Machine Learning Seminar

Mohit Gupta (110050085)
Deepali Adlakha (11D170020)

On the Relation Between Universality, Characteristic Kernels and RKHS Embedding of Measures

Bharath K. Sriperumbudur[⋆], Kenji Fukumizu[†] and
Gert R. G. Lanckriet[⋆]

[⋆]UC San Diego [†]The Institute of Statistical Mathematics

AISTATS 2010

Outline

- Kernel Based Learning
- RKHS embedding of probability measures
- Characteristic kernels
- Universal kernels
- Various notions of universality
- Novel characterization of universality
- Relation to RKHS embedding of signed measures

SVM Optimisation Problem

We want to optimise the following problem

$$\min_{w \in \mathbb{R}^n} \|w\| + C \sum_{i=1}^{m} l(x_i, y_i, w),$$

Where C is a *regularisation constant*

Note: SVM considers linear boundary functions only.

Non Linear functions

To allow for non-linear functions as our separator function:

Define a feature map :

$$\boldsymbol{\varphi}: \mathrm{X} \rightarrow \mathcal{H}$$

Where $\boldsymbol{\varphi}$ can have non-linear mappings of components of X

Kernel Trick

- We only need inner product of two points in the hilbert space
- Reproducing Kernel Hilbert Space

Define the mapping as

$$X \longrightarrow \varphi(X) := K(\cdot, x)$$

Where $K(\cdot, x)$ measures the similarity of x another point

Representer Theorem

Solutions to the optimisation problem is of the form

$$f := \sum_{j \in \mathbb{N}_n} c_j k(\cdot, x_j)$$

Can f approximate any target function arbitrarily well as $n \to \infty$?

RKHS Embeddings of Probability Measures

- *Input space* : X
- *Feature space* : $\mathcal{H}$ (with reproducing kernel, k)
- *Feature map* : Φ

$$\Phi : X \to \mathcal{H} \qquad\qquad x \mapsto \Phi(x) := k(\cdot, x)$$

Extension to probability measures:

$$\mathbb{P} \mapsto \Phi(\mathbb{P}) := \underbrace{\int_X k(\cdot, x)\, d\mathbb{P}(x)}_{E_{Y\sim\mathbb{P}}[\Phi(Y)] = E_{Y\sim\mathbb{P}}[k(\cdot, Y)]}$$

Advantage: $\Phi(\mathbb{P})$ can distinguish $\mathbb{P}$ by *high-order moments.*

$$\begin{aligned} k(y,x) &= c_0 + c_1(xy) + c_2(xy)^2 + \cdots \quad (c_i \neq 0) \qquad \text{e.g. } k(y,x) = e^{xy} \\ \Phi(\mathbb{P})(y) &= c_0 + c_1\left(\int_X x\, d\mathbb{P}(x)\right) y + c_2\left(\int_X x^2\, d\mathbb{P}(x)\right) y^2 + \cdots \end{aligned}$$

Characteristic Kernels

Define: k is *characteristic* if

$$\mathbb{P} \mapsto \int_X k(\cdot, x)\, d\mathbb{P}(x) \quad \textit{is injective.}$$

In other words,

$$\int_X k(\cdot, x)\, d\mathbb{P}(x) = \int_X k(\cdot, x)\, d\mathbb{Q}(x) \Leftrightarrow \mathbb{P} = \mathbb{Q}.$$

Applications

Two-sample problem:

- Given random samples $\{X_1, \ldots, X_m\}$ and $\{Y_1, \ldots, Y_n\}$ drawn i.i.d. from $\mathbb{P}$ and $\mathbb{Q}$, respectively.
- *Determine:* are $\mathbb{P}$ and $\mathbb{Q}$ different?
- $\gamma(\mathbb{P}, \mathbb{Q}) = \|\Phi(\mathbb{P}) - \Phi(\mathbb{Q})\|_{\mathcal{H}}$: distance metric between $\mathbb{P}$ and $\mathbb{Q}$.

$$\begin{array}{c} H_0 : \mathbb{P} = \mathbb{Q} \\ \\ H_1 : \mathbb{P} \neq \mathbb{Q} \end{array} \equiv \begin{array}{c} H_0 : \gamma(\mathbb{P}, \mathbb{Q}) = 0 \\ \\ H_1 : \gamma(\mathbb{P}, \mathbb{Q}) > 0 \end{array}$$

- *Test:* Say H_0 if $\widehat{\gamma}(\mathbb{P}, \mathbb{Q}) < \varepsilon$. Otherwise say H_1.

Other applications:

- *Hypothesis testing* : Independence test, Goodness of fit test, etc.
- Feature selection, message passing, density estimation, etc.

c- Universality [Steinwart, 2001]

- X : compact metric space
- k : continuous on $X \times X$
- Target function space : C(X), continuous functions on X

Define k to be c-universal if H is dense in C(X) w.r.t. the uniform norm

$(\|f\|_u \coloneqq \sup_{x \in X} |f(x)|)$.

Sufficient conditions are obtained based on the Stone-Weierstraß theorem.

Examples: Gaussian and Laplacian kernels on any compact subset of R^d

Stone Weierstraß Theorem

Theorem 1 *Let (X, d) be a compact metric space and $A \subset C(X)$ be an algebra. Then A is dense in $C(X)$ if both A does not vanish, i.e. for all $x \in X$ there exists an $f \in A$ with $f(x) \neq 0$, and A separates points, i.e. for all $x, y \in X$ with $x \neq y$ there exists an $f \in A$ with $f(x) \neq f(y)$.*

However, one limitation in the c-universality is that X is assumed to be compact, which excludes many interesting spaces, such as R^d and infinite discrete sets

cc- Universality [Micchelli et al., 2006]

- X : Hausdorff space
- k : continuous on $X \times X$
- Target function space : C(X)

Define k to be cc-universal if H is dense in C(X) endowed with the topology of compact convergence.

Necessary and sufficient conditions related to the injectivity of RKHS embedding of measures are obtained

Examples: Gaussian, Laplacian and Sinc kernels on R^d.

Issue

Topology of compact convergence is weaker than the topology of uniform convergence.

The question is whether we can characterize H that are rich enough to approximate any f^* on non-compact X in a stronger sense, i.e., uniformly, by some $g \in H$.

Proposed Notion: c_0 universality

- X : locally compact Hausdorff (LCH) space
- Target function space : $C_0(X)$, the space of bounded continuous functions that "vanish at infinity" (for every $\varepsilon > 0$, $\{x \in X : |f(x)| \geq \varepsilon\}$ is compact).
- k is bounded and $k(., x) \in C_0(X)$ for all $x \in X$.

Define k to be c_0-universal if H is dense in $C_0(X)$ w.r.t. $\|.\|_u$.

Handles non-compact X and ensures uniform convergence over entire X.

Radon Measure

A Radon measure μ on a Hausdorff space X is a Borel measure on X satisfying :

- $\mu(C) < \infty$ for each compact subset $C \subset X$ and
- $\mu(B) = \sup \{\mu(C) | C \subset B, C \text{ compact}\}$ for each B in the Borel σ-algebra of X

Radon Measure notations

- μ is said to be finite if $\|\mu\| := |\mu|(X) < \infty$, where $|\mu|$ is the total variation of μ
- $M_b(X)$ denotes the space of all finite signed Radon measures on X
- $M^1_+(X)$ denotes the space of all Radon probability measures
- $M_{bc}(X)$ denotes the space of all compactly supported finite signed Radon measures on X.
- For $\mu \in M_b(X)$,the support of μ is defined as $\mathrm{supp}(\mu) = \{x \in X \mid$ for any open set U such that $x \in U$, $|\mu|(U) \neq 0\}$

Embedding Characterization of Universality

- k is c_0-universal if and only if

$$\mu \rightarrow \int_X k(\cdot, x)\, d\mu(x), \qquad \mu \in M_b(X),$$

is injective. $M_b(X)$ is the space of finite signed Radon measures on X

- k is cc-universal if and only if

$$\mu \rightarrow \int_X k(\cdot, x)\, d\mu(x), \qquad \mu \in M_{bc}(X),$$

is injective. $M_{bc}(X) = \{\mu \in M_b(X) \mid \mathrm{supp}(\mu) \text{ is compact}\}$

- k is c-universal if and only if

$$\mu \rightarrow \int_X k(\cdot, x)\, d\mu(x), \qquad \mu \in M_b(X),$$

is injective

Positive Definite Characterization of Universality

- k is c_0-universal (resp. c-universal) if and only if

$$\int_{\mathbf{x}}\int_{\mathbf{y}} k(x, y)\ d\mu(x)\ d\mu(y) > 0, \qquad \forall\ \mu \in M_b(X)\backslash\{0\}$$

- k is cc-universal if and only if

$$\int_{\mathbf{x}}\int_{\mathbf{y}} k(x, y)\ d\mu(x)\ d\mu(y) > 0, \qquad \forall\ \mu \in M_{bc}(X)\backslash\{0\}$$

- If k is c-, cc- or c_0-universal, then it is *strictly positive definite*

X is an LCH space: Summary

$$
\begin{array}{ccc}
\begin{array}{c}\clubsuit \text{ holds} \\ \forall\, \mu \in M_b(X)\end{array} & \begin{array}{c}\Longrightarrow \\ \not\Longleftarrow\end{array} & \begin{array}{c}\clubsuit \text{ holds} \\ \forall\, \mu \in M_{bc}(X)\end{array} \\
\Updownarrow & & \Updownarrow \\
c_0\textit{-universal} & \begin{array}{c}\Longrightarrow \\ \not\Longleftarrow\end{array} & cc\textit{-universal} \\
\Downarrow \; \not\Uparrow & & \Downarrow \; \not\Uparrow \\
\textit{characteristic} & \begin{array}{c}\overset{?}{\Longrightarrow} \\ \not\Longleftarrow\end{array} & \textit{strictly pd}
\end{array}
$$

$$\clubsuit \; : \; \iint_X k(x,y)\, d\mu(x)\, d\mu(y) > 0$$

Translation invariant Kernels on R^d

$X = R^d$ and $k(x, y) = \psi(x - y)$, where

$$\psi(x) = \int_{\mathbb{R}^d} e^{\sqrt{-1}x^T\omega}\, d\Lambda(\omega),\ x \in \mathbb{R}^d,$$

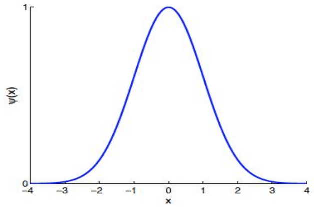
and Λ is a non-negative finite Borel measure

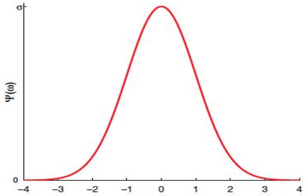
Theorem

- k is c_0-universal if and only if $\text{supp}(\Lambda) = R^d$
- k is c_0-universal if and only if it is characteristic
- If $\text{supp}(\Lambda)$ has a non-empty interior, then k is cc-universal. [Micchelli et al., 2006]

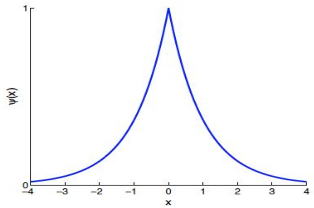
Examples

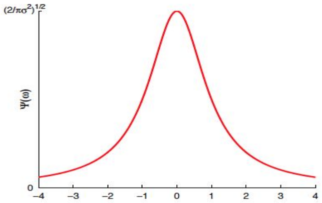
- Gaussian kernel: $\psi(x) = e^{-x^2/2\sigma^2}$; $\Psi(\omega) = \sigma e^{-\sigma^2\omega^2/2}$; $d\Lambda(\omega) = \Psi(\omega)\, d\omega$.





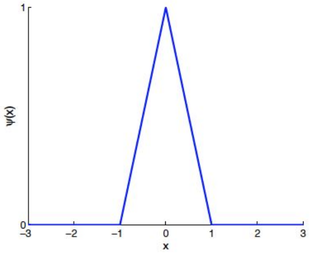
- Laplacian kernel: $\psi(x) = e^{-\sigma|x|}$; $\Psi(\omega) = \sqrt{\frac{2}{\pi}} \frac{\sigma}{\sigma^2+\omega^2}$.

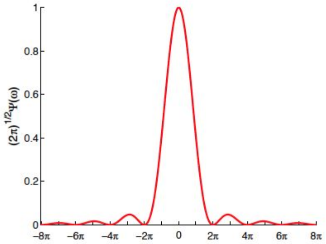




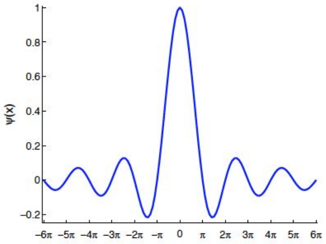
Examples

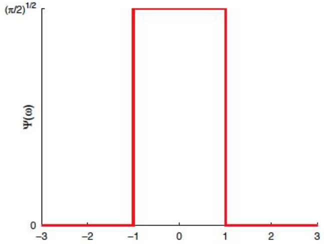
- B_1-spline kernel: $\psi(x) = (1 - |x|)\mathbb{1}_{[-1,1]}(x)$; $\Psi(\omega) = \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{\sin^2(\frac{\omega}{2})}{\omega^2}$.





- Sinc kernel: $\psi(x) = \frac{\sin(\sigma x)}{x}$; $\Psi(\omega) = \sqrt{\frac{\pi}{2}}\mathbb{1}_{[-\sigma,\sigma]}(\omega)$.





Summary

Characteristic kernel

- Injective RKHS embedding of probability measures
- Applications: Hypothesis testing, feature selection, etc

Universal kernel

- Consistency of learning algorithms
- Injective RKHS embedding of finite signed Radon measures

Conclusion

The characteristic and universal kernels are essentially the same except that universal kernels deal with some subset of M_b (X) while characteristic kernels deal with probability measures.

For a set of kernels such as Translation Invariant Kernels, the concepts of characteristic and universal kernels are equivalent.

References

- Sriperumbudur, Bharath K., Kenji Fukumizu, and Gert Lanckriet. "On the relation between universality, characteristic kernels and RKHS embedding of measures." *International Conference on Artificial Intelligence and Statistics*. 2010.
- Steinwart, Ingo. "On the influence of the kernel on the consistency of support vector machines." *The Journal of Machine Learning Research* 2 (2002): 67-93.
- van Gaans, Onno. "Probability measures on metric spaces, Notes of the seminar ‘Stochastic Evolution Equations’." *Delft University of Technology*(2003).
- Nath, Saketh. “Notes on Topics in Machine Learning” Indian Institute of Technology, Bombay (2014).

Appendix

Banach Space

A **Banach space** is a complete normed vector space. Thus, a Banach space is a vector space with a metric that allows the computation of vector length and distance between vectors and is complete in the sense that a Cauchy sequence of vectors always converges to a well defined limit in the space.

Hausdorff Space

Points x and y in a topological space X can be *separated by neighbourhoods* if there exists a neighbourhood U of x and a neighbourhood V of y such that U and V are disjoint ($U \cap V = \varnothing$). X is a **Hausdorff space** if any two distinct points of X can be separated by neighborhoods.

Sigma Algebra

A σ-algebra on a set X is a set of subsets of X that contains ∅ and is closed under elementary, countable set operations such as complements and countable intersections.

Borel Sigma Algebra

Let (X, d) be a metric space. The *Borel σ-algebra* (σ-field) $\mathcal{B} = \mathcal{B}(X)$ is the smallest σ-algebra in X that contains all open subsets of X. The elements of $\mathcal{B}$ are called the *Borel sets* of X.

Let (X, d) be a metric space. A *finite Borel measure* on X is a map $\mu : \mathcal{B}(X) \to [0, \infty)$ such that

$$\mu(\emptyset) = 0, \text{ and}$$
$$A_1, A_2, \ldots \in \mathcal{B} \text{ mutually disjoint } \implies \mu(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mu(B_i).$$

μ is called a *Borel probabiliy measure* if in addition $\mu(X) = 1$.

Locally compact spaces

A locally compact space is a Hausdorff topological space with the

property :

Every point has a compact neighborhood.

Let X be a locally compact space, let K be a compact set in X, and let D be an open

subset, with $K \subset D$. Then there exists an open set E with:

(i) E compact

(ii) $K \subset E \subset E^{c} \subset D$

Hahn-Banach

Theorem 2 (Hahn-Banach). *Suppose A be a subspace of a locally convex topological vector space Y. Then A is dense in Y if and only if $A^{\perp} = \{0\}$, where*

$$A^{\perp} := \{T \in Y' : \forall x \in A,\ T(x) = 0\}.$$

Riesz representation theorem

It says that any bounded linear functional T on the space of compactly supported continuous functions on X is the same as integration against a measure μ,

$$T f = \int f \, d\mu$$

Here, the integral is the Lebesgue integral.

C_0 universality

Theorem 3 (Characterization of c_0-*universality*). k *is* c_0-*universal if and only if the embedding,*

$$\mu \mapsto \int_X k(\cdot, x)\, d\mu(x), \;\; \mu \in M_b(X),$$

is injective.

C_0 universality

Proposition 4. *k is c_0-universal if and only if*

$$\iint_X k(x,y)\,d\mu(x)\,d\mu(y) > 0,\ \forall\, 0 \neq \mu \in M_b(X).$$

$$\begin{aligned}
0 &= \left\langle \int_X k(\cdot,x)\,d\mu(x), \int_X k(\cdot,x)\,d\mu(x) \right\rangle_{\mathcal{H}} \\
&\overset{(a)}{=} \iint_X k(x,y)\,d\mu(x)\,d\mu(y),
\end{aligned}$$

pd kernels

Proposition 5 (c_0-*universal* kernels are strictly pd).
If k is c_0-universal, then it is strictly pd.

Special Kernels

(A_1) k is translation invariant on $\mathbb{R}^d \times \mathbb{R}^d$, i.e., $k(x, y) = \psi(x - y)$, where $0 \neq \psi \in C_b(\mathbb{R}^d)$ is a pd function on $\mathbb{R}^d$.

(A_2) k is a radial kernel on $\mathbb{R}^d \times \mathbb{R}^d$, i.e., $k(x, y) = \varphi(\|x - y\|_2^2)$, $x, y \in \mathbb{R}^d$, where $\varphi \in C_0(\mathbb{R})$ is *completely monotone* (Wendland, 2005, Chapter 7) on $[0, \infty)$.

(A_3) X is an LCH space with bounded k. Let $k(x, y) = \sum_{j \in I} \phi_j(x)\phi_j(y)$, $(x, y) \in X \times X$, where we assume that series converges uniformly on $X \times X$. $\{\phi_j : j \in I\}$ is a set of continuous real-valued functions on X where I is a countable index set.

Bochner Theorem

Theorem 7 (Bochner). *$\psi \in C_b(\mathbb{R}^d)$ is pd on $\mathbb{R}^d$ if and only if it is the Fourier transform of a finite non-negative Borel measure Λ on $\mathbb{R}^d$, i.e.,*

$$\psi(x) = \int_{\mathbb{R}^d} e^{-ix^T\omega}\, d\Lambda(\omega),\ x \in \mathbb{R}^d. \tag{6}$$

Translation Invariant Kernels

Proposition 8 (Translation invariant kernels on $\mathbb{R}^d$). *Suppose* (A_1) *holds and* $\psi \in C_0(\mathbb{R}^d)$. *Then* k *is* c_0*-universal if and only if* $\operatorname{supp}(\Lambda) = \mathbb{R}^d$.

$$\begin{aligned}
B &= \iint_{\mathbb{R}^d} \psi(x-y)\, d\mu(x)\, d\mu(y) \\
&\overset{(a)}{=} \iiint_{\mathbb{R}^d} e^{-i(x-y)^T\omega}\, d\Lambda(\omega)\, d\mu(x)\, d\mu(y) \\
&\overset{(b)}{=} \iint_{\mathbb{R}^d} e^{-ix^T\omega}\, d\mu(x) \int_{\mathbb{R}^d} e^{iy^T\omega}\, d\mu(y)\, d\Lambda(\omega) \\
&= \int_{\mathbb{R}^d} \hat{\mu}(\omega)\overline{\hat{\mu}(\omega)}\, d\Lambda(\omega) = \int_{\mathbb{R}^d} |\hat{\mu}(\omega)|^2\, d\Lambda(\omega), (7)
\end{aligned}$$

Translation Invariant Kernels

Assumption 1 *$k(x, y) = \psi(x - y)$ where ψ is a bounded continuous real-valued positive definite function*[4] *on $M = \mathbb{R}^d$.*

Theorem 7 *Let $\mathcal{F}$ be a unit ball in an RKHS $(\mathcal{H}, k)$ defined on $\mathbb{R}^d$. Suppose k satisfies Assumption 1. Then k is a characteristic kernel to the family, $\mathfrak{S}$, of all probability measures defined on $\mathbb{R}^d$ if and only if $supp(\Lambda) = \mathbb{R}^d$.*