# Characteristic Kernel

Pan Chao

March 26, 2014

The idea of kernel method is to apply linear methods or calculate linear statistic after the data is mapped into a feature space which is restricted to be a RKHS $\mathcal{H}$. The mean, variance and covariance have their counterparts in the feature space. In infinite dimensional RKHSs, they are defined as **mean element**, **covariance operator** and **cross-covariance** respectively.

The notion of **characteristic kernel** (RKHS) is related to the mean element. In FBJ08, let $(\Omega, \mathcal{B})$ be a measurable space and $(\mathcal{H}, k)$ be an RKHS over $\Omega$ with the kernel $k$ measurable and bdd, and let $\mathcal{S}$ be the set of all probability measure on $(\Omega, \mathcal{B})$, then the RKHS is called characteristic (w.r.t $\mathcal{B}$) if the following map is 1-1

$$\mathcal{S} \ni P \to m_P = \mathbb{E}_{X\sim P}\left[k(\cdot, X)\right] \in \mathcal{H}, \tag{1}$$

where $m_P$ is the mean element of the random variable $k(\cdot, X)$ with law $P$. In orther words,

$$m_{P_X} = m_{P_Y} \iff P_X = P_Y.$$

The mean element is defined by Equation (4). For a random variable $X : \Omega \to \mathcal{X}$, the mean element $m_{P_X} \in \mathcal{H}_{\mathcal{X}}$ is defined as

$$\langle f, m_{P_X}\rangle_{\mathcal{H}_{\mathcal{X}}} = \mathbb{E}_X\left[f(X)\right] \quad \forall f \in \mathcal{H}_{\mathcal{X}}.$$

The existence and uniqueness of the mean element in the RKHS is guaranteed by Riesz Representation Theorem ($\mathbb{E}\left[\cdot\right]$ is a linear and bounded operator in the RKHS).

The sample version of the mean element is given by

$$\widehat{m}_{P_X} = \frac{1}{n}\sum_{i=1}^{n} k(\cdot, \mathbf{x}_i),$$

so that

$$\langle \widehat{m}_{P_X}, f\rangle = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i) = \widehat{\mathbb{E}}[f(X)], \quad \forall f \in \mathcal{H}_{\mathcal{X}},$$

where the first equality is due to linearity of inner-product and reproducing property of the kernel.

The mean element contains the information of all moments of the original random variable $X$. Let's consider the following type of pd kernel

$$k(x, y) := e^{xy}. \tag{2}$$

Remark: Thus the feature map is $\Phi_t : X \to k(X, t)$ and $\Phi_t(X) = e^{tX}$.

Its Taylor expansion is

$$\Phi_t(X) = 1 + tX + \frac{t^2}{2!}X^2 + \frac{t^3}{3!}X^3 + \cdots. \tag{3}$$

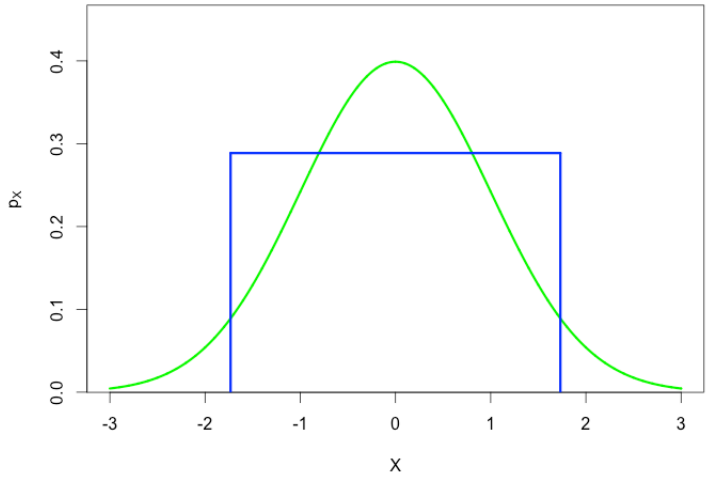
If we take the expectation, it becomes

$$\mathbb{E}_X\left[\Phi_t(X)\right] = 1 + t\mathbb{E}_X\left[X\right] + \frac{t^2}{2!}\mathbb{E}_X\left[X^2\right] + \frac{t^3}{3!}\mathbb{E}_X\left[X^3\right] + \cdots. \tag{4}$$

**Remark:** Since $m_{P_X}(t) = \mathbb{E}_X\left[k(t, X)\right]$ in the definition (1), with kernel (2) the mean element is exactly the same as moment-generating function of $X$ which contains all moments' information (it suggests some richness of the RKHS). This means that mean element of characteristic kernel is an generalized notion of moment-generating function in a real domain. $\Phi_t$ is no longer restricted to the transformation function of the random variable used for moment-generating function. As long as $\Phi_t$ (or equivently $k$) can uniquely determine a distribution, it is characteristic. Since moment-generating function can be generalized to the mean element in the real domain. I guess there is similar result in the complex domain, which generalize characteristic function. Besides, while moment-generating function is defined only for Eculidean space, mean element can be defined for any metric space as long as inner-product can be defined.

However, if some other kernel is used, for example, polynomial kernel, then the Taylor expansion is of finite sum. In this case, the mean element does not include the information of all moments of $X$. So the kernel is not characteristic.

For example, if $X_1 \sim N(0, 1)$ and $X_2 \sim Unif(-\sqrt{3}, \sqrt{3})$ are two univariate distributions. If we use the polynomial kernel of order three, $k(x, y) = (xy + c)^3$, the feature map expansion (3) would be an polynomial of order three and so is the mean element expansion (4). Since these two distributions have the same moments upto the third degree, their mean elements are the same. So this kernel cannot differetiate the two distributions.




After knowning what a mean element of a kernel does, a question is **when a kernel is characteristic or how to characterize a characteristic kernel**. The answer is related to the denseness assumption of the RKHS elaborated by the following Theorem (Proposition 5 of FBJ08).

**Theorem**

Let $(\Omega, \mathcal{B})$ be a measurable space and $(k, \mathcal{H})$ be a bdd meas. pd kernel on $\Omega$ and its RKHS. Then, $k$ is characteristic if and only if $\mathcal{H} + \mathbb{R}$ is dense in $L_P^2$ for any probability measure $P$ on $(\Omega, \mathcal{B})$.

Two examples of characteristic kernels are Gaussian kernel and Laplacian kenel:

$$K_G(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

$$K_L(x, y) = \exp\left(-\lambda \sum_{i=1}^{m} |x_i - y_i|\right)$$

**Not relevant now:** A continuous kernel $k$ on a compact metric space is called **universal** if the corresponding RKHS $\mathcal{H}$ is dense in the class of continuous functions of the space. It can be shown that a universal kernel is characteristic. So the set of universal kernels is a subset of the class of characteristic kernels. Besides, unversal kernels are only defined on compact sets while characteristic kernel can be on any kind of sets.

The idea of mean element is applied in two sample test. Given two samples, if the sample mean elements are significantly different from each other, then we conclude that they

are from different distributions. This method is proposed by Arthur Gretton and the statistic he used is called **Maximum Mean Discrepancy (MMD)**

$$MMD^2 = \|m_X - m_Y\|_{\mathcal{H}}^2.$$

Remark: The original definition of MMD is given by $MMD^2 = \sup_{f \in \mathcal{H}} \left[ \|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\|^2 \right]$. It can be shown they are equivalent.