

Reproducing Kernel Hilbert Spaces(RKHS)

Roberto Alcover Couso

29/9/2018

1 Introduction

In this document we will study a special type of Hilbert spaces, RKHS, with a kernel which meets the reproducing property. Understanding this is key for our study due to its relevance in statistical models and the ideas behind algorithms such as RDC, HSIC... Furthermore we will define a homogeneity test based on embeddings of probability distributions on RKHSs, where the distance between distributions corresponds to the distance between their embeddings. We will see that the unit ball of an RKHS is a rich enough space so that the expression for the discrepancy vanishes only if the two probability distributions are equal. At the same time it is restrictive enough for the empirical estimate at the discrepancy to converge quickly to its population counterpart as the sample size increases.

1.1 Preliminar knowledge

1. **Feature map:** ϕ is known as a feature map if is a function which maps the data to a Hilbert space \mathcal{H} (feature space).

$$\begin{aligned}\varphi : \mathcal{X} &\rightarrow \mathcal{H} \\ x &\mapsto \varphi\end{aligned}$$

2. **Kernel function:** k is called a kernel function if it is the dot product defined on a feature space.

Then, we can rewrite the dot product of the space in terms of this mapping:

$$\begin{aligned}k : \mathcal{X} \times \mathcal{X} &\rightarrow \mathcal{H} \\ (x, x') &\mapsto k(x, x') = \langle \phi(x), \phi(x') \rangle\end{aligned}$$

3. **Reproducing kernel:** a function k is a reproducing kernel of the Hilbert space \mathcal{H} if and only if it satisfies:

(a) $k(x, \cdot) \in \mathcal{H}, \forall x \in \mathcal{X}$

(b) *Reproducing property:* $\langle f, k(x, \cdot) \rangle = f(x) \forall f \in \mathcal{H}, \forall x \in \mathcal{X}$

Proposition 1. If k is a reproducing kernel then: $k(x, x') = \langle k(x, \cdot), k(x', \cdot) \rangle$

2 Maximum mean discrepancy

In this section it'll be shown how RKHSs can be used to define a homogeneity test in terms of the embeddings of the probability measures. This test consist in maximizing the measure of discrepancy between functions that belong to a certain family \mathcal{F} which must be rich enough to detect all the possible differences between the two probability measures.

2.1 Mean embedding

Lemma 1. Given two Borel probability measures \mathbb{P} and \mathbb{Q} are equal if and only if $\mathbb{E}f(X) = \mathbb{E}f(Y) \forall f \in \mathcal{C}(\mathcal{X})$

$$X \sim \mathbb{P} \text{ and } Y \sim \mathbb{Q}$$

This condition is pretty difficult to prove therefore we will keep our study in order to simplify this evaluation.

Definition 1. MMD

Let \mathcal{F} be a class of functions $f: X \rightarrow \mathbb{R}$ the MMD based on \mathcal{F} is

$$\gamma(\mathbb{P}, \mathbb{Q}) = MMD(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{F}} \{\mathbb{E}f(X) - \mathbb{E}f(Y)\}$$

This \mathcal{F} must be rich enough for it to ensure that $\mathbb{P} = \mathbb{Q} \leftrightarrow \gamma(\mathbb{P}, \mathbb{Q}) = 0$. And restrictive enough for the empirical estimate to converge quickly as the sample size increases. This will be done through RKHS with a characteristic kernel K

Riesz representation

If T is a bounded linear operator on a Hilbert space \mathcal{H} , then there exist some $g \in \mathcal{H}$ such that $\forall f \in \mathcal{H}$:

$$T(f) = \langle f, g \rangle_{\mathcal{H}}$$

Lemma 2.

Given a $K(s, \cdot)$ semi positive definite, measurable and $\mathbb{E}\sqrt{k(X, X)} < \infty$, where $X \sim \mathbb{P}$ then $\mu_p \in \mathcal{H}$ exist and fulfills the next condition $\mathbb{E}f(X) = \langle f, \mu_p \rangle$ for all $f \in \mathcal{H}$

Proof Lets define the linear operator $T_{\mathbb{P}}f \equiv \mathbb{E}(\sqrt{k(X, \cdot)}) < \infty \forall f \in \mathcal{H}$

$$|T_{\mathbb{P}}f| = |\mathbb{E}(f(X))| \leq \mathbb{E}(|f(X)|) = \mathbb{E}|\langle f, k(\cdot, X) \rangle| \leq \mathbb{E}\|f\|_{\mathcal{H}} \cdot \mathbb{E}(\sqrt{k(X, X)})^{1/2} < \infty$$

¹Reproducing property of the kernel

²Cauchy Schwarz inequality

³The expectation under \mathbb{P} of the kernel is bounded

Then using the Riesz representation theorem applied to T_p , there exist a $\mu_p \in \mathcal{H}$ such that $T_p f = \langle f, \mu_p \rangle_{\mathcal{H}}$

Definition 2. Mean embedding

Given a probability distribution \mathbb{P} we will define the mean embedding of \mathbb{P} as an element $\mu_{\mathbb{P}} \in \mathcal{H}$ such that

$$\mathbb{E}(f(X)) = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}, \forall f \in \mathcal{H}$$

If $f \in \mathcal{H}$ and $\mu_{\mathbb{P}} \in \mathcal{H}$ $\mathbb{E}(f(X)) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N f(x_n)$

Applying the Riesz representation theorem to represent $f(x_n)$
 $\forall x_n$ then:

$$f(x_n) = \langle f, K(\cdot, x_n) \rangle_{\mathcal{H}}$$

then

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N f(x_n) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \langle f, K(\cdot, x_n) \rangle_{\mathcal{H}} = \langle f, \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N K(\cdot, x_n) \rangle_{\mathcal{H}}$$

which leads to the final conclusion:

$$\mu_{\mathbb{P}} \equiv \mathbb{E}_{X \sim \mathbb{P}}(K(t, X)) \quad t \in [0, T]$$

SECOND INTERPRETATION OF THE MEAN EMBEDDING

$$\mu_{\mathbb{P}} = \mathbb{E}(K(\cdot, X))$$

2.2 Introduction to MMD

Lemma 3. Given the conditions of Lemma 2 ($\mu_{\mathbb{P}}$ and $\mu_{\mathbb{Q}}$ exist) then:

$$X \sim \mathbb{P} \mu_{\mathbb{P}} \equiv \mathbb{E}_{X \sim \mathbb{P}}(K(\cdot, X)) \quad Y \sim \mathbb{Q} \mu_{\mathbb{Q}} \equiv \mathbb{E}_{Y \sim \mathbb{Q}}(K(\cdot, Y))$$

and:

$$MMD(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

Proof

$$\begin{aligned} MMD &\equiv \sup_{f \in \mathcal{H} \quad \|f\| \leq 1} \{\mathbb{E}(f(x)) - \mathbb{E}(f(y))\} \\ &= \sup_{f \in \mathcal{H} \quad \|f\| \leq 1} \{\langle f, \mu_{\mathbb{P}} \rangle - \langle f, \mu_{\mathbb{Q}} \rangle\} \\ &= \sup_{f \in \mathcal{H} \quad \|f\| \leq 1} \langle f, (\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}) \rangle \end{aligned}$$

$$\begin{aligned} &\leq^4 \sup_{f \in \mathcal{H} \text{ } \|f\| \leq 1} \{ \|f\|_{\mathcal{H}}, \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \} \\ &\leq \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \end{aligned}$$

But on the other side, if we choose f as:

$$f = \frac{1}{\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|} (\mu_{\mathbb{P}} - \mu_{\mathbb{Q}})$$

then we have:

$$\sup_{f \in \mathcal{H} \text{ } \|f\| \leq 1} \{ \|f\|_{\mathcal{H}}, \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \} \geq \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

therefore

$$MMD = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$$

Proposition 1

Given: $X, X' \sim \mathbb{P}$ and $Y, Y' \sim \mathbb{Q}$ and X and Y are independent then:

$$MMD^2(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = \mathbb{E}(K(X, X')) + \mathbb{E}(K(Y, Y')) - 2\mathbb{E}K(X, Y).$$

proof

$$\begin{aligned} MMD^2(\mathcal{F}, \mathbb{P}, \mathbb{Q}) &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}^2 \\ &= \langle \mu_{\mathbb{P}} - \mu_{\mathbb{Q}}, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \\ &= \langle \mathbb{E}(K(\cdot, X)) - K(\cdot, Y), \mathbb{E}(K(\cdot, X')) - K(\cdot, Y') \rangle \\ &= \mathbb{E}(\langle K(\cdot, X), K(\cdot, X') \rangle + \langle K(\cdot, Y), K(\cdot, Y') \rangle - 2\langle K(\cdot, X), K(\cdot, Y) \rangle) \\ &=^5 \mathbb{E}(K(X, X') + K(Y, Y') - 2K(X, Y)) \\ &= \mathbb{E}(K(X, X')) + \mathbb{E}(K(Y, Y')) - 2\mathbb{E}(K(X, Y)) \\ &= \int \int K(s, t) \underbrace{d(\mathbb{P} - \mathbb{Q})(s)}_{\text{Signedmeasure}} d(\mathbb{P} - \mathbb{Q})(t) \end{aligned}$$

Prooving that MMD defines an homogeneity test

Definition 3. Characteristic kernel

A reproducing kernel k is a characteristic kernel if the induced γ_k is a metric.

Theorem 2.

⁴Cauchy Schwarz inequality

⁵is due to the reproductive property of the kernel.

If X is a compact metric space, k is continuous and \mathcal{H} is dense in $\mathcal{C}(X)$ with respect to the supremum norm, then \mathcal{H} is characteristic.

Proof. Being characteristic means that $MMD(\mathcal{F}, \mathbb{P}, \mathbb{Q}) = 0 \leftrightarrow \mathbb{P} = \mathbb{Q}$

→

By lemma 1 we know that \mathbb{P} and \mathbb{Q} are equal if and only if $\mathbb{E}f(X) = \mathbb{E}f(Y)$
 $\forall f \in \mathcal{C}(\mathcal{X})$

Given that \mathcal{H} is dense in $\mathcal{C}(X)$ then:

$$\forall \epsilon > 0, f \in \mathcal{C}(X), \exists g \in \mathcal{H} : \|f - g\|_\infty < \epsilon$$

$$\begin{aligned} |\mathbb{E}(f(X)) - \mathbb{E}(f(Y))| &= |\mathbb{E}(f(X)) - \mathbb{E}(g(X)) + \mathbb{E}(g(X)) - \mathbb{E}(g(Y)) + \mathbb{E}(g(Y)) - \mathbb{E}(f(Y))| \\ &\leq |\mathbb{E}(f(X)) - \mathbb{E}(g(X))| + |\mathbb{E}(g(X)) - \mathbb{E}(g(Y))| + |\mathbb{E}(g(Y)) - \mathbb{E}(f(Y))| \\ &= |\mathbb{E}(f(X)) - \mathbb{E}(g(X))| + |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + |\mathbb{E}(g(Y)) - \mathbb{E}(f(Y))| \\ &\leq \mathbb{E}|f(X) - g(X)| + |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + \mathbb{E}|g(Y) - f(Y)| \\ &\leq^1 \|f - g\|_\infty + |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + \|f - g\|_\infty \\ &\leq |\langle g, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| + 2\epsilon \end{aligned}$$

By lemma 3 we know that if $MMD = 0$ then $\mu_{\mathbb{P}} = \mu_{\mathbb{Q}}$. Hence:

$$|\mathbb{E}(f(X)) - \mathbb{E}(f(Y))| \leq 2\epsilon$$

Then by lemma 1 \mathbb{P} and \mathbb{Q} are equal.

←

By definition of MMD.

3 HSIC

From the MMD criterion we will develop an independence criterion which will be conducted by the following idea: Given $\mathcal{X} \sim \mathbb{P}$ and $\mathcal{Y} \sim \mathbb{Q}$ whose joint distribution is $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$ then the test of independence between these variables will be determining if $\mathbb{P}_{\mathcal{X}\mathcal{Y}}$ is equal to the product of the marginals $\mathbb{P}\mathbb{Q}$. Therefore:

$\mathcal{MMD}(\mathcal{F}, \mathbb{P}_{\mathcal{X}\mathcal{Y}}, \mathbb{P}\mathbb{Q}) = 0$ if and only if \mathcal{X} and \mathcal{Y} are independent. To characterize this independence test we need to introduce a new RKHS, which is a tensor product of the RKHSs in which the marginal distributions of the random variables are embedded. Let \mathcal{X} and \mathcal{Y} be two topological spaces and let k and l be kernels on these spaces, with respective RKHS \mathcal{H} and \mathcal{G} . Let us denote

as $v((x, y), (x', y'))$ a kernel on the product space $\mathcal{X} \times \mathcal{Y}$ with RKHS \mathcal{H}_v . This space is known as the tensor product space $\mathcal{H} \times \mathcal{G}$. Tensor product spaces are defined as follows:

Definition 4. Tensor product The tensor product of Hilbert spaces \mathcal{H}_1 and \mathcal{H}_2 with inner products $\langle \cdot, \cdot \rangle_1$ and $\langle \cdot, \cdot \rangle_2$ is defined as the completion of the space $\mathcal{H}_1 \times \mathcal{H}_2$ with inner product $\langle \cdot, \cdot \rangle_1 \otimes \langle \cdot, \cdot \rangle_2$ extended by linearity. The resulting space is also a Hilbert space.

Lemma 4.

A kernel v in the tensor product space $\mathcal{H} \times \mathcal{G}$ can be defined as:

$$v((x, y), (x', y')) = k(x, x')l(y, y')$$

Useful definitions for the following content

$$\mathbb{E}_{\mathcal{X}} f(\mathcal{X}) = \int f(x) d\mathbb{P}(x)$$

$$\mathbb{E}_{\mathcal{Y}} f(\mathcal{Y}) = \int f(y) d\mathbb{Q}(y)$$

$$\mathbb{E}_{\mathcal{XY}} f(\mathcal{XY}) = \int f(x, y) d\mathbb{P}_{\mathcal{XY}}(x, y)$$

Using this notation, the mean embedding of $\mathbb{P}_{\mathcal{XY}}$ and $\mathbb{P}_{\mathcal{Q}}$ are:

$$\mu_{\mathbb{P}_{\mathcal{XY}}} = \mathbb{E}_{\mathcal{XY}} v((\mathcal{X}, \mathcal{Y}), \cdot)$$

$$\mu_{\mathbb{P}_{\mathcal{Q}}} = \mathbb{E}_{\mathcal{Y}} v((\mathcal{X}, \mathcal{Y}), \cdot)$$

In terms of these embeddings:

$$\mathcal{MMD}(\mathcal{F}, \mathbb{P}_{\mathcal{XY}}, \mathbb{P}_{\mathcal{Q}}) = \|\mu_{\mathbb{P}_{\mathcal{XY}}} - \mu_{\mathbb{P}_{\mathcal{Q}}}\|_{\mathcal{H}_v}$$