Mathematisches Forschungsinstitut Oberwolfach

Report No. 30/2008

Learning Theory and Approximation

Organised by
Kurt Jetter, Hohenheim
Steve Smale, Berkeley
Ding-Xuan Zhou, Hong Kong

June 29th – July 5th, 2008

Mathematics Subject Classification (2000): AMS-CLASSIFICATION.

Workshop: Learning Theory and Approximation

Table of Contents

Abstracts

RKHS representation of measures applied to homogeneity, independence, and Fourier optics

Bernhard Schölkopf, Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu

A symmetric function $k : \mathcal{X}^2 \to \mathbb{R}$, where $\mathcal{X}$ is a nonempty set, is called a positive definite (pd) kernel if for arbitrary points $x_1, \ldots, x_m \in \mathcal{X}$ and coefficients $a_1, \ldots, a_m \in \mathbb{R}$, we have

$$\sum_{i,j} a_i a_j k(x_i, x_j) \geq 0.$$

The kernel is called strictly positive definite if for pairwise distinct points, the implication $\sum_{i,j} a_i a_j k(x_i, x_j) = 0 \Longrightarrow \forall i : a_i = 0$ is valid.

Any positive definite kernel induces a mapping

$$x \mapsto k(x, .)$$

into a reproducing kernel Hilbert space (RKHS) satisfying

$$\langle k(x, .), k(x', .)\rangle = k(x, x')$$

for all $x, x' \in \mathcal{X}$.

Consider two sets of points $X := \{x_1, \ldots, x_m\} \subset \mathcal{X}, Y := \{y_1, \ldots, y_n\} \subset \mathcal{X}$. We define the mean map μ by

$$\mu(X) = \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot).$$

One can define a classification rule in $\mathcal{H}$ based on the closest mean, i.e., using a hyperplane with normal vector $\mu(X) - \mu(Y)$ [4]. This begs the question: when is this normal vector zero (in which case it does not define a hyperplane)? For polynomial kernels $k(x, x') = (\langle x, x'\rangle + 1)^d$, this amounts to all empirical moments up to order d vanishing. For strictly positive definite kernels, the means coincide only if $X = Y$, rendering μ injective:

Lemma. *Assume X, Y are defined as above, k is strictly pd, and for all i, j, $x_i \neq x_j$, and $y_i \neq y_j$. If for some $\alpha_i, \beta_j \in \mathbb{R} - \{0\}$, we have*

$$\sum_{i=1}^{m} \alpha_i k(x_i, .) = \sum_{j=1}^{n} \beta_j k(y_j, .), \tag{1}$$

then $X = Y$.

To see this, assume w.l.o.g. that $x_1 \notin Y$. Subtract $\sum_{j=1}^{n} \beta_j k(y_j, .)$ from (1), and make it a sum over pairwise distinct points, to get

$$0 = \sum_i \gamma_i k(z_i, .),$$

where $z_1 = x_1, \gamma_1 = \alpha_1 \neq 0$, and $z_2, \dots \in X \cup Y - \{x_1\}$, $\gamma_2, \dots \in \mathbb{R}$. Take the RKHS dot product with $\sum_j \gamma_j k(z_j, .)$ to get

$$0 = \sum_{ij} \gamma_i \gamma_j k(z_i, z_j),$$

with $\gamma \neq 0$, hence k cannot be strictly pd. ■

The mean map has some other interesting properties. Among them is the fact that $\mu(X)$ represents the operation of taking a mean of a function on the sample X:

$$\langle \mu(X), f \rangle = \left\langle \frac{1}{m} \sum_{i=1}^{m} k(x_i, \cdot), f \right\rangle = \frac{1}{m} \sum_{i=1}^{m} f(x_i)$$

Moreover, we have

$$\|\mu(X) - \mu(Y)\| = \sup_{\|f\| \leq 1} |\langle \mu(X) - \mu(Y), f \rangle| = \sup_{\|f\| \leq 1} \left| \frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(y_i) \right|.$$

If $\mathbf{E}_{x,x' \sim p}[k(x, x')]$, $\mathbf{E}_{x,x' \sim q}[k(x, x')] < \infty$, then the above statements generalize to Borel measures p, q, with the difference being that the mean map is defined as

$$\mu\colon p \mapsto \mathbf{E}_{x \sim p}[k(x, \cdot)],$$

and the notion of strictly pd kernels is replaced by that of characteristic kernels [1]. In this case, the mean map can be viewed as a generalization of the *moment generating function* M_p of a random variable x with distribution p,

$$M_p(.) = \mathbf{E}_{x \sim p}\left[e^{\langle x,\, \cdot\, \rangle}\right].$$

If we restrict the class of distributions, the class of kernels for which μ is injective gets larger. To see this, consider a bounded translation invariant kernel $k(x, x') = \psi(x - x')$, with continuous $\psi : \mathbb{R}^d \to \mathbb{R}$, which by Bochner's theorem corresponds to a finite nonnegative Borel measure Λ. In that case, we have

$$\|\mu(p) - \mu(q)\| = \|F^{-1}[(\bar{\phi}_p - \bar{\phi}_q)\Lambda]\|,$$

where ϕ_p is the characteristic function of the measure p, $\|.\|$ is the norm of the RKHS, F^{-1} is the inverse Fourier transform, and the bar denotes complex conjugation. Roughly speaking, this shows that p and q can be distinguished as long as the spectrum Λ of the kernel is nonzero wherever the spectra of the distributions might differ. If $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, the kernel can distinguish all Borel distributions; if $\mathrm{supp}(\Lambda) \subset \mathbb{R}^d$ has a non-empty interior, it can still distinguish Borel distributions with compact support, subject to certain technical conditions (for details, see [5]).

The map μ has applications in a number of tasks including testing of homogeneity and independence [2, 3]. One can also establish a link to wave optics, which we will briefly sketch presently. We consider p as the intensity distribution of the light coming from an object which we would like to image. On the way to the sensor, there is an aperture with indicator function L (i.e., L takes the value 1 in the aperture, and 0 elsewhere). In the setting of Fraunhofer diffraction, the image intensity arising from a point source is the squared Fourier transform of L, i.e., the Fourier transform of the convolution of L with itself, $\Lambda := L * L$. For instance, in the 1-D case, if L is the indicator function of an interval, then Λ is a B_1-spline. Under the assumption of incoherent light, the image of p would thus be the convolution of p with the Fourier transform of Λ, equalling the map $\mu(p)$ induced by the translation invariant kernel associated with the Fourier transform of Λ. If the image has compact support, and the aperture has non-empty interior, then the imaging process is thus invertible.

References

[1] K. Fukumizu, A. Gretton, X. Sun and B. Schölkopf, *Kernel Measures of Conditional Dependence*, Advances in Neural Information Processing Systems 20, 489-496, MIT Press, Cambridge, MA, USA (2008)

[2] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf and A. Smola, *A Kernel Method for the Two-Sample-Problem*, Advances in Neural Information Processing Systems 19 (2007)

[3] A. Gretton, K. Fukumizu, C. H. Teo, L. Song, B. Schölkopf and A. Smola, *A Kernel Statistical Test of Independence*, Advances in Neural Information Processing Systems 20 (2008)

[4] B. Schölkopf and A. J. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, USA (2002)

[5] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet and B. Schölkopf, *Injective Hilbert Space Embeddings of Probability Measures*, Proceedings of the 21st Annual Conference on Learning Theory, USA (2008)

Reporter: Kurt Jetter