Stable Learning

Foundations and Applications

Liping Wang

CRIPAC, CASIA, UCAS

Institute of Automation 2022.4

中国科学院自动化研究所
Institute of Automation
Chinese Academy of Sciences

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

Contents

Background Knowledge

Hilbert Space & Kernel

Definition (Hilbert Space)

A Hilbert space is a real or complex **inner product space** that is also a **complete** metric space with respect to the distance function induced by the inner product (i.e. complete inner space).

Motivation: to generalize methods of linear algebra and calculus from the finite-dimensional Euclidean spaces to infinite-dimensional spaces.

Definition (Kernel)

Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if there exists an $\mathbb{R}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x_1, x_2 \in \mathcal{X}$,

$$k(x_1, x_2) := \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}}, \tag{1}$$

i.e. functions that can be written as an inner product in a feature space.

RKHS(Reproducing Kernel Hilbert Spaces)

Definition (RKHS)

Let $\mathcal{H}$ be a Hilbert space of of $\mathbb{R}$-valued functions defined on non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space, if k satisfies

- $\forall x \in \mathcal{X}, k(\cdot, x) \in \mathcal{H}$
- $\forall x \in \mathcal{X}, \langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ (the reproducing property).

Notions:

- RKHS is a specific kind of Hilbert space.
- In particular, $\forall x, y \in \mathcal{X}, k(x, y) = \langle k(\cdot, x), k(\cdot, y)\rangle$

Further: We can prove the uniqueness of reproducing kernel.

Reproducing Kernel

Actually, the following three concepts are the equivalent:

- Kernel
- Reproducing kernel
- Positive definite function

Definition (Positive definite function)

A symmetric function $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if $\forall n \geq 1, \forall (a_1, a_2, \ldots, a_n) \in \mathbb{R}^n$, $\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) \geq 0. \tag{2}$$

Cross-Covariance Operator

Definition (Cross-covariance operator)

Suppose we have a random variable (X, Y) on $\mathcal{X} \times \mathcal{Y}$, and RKHSs $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ on $\mathcal{X}, \mathcal{Y}$, respectively, with reproducing kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$, assuming $E\left[k_{\mathcal{X}}(X, X)\right] < \infty, E\left[k_{\mathcal{Y}}(Y, Y)\right] < \infty$, the cross-covariance operator $\Sigma_{YX} : \mathcal{H}_{\mathcal{X}} \to \mathcal{H}_{\mathcal{Y}}$ is defined by the unique bounded operator that satisfies

$$\langle g, \Sigma_{YX} f \rangle_{\mathcal{H}Y} = \mathrm{Cov}[f(X), g(Y)] \quad (= E[f(X)g(Y)] - E[f(X)]E[g(Y)]) \tag{3}$$

Hilbert-Schmidt

Definition (Hilbert-Schmidt Norm)

Denote by $C : \mathcal{F} \rightarrow \mathcal{G}$ a linear operator. Then provided the sum converges, the Hilbert-Schmidt (HS) norm of C is defined as

$$\|C\|_{HS} := \sum_{i,j} \langle Cu_i, v_j \rangle_{\mathcal{G}}^2 \tag{4}$$

Definition (Hilbert-Schmidt Operator)

A linear operator $C : \mathcal{F} \rightarrow \mathcal{G}$ is called a Hilbert-Schmidt operator if its HS norm exists.

HSIC(Hilbert-Schmidt Independence Criterion)[1]

Definition (HSIC)

Given random variable (X, Y) on $\mathcal{X} \times \mathcal{Y}$, and RKHSs $\mathcal{H}_{\mathcal{X}}$ and $\mathcal{H}_{\mathcal{Y}}$ on $\mathcal{X}, \mathcal{Y}$, HSIC is defined as the squared HS-norm of cross-covariance operator Σ_{YX}:

$$HSIC(p_{xy}, \mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}) := \|\Sigma_{YX}\|_{HS}^2. \tag{5}$$

Definition (Empirical HSIC)

Let $Z := \{(x_1, y_1), \ldots, (x_m, y_m)\} \subset \mathcal{X} \times \mathcal{Y}$ be a series of m independent drawn from p_{xy}. An estimator of HSIC is given by

$$HSIC(Z, \mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}) := (m-1)^{-2} \operatorname{tr} KHLH, \tag{6}$$

where $H, K, L \in \mathbb{R}^{m \times m}, K_{ij} = k(x_i, x_j), L_{ij} = l(y_i, y_j)$ and $H_{ij} = \delta_{ij} - m^{-1}$

[1] Arthur Gretton et al. "Measuring Statistical Dependence with Hilbert-Schmidt Norms". In: *Proceedings of the 16th International Conference on Algorithmic Learning Theory*. 2005.

Approximation of Empirical HSIC

Theorem ($O(m^{-1})$ Bias of Estimator)

$$\operatorname{HSIC}\left(p_{xy}, \mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}}\right) = \mathrm{E}_Z[\operatorname{HSIC}(Z, \mathcal{H}_{\mathcal{X}}, \mathcal{H}_{\mathcal{Y}})] + O\left(m^{-1}\right) \tag{7}$$

Applications of HSIC

Deep Stable Learning for Out-Of-Distribution Generalization.(CVPR 2021)[2]

[2] Xingxuan Zhang et al. "Deep Stable Learning for Out-Of-Distribution Generalization". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Conclusion

Conclusion

Thank you!