"Hilbert Space Embedding of Conditional Distribution with Application to Dynamical Systems"

report by Mohsen Nabian

The main contributions of this paper is as follows:

① the paper introduced the concept of embedding conditional distributions in an RKHS (Reproducing kernel Hilbert Space) and presented a novel method for estimating such embeddings from training data.

② The paper consider several usefull probabolistic inference operations such as marginalization and conditioning and show, using the author's theory that these operations can be performed only in RKHS.

(3) The article, applied their own inference algorithm to learn non-parametric models and perform inferenc for dynamical systems. These algorithms are general because it handels wide variety of possible nonline nongaussian models and also they apply in any setting in which an appropriate kernel function can be defined.

---

reproducing kernel Hilbert space (RKHS) F on $\mathcal{X}$ with kernel K is a Hilbert space of function $f: X \to \mathbb{R}$. Its dot product $\langle \cdot, \cdot \rangle_F$ satisfy the reproducing property:

$$\text{(1)} \quad \langle f(\cdot), K(x, \cdot) \rangle_F = f(x)$$

$$\Downarrow$$

$$\text{(2)} \quad \langle K(x, \cdot), K(x', \cdot) \rangle_F = K(x, x')$$

Scanned by CamScanner

also true map and its emerical estimates are

(a) $\mu_x := E_x[\varphi(X)]$ Ex : expectation

(b) $\hat{\mu}_X := \frac{1}{m} \sum_{i=1}^{m} \varphi(x_i)$ → feature map

Def 1) when the mean map $\mu_x : P \rightarrow F$ is injective, the kernel function k is called characteristic

The 2) The emperical mean $\hat{\mu}_x$ converges to μ_x in the RKHS. norm at rate of $O_p(m^{-1/2})$

m ↓ size of training set

$D_X = \{x_1, \ldots, x_m\}$

Cross-Covarience Operator:

$C_{XY} : \mathcal{G} \rightarrow F$

$C_{XY} = E_{XY}[\varphi(X) \otimes \Phi(Y)] - \mu_X \otimes \mu_Y$

$\otimes$ ⇒ Tensor product

given two functions $f \in F$ and $g \in \mathcal{G}$ their cross-Covarience

$$\mathrm{Cov}_{xy}[f(X), g(Y)] := E_{XY}[f(X)g(Y)] - E_X[f(X)]\,E_Y[g(Y)]$$ will be

computed as: $\langle f, C_{XY}\,g\rangle_F$ or $\langle f \otimes g, C_{XY}\rangle_{F\otimes g}$

for example:

given ~~map~~ m pairs of training examples

$$D_{XY} = \{(x_1, y_1), \dots, (x_m, y_m)\}$$

iid from $P(X, Y)$,

$$\Upsilon = (\varphi(x_1), \dots, \varphi(x_m))$$

$$\Phi = (\phi(y_1), \dots, \phi(y_m))$$

Theorem 5: let $k_x := \Upsilon^T \varphi(x)$

then $\hat{\mu}_{Y|x}$ can be estimated as ~~$\mu_{Y|x}$~~

$$\hat{\mu}_{Y|x} = \Phi(HK + \lambda m I)^{-1} H k_x = \sum_{i=1}^{m} \beta_i(x)\,\phi(y_i)$$

real values GD

Theorem 6:

Assume $k(x,\cdot)$ is in range of C_{XX}.
The emperical conditional embedding $\hat{\mu}_{Y|x}$ converges to $\mu_{Y|x}$ in the RKHS norm.

Operations on RKHS Embeddings:

we have the following:

$$\mu_X = E_Y[\mathcal{U}_{X|Y}\phi(Y)] = \mathcal{U}_{X|Y} E_Y[\phi(Y)] =$$
$$= \mathcal{U}_{X|Y}\mu_Y$$

Chain Rule:

Similar to $P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$

we have:

$$\mu_{XY} = \mathcal{U}_{X|Y}\mu_Y^{\otimes} = \mathcal{U}_{Y|X}\mu_X^{\otimes}$$

Here: $\mu_X^{\otimes} = E_X[\varphi(X) \otimes \varphi(X)]$ & $\mu_Y^{\otimes} = E_Y[\phi(Y) \otimes \phi(Y)]$

Appeli

The paper also extract kernel formulation for Conditional cross validation.

Application: Dynamical Systems:

obj: Learning and inference in a dynamical system

The article models uncertainity in a dynamic system using partially observable Markov model which is:

$P(s^1, \ldots, s^T, o^1, \ldots, o^T)$ s^t is hidden state at time t.

&

o^t corresponding observation

Theorem: Hilbert space prediction is given by:

$$\mu_{s^{t+1}|h} = U_{s^{t+1}|s^t} \mu_{s^t|h}$$

Theorem: Hilbert space conditioning step is given by:

$$\mu_{s^{t+1}|h}$$

Learning Algorithm:

input $\Upsilon = (\varphi(s^t))$ $\quad \Phi = (\phi(s^{t+1}))$

$\Psi = (\psi(o^{t+1}))$ $\quad t = 1, 2, \dots m$

① Compute $K = \Upsilon^T \Upsilon$, $U = \Psi^T \Psi$, $G_\Upsilon^\Phi = \Upsilon^T \Phi$

② Comput $T_2 = (HK + HU + \lambda m I)^{-1} H$

③ Compute $T_1 = T_2 G_\Upsilon^\Phi$

Experiments:

Experimet on Two dynamical systems

① a synthetic dataset generated from a linear ~~[illegible]~~ dynamical sys.

② Camera Tracking Problem

---

① Synthetic data exp:

a particle rotating around the origin:

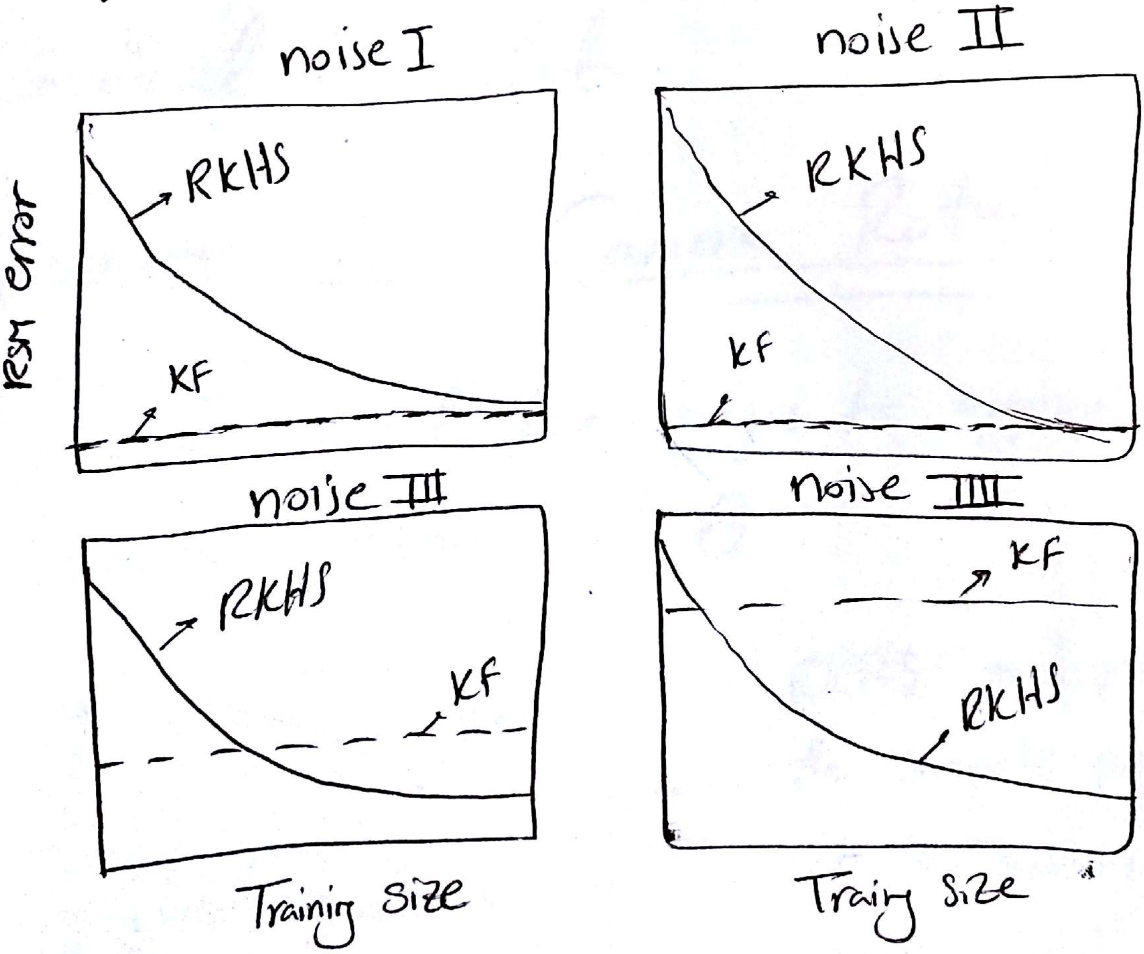
$$s^{t+1} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} s^t + \xi \rightarrow \text{Process noise}$$

$$o^{t+1} = s^{t+1} + \eta \rightarrow \text{observation noise}$$

$\theta = 0.02$

goal: Compare the performance of ~~our~~ the article method with the Kalman filter in estimating the position of particle.

Four Types of Process noise and Observation noise were tested. The result of KF & RKHS were Compared:



nonliner

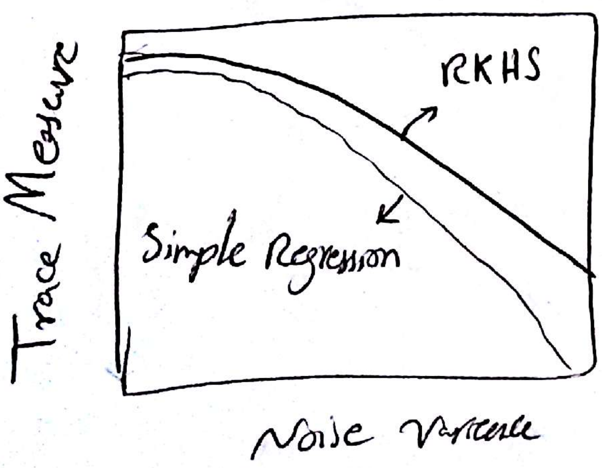
for

Result:

in cases when kalman filter is optimal (Case I, II) the performance of the RKHS method approaches the performance of KF with higher training data However for nonlinear systems (Case III & ~~I~~III) ~~KF~~ RKHS out perform KF significantly with sufficient training data.

Experiment 2 Camera Rotation

used Cornell box images to approximate camera rotation angle θ.



RKHS outperforms the simple regression in estimating Camera rotation