Galyna Kriukova

Kernel methods for Hidden Markov Model

National University of "Kyiv-Mohyla Academy", Kyiv, Ukraine.

Development of proper models for time series of stochastic semi-observable processes is crucial for solving a wide variety of problems in learning theory. Most of the observed data from system does not depict the true states but rather noisy variates of them. Moreover, the observed state space is generally only a subset of the true state space, as the sensory equipment of most systems is limited.

Hidden Markov model (HMM) is widely used probabilistic graphical model for time series of discrete, partially observable stochastic processes. It makes the basic assumptions that a fixed number of preceding hidden states suffices to reason about the current hidden state (Markovian property) and that a certain observation is conditionally dependent solely on its corresponding hidden state. At its core, the HMM evolves its belief over the state by advancing it according to the system dynamics (transition model) and by conditioning on new observations (observation model). Accordingly, HMM has a bunch of disadvantages, among which large number of unstructured parameters, limitations caused by markov property for first order HMMs, and the most critical is that only a small portion of distribution can be represented by HMM due to assumption of discrete number of hidden states.

A well established concept that extends the ideas of HMMs to continuous domains is the Kalman filter (KF), which assumes linear system dynamics and represents the state as a Gaussian random variable. Considering of non-linear system dynamics by means of its sequential linearization leads to Extended Kalman filter (EKF), notwithstanding, assuming zero mean multivariate Gaussian noises for transition and observation models. As further step to address complex problems with non-linear models and non-Gaussian noise, the particle filter has been proposed. Particle filter is a technique for implementing recursive Bayesian filter by Monte Carlo sampling representing the posterior density by a set of random particles with associated weights, thereby, estimates are computed based on these samples and weights. Despite of its undoubted ability to represent arbitrary densities and deal with non-Gaussian noise, there is list of disadvantages of particle filter, among which high computational

complexity, difficulties while determining optimal number of particles, number of particles increase with increasing model dimention, a vital role of proper importance density choice, and necessity of resampling to avoid potential risk of degeneracy and loss of diversity. Various modifications of particle filter have been proposed, nevertheless, there is still research challenge to develop optimal algorithm with reduced complexity.

In our study we propose a nonparametric HMM that extends traditional HMMs to structured and non-Gaussian continuous distributions by means of embedding HMM into Reproducing Kernel Hilbert Space (RKHS).

Definition 1. RKHS is a Hilbert space of functions $f : \Omega \to \mathbb{R}$ with a scalar product $\langle \cdot, \cdot \rangle$ that is implicitly defined by Mercer kernel function $k : \Omega \times \Omega \to \mathbb{R}$ as $\langle \varphi(x), \varphi(y) \rangle = k(x, y)$, where $\varphi(x)$ is a feature mapping into space corresponding to the kernel function. According to reproducing property $\forall x \in \mathcal{X}, \forall f \in \mathcal{H} \rangle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ we have for any element f from RKHS $f(y) = \sum_{i \in I} \alpha_i k(x_i, y)$, $\alpha_i \in \mathbb{R}$.

Embedding distribution functions into reproducing kernel Hilbert spaces leverages the generalization of machine learning methods to arbitrary probability densities, not only Gaussian ones, by providing a uniform representation of functions and, thus, probability densities as elements of a RKHS. Joint and conditional distributions may be embedded into a RKHS and manipulate the probability densities, by means of the chain, sum and Bayes' rule, entirely in Hilbert space.

Remark 1. Given a set of feature mappings $\Phi = [\varphi(x_1), \ldots, \varphi(x_m)]$ any distribution $q(x)$ may be embedded as a linear combination $\hat{\mu}_q = \Phi\beta$, with weight vector $\beta \in \mathbb{R}^m$. The mean embedding of a distribution can be used to evaluate expectation of any function f in the RKHS, e.g. if $f = \Phi\alpha$, then

$$\mathbb{E}_q[f(x)] = \langle \hat{\mu}_q, f \rangle = \langle \Phi\beta, \Phi\alpha \rangle = \beta^\top \Phi^\top \Phi\alpha = \beta^\top \mathsf{K}\alpha,$$

where $\mathsf{K} = \Phi^\top \Phi$ is Gramian matrix, $\mathsf{K}_{ij} = k(x_i, x_j)$.

Theorem 1 ([3]). *Assume $k(x, x')$ is bounded. With probability $1 - \delta$*

$$||\hat{\mu}_q - \mu_q||_{\mathcal{H}} = O\left(m^{-1/2}\sqrt{-\log \delta}\right).$$

Now we are ready to consider RKHS embedding for HMM.

Definition 2. Assuming RKHS $\mathcal{F}$ with kernel $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$ defined on the observations, and RKHS $\mathcal{G}$ with kernel $l(h, h') = \langle \phi(h), \phi(h') \rangle_{\mathcal{G}}$ defined on the hidden states, *observable operator* $\mathcal{A}_x : \mathcal{G} \to \mathcal{G}$ is defined as

$$\mathcal{A}_x \phi(h_t) = p(X_t = x | h_t) \mathbb{E}_{H_{t+1}|h_t}[\phi(H_{t+1})].$$

The *observation operator* is defined as a conditional operator $\mathcal{C}_{X_{t+1}|H_{t+1}} = \mathcal{C}_{X_t|H_t}$ that maps distribution function over hidden states embedded into $\mathcal{G}$ to a distribution function over emissions embedded in $\mathcal{F}$.

Straightforward from Theorem 1 we have

Corollary 2. *Assume $k(x, x')$ and $l(x, x')$ are bounded. Then with probability $1 - \delta$*

$$||\hat{\mathcal{C}}_{XY} - \mathcal{C}_{XY}||_{\mathcal{F} \otimes \mathcal{G}} = O\left(m^{-1/2} \sqrt{-\log \delta}\right).$$

For conditional embedding operator use of regularization is needed. Thus, for Tikhonov regularization and given regularization parameter λ we have

Corollary 3. *Assume $k(x, x')$ and $l(x, x')$ are bounded. Then with probability $1 - \delta$*

$$||\hat{\mu}_{Y|x} - \mu_{Y|x}||_{\mathcal{G}} = O\left(\sqrt{\lambda} + \sqrt{\frac{-\log \delta}{\lambda m}}\right).$$

Appropriate value of regularization parameter λ may be selected by means of classical approaches, such as Morozov's discrepancy principle, or using Linear Functional Strategy considered in [1].

We evaluate our approach on seizure prediction on EEG signal. For more details on experiment setting we refer to [5].

[1] Kriukova G. A linear functional strategy for regularized ranking / Kriukova G., Panasiuk V., Pereverzyev S.V., Tkachenko P. // Neural Networks. — Vol. 73. — 2016. — P. 26–35.

[2] Kriukova G. Nyström type subsampling analyzed as a regularized projection / Kriukova G., Pereverzyev S., Tkachenko P. // Inverse Problems. Special issue on learning and inverse problems. — Vol. 33. — Num. 7. — 2017. — P. 074001.

[3] Smola A. A Hilbert space embedding for distributions / Smola, A., Gretton, A., Song, L., Schölkopf, B. // Algorithmic Learning Theory, Lecture Notes on Computer Science. Springer. — 2007.

[4] Song L. Hilbert space embeddings of conditional distributions / Song, L., Huang, J., Smola, A., Fukumizu, K. // International Conference on Machine Learning. — 2009.

[5] Sudakov O. Distributed System for Sampling and Analysis of Electroencephalograms / Sudakov O., Kriukova G., Natarov R., Gaidar V., Maximyuk O., Radchenko S., Isaev D. // The 9-th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications. — Vol. 1. — 2017. — P. 306–310.
[6] Vapnik V.N. Statistical Learning Theory // Wiley, New York, 1998.

E-mail: ✉ *kriukovagv@ukma.edu.ua.*