \in \mathbb{P}(\mathcal{X}) \mid \int \sqrt{K(x,x)}\,\mu(dx) < +\infty\}$.

Now let $\mu, \mu', \mu'' \in \mathbb{P}_K(\mathcal{X})$ then

(Sym) 1) $MMD(\mu,\mu') = \|\mu_K - \mu'_K\|_K = MMD(\mu',\mu)$

(Pos.) 2) $MMD(\mu,\mu') \geq 0$

(Δ-ineq) 3) $MMD(\mu,\mu') = \|\mu_K - \mu'_K\|_K \leq \|\mu - \mu''\|_K + \|\mu'' - \mu'\|_K$

$= MMD(\mu,\mu'') + MMD(\mu'',\mu)$

4) $MMD(\mu,\mu) = 0$ but we may also have $MMD(\mu,\mu') = 0$ while $\mu \neq \mu'$! ie, the divergence property may fail!

Given a distance-like function $D : \mathbb{P}(\mathcal{X}) \times \mathbb{P}(\mathcal{X}) \to \mathbb{R}$ we say D has the divergence property if

$$D(\mu,\mu') = 0 \text{ iff } \mu = \mu'$$

See Birrell et al. "(f,Γ)-divergences: interpolating between f-divergences & integral probability metrics.

So in general MMDs may not satisfy the div. property which means MMD may not be a metric (hence the phrase "discrepancy"). This essentially comes down to our choice of the kernel.

Def^n: We say a kernel $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (or $\mathbb{C}$) is Characteristic if the map

$$\mathbb{P}(\mathcal{X}) \ni \mu \longmapsto \mu_K \in \mathcal{H}_K$$

is injective. In that case we may also say $\mathcal{H}_K$ is characteristic.

Intuitively, a kernel / RKHS is charateristic if $\mathcal{H}_K$ is sufficiently rich/large so that its elements can capture higher moments of prob. mesures in $\mathbb{P}(\mathcal{X})$.

The best eg. of a characteristic kernel is the Fourier kernel $K(\underline{x}, \underline{x}') = \exp(i \underline{x}^T \underline{x}')$ since its mean embedding is simply the Fourier transform.

Table 3.1: Various characterizations of well-known kernel functions. The columns marked 'U', 'C', 'TI', and 'SPD' indicate whether the kernels are universal, characteristic, translation-invariant, and strictly positive definite, respectively, w.r.t. the domain $\mathcal{X}$. For the discrete kernel, $\#_s(\mathbf{x})$ is the number of times substrings s occurs in a string $\mathbf{x}$. K_ν is the modified Bessel function of the second kind of order ν and Γ is the Gamma function.

Kernel Function	$k(\mathbf{x}, \mathbf{y})$	**Domain** $\mathcal{X}$	**U**	**C**	**TI**	**SPD**
Dirac	$\mathbb{1}_{\mathbf{x}=\mathbf{y}}$	$\{1, 2, \ldots, m\}$	✓	✓	✗	✓
Discrete	$\sum_{s\in\mathcal{X}} w_s \#_s(\mathbf{x})\#_s(\mathbf{y})$ with $w_s > 0$ for all s	$\{s_1, s_2, \ldots, s_m\}$	✓	✓	✗	✓
Linear	$\langle \mathbf{x}, \mathbf{y}\rangle$	$\mathbb{R}^d$	✗	✗	✗	✗
Polynomial	$(\langle \mathbf{x}, \mathbf{y}\rangle + c)^p$	$\mathbb{R}^d$	✗	✗	✗	✗
Gaussian	$\exp(-\sigma\|\mathbf{x}-\mathbf{y}\|_2^2)$, $\sigma > 0$	$\mathbb{R}^d$	✓	✓	✓	✓
Laplacian	$\exp(-\sigma\|\mathbf{x}-\mathbf{y}\|_1)$, $\sigma > 0$	$\mathbb{R}^d$	✓	✓	✓	✓
Rational quadratic	$(\|\mathbf{x}-\mathbf{y}\|_2^2 + c^2)^{-\beta}$, $\beta > 0, c > 0$	$\mathbb{R}^d$	✓	✓	✓	✓
B_{2l+1}-splines	$B_{2l+1}(\mathbf{x}-\mathbf{y})$ where $l \in \mathbb{N}$ with $B_i := B_i \otimes B_0$	$[-1, 1]$	✓	✓	✓	✓
Exponential	$\exp(\sigma\langle \mathbf{x}, \mathbf{y}\rangle)$, $\sigma > 0$	compact sets of $\mathbb{R}^d$	✗	✓	✗	✓
Matérn	$\frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{x}'\|_2}{\sigma}\right) K_\nu\left(\frac{\sqrt{2\nu}\|\mathbf{x}-\mathbf{x}'\|_2}{\sigma}\right)$	$\mathbb{R}^d$	✓	✓	✓	✓
Poisson	$1/(1 - 2\alpha\cos(\mathbf{x}-\mathbf{y}) + \alpha^2)$, $0 < \alpha < 1$	$([0, 2\pi), +)$	✓	✓	✓	✓

(Table from Muandet et al (2017)