Linear-time Learning on Distributions with Approximate Kernel Embeddings

Dougal J. Sutherland*, Junier B. Oliva*, Barnabás Póczos, and Jeff Schneider
*: Equal contribution.
{dsutherl,joliva,bapoczos,schneide}@cs.cmu.edu

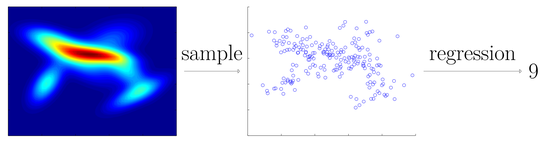
Auton Lab

Carnegie Mellon University

Learning on Distributions

We want to do machine learning where the inputs are distributions, instead of fixed-dimensional vectors:

- Predict galaxy cluster mass from galaxy velocities [5].
- Model political feeling with sets of individual demographics [3].
- Classify images with distributions of patch descriptors [4, 7].
- Approximate expectation propagation messages [2].
- Perform parametric statistical inference faster and/or better:



Distribution Kernels

Use a distributional distance d in a generalized RBF kernel

$$K(p,q) = \exp\left(-\frac{1}{2\sigma^2} d^2(p,q)\right)$$

where p, q are densities on $\mathcal{X}$.

Then estimate K based on samples $X \sim p$, $Y \sim q$.

But doing pairwise kernel estimates means with N input distributions, we need N^2 estimates, and usually N^3 computation in kernelized SVMs, GPs, ...

Approximate Kernel Embeddings

Instead, we can find $z : \mathcal{X} \to \mathbb{R}^D$ so $K(p,q) \approx z(X)^\mathsf{T} z(Y)$:

$$K(\,,\,) \approx z(\,)^\mathsf{T} z(\,)$$

Previous work has developed embeddings for kernels based on L_2 distance [6] and mean map distances [2, 3, others]. We want to consider more possibilities for the distance.

Our Contribution

We give a new embedding z for kernels based on a class of distances including total variation, Jensen-Shannon, and Hellinger distances.

This work was funded in part by NSF grant IIS1247658 and DARPA grant FA87501220324. DJS is also supported by a Sandia Campus Executive Program fellowship.

Homogeneous Density Distances

We'll embed *homogeneous density distances* (HDDs):

$$d^2(p,q) = \int_{\mathcal{X}} \kappa(p(x), q(x))\, \mathrm{d}x$$

where $\kappa(x,y) : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ satisfies:

- κ is a negative-type kernel (a squared Hilbertian metric).
- $\kappa(tx, ty) = tk(x,y)$ for all $t > 0$.

Name	$\kappa(p(x), q(x))$	$\mathrm{d}\mu(\lambda)$
Jensen-Shannon	$\Sigma_{r\in\{p,q\}} \frac{1}{2} r(x) \log\left(\frac{2r(x)}{p(x)+q(x)}\right)$	$\frac{\mathrm{d}\lambda}{\cosh(\pi\lambda)(1+\lambda^2)}$
Squared Hellinger	$\frac{1}{2}\left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2$	$\frac{1}{2}\delta(\lambda = 0)\,\mathrm{d}\lambda$
Total Variation	$\lvert p(x) - q(x)\rvert$	$\frac{2}{\pi}\frac{1}{1+4\lambda^2}\,\mathrm{d}\lambda$

We'll embed:

1. d into $L_2(\mathcal{X})^{2M}$ (based on [1]).
2. $L_2(\mathcal{X})$ into $\mathbb{R}^{|V|}$ (using projection coefficients).
3. RBF kernel on $\mathbb{R}^{2M|V|}$ into $\mathbb{R}^D$ (random Fourier features).

1: Embedding HDDs into L_2

First, embed κ into L_2: κ corresponds to measure $\mu(\lambda)$ so that with $Z = \mu(\mathbb{R}_{\geq 0})$, $g_\lambda(x) = \sqrt{Z}\frac{-\frac{1}{2}+\mathrm{i}\lambda}{\frac{1}{2}+\mathrm{i}\lambda}\left(x^{\frac{1}{2}+\mathrm{i}\lambda} - 1\right)$ [1]:

$$\kappa(x,y) = \mathbb{E}_{\lambda\sim\frac{\mu}{Z}} \left|g_\lambda(x) - g_\lambda(y)\right|^2 \approx \frac{1}{M}\sum_{j=1}^{M} \left|g_{\lambda_j}(x) - g_{\lambda_j}(y)\right|,$$

sampling $\{\lambda_j\}_{j=1}^M \overset{iid}{\sim} \frac{\mu}{Z}$. Then

$$\begin{aligned} d^2(p,q) &= \int_{\mathcal{X}} \kappa(p(x), q(x))\, \mathrm{d}x \\ &\approx \frac{1}{M}\sum_{j=1}^{M} \int_{\mathcal{X}} \left|g_{\lambda_j}(p(x)) - g_{\lambda_j}(q(x))\right|^2 \mathrm{d}x \\ &= \frac{1}{M}\sum_{j=1}^{M} \|p^R_{\lambda_j} - q^R_{\lambda_j}\|^2_{L_2} + \|p^I_{\lambda_j} - q^I_{\lambda_j}\|^2_{L_2} \end{aligned}$$

where $p^R_\lambda(x) = \Re(g_\lambda(p(x)))$, $p^I_\lambda(x) = \Im(g_\lambda(p(x)))$.

2: Embedding L_2 into $\mathbb{R}^{|V|}$

Let $\{\varphi_\alpha(x)\}_{\alpha\in\mathcal{Z}^d}$ be an orthonormal basis for $L_2(\mathcal{X})$, and $a_\alpha(f) = \langle\varphi_\alpha, f\rangle = \int_{\mathcal{X}} \varphi_\alpha(x) f(x)\,\mathrm{d}x$. Then:

$$\begin{aligned} \langle f, g\rangle &= \left\langle \sum_{\alpha\in\mathcal{Z}^d} a_\alpha(f)\varphi_\alpha, \sum_{\beta\in\mathcal{Z}^d} a_\beta(g)\varphi_\alpha \right\rangle = \sum_{\alpha\in\mathcal{Z}^d} a_\alpha(f) a_\alpha(g) \\ &\approx \sum_{\alpha\in V} a_\alpha(f) a_\alpha(g) = \vec{a}(f)^\mathsf{T}\vec{a}(g), \end{aligned}$$

and $d^2(p,q) \approx \|A(p) - A(q)\|^2$, where

$$A : L_2(\mathcal{X}) \to \mathbb{R}^{2M|V|} \quad p \mapsto A(p) = \tfrac{1}{\sqrt{M}}\left(\vec{a}(p^R_{\lambda_1}), \vec{a}(p^I_{\lambda_1}), \dots\right).$$

3: Embedding RBF kernels on $\mathbb{R}^m$ into $\mathbb{R}^D$

We'll use method of [8]: sample $\{\omega_i\}_{i=1}^{D/2} \overset{iid}{\sim} \mathcal{N}(0, \sigma^{-2} I_m)$. Define

$$z : \mathbb{R}^m \to \mathbb{R}^D \quad x \mapsto z(x) = \sqrt{\frac{2}{D}}\left(\sin(\omega_1^\mathsf{T} x), \cos(\omega_1^\mathsf{T} x), \dots\right).$$

Then $z(x)^\mathsf{T} z(y) \approx \exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right)$, and so

$$K(p,q) \approx z(A(p))^\mathsf{T} z(A(q)).$$

Finite Sample Estimates

Estimate $\vec{a}(p^R_\lambda)$, $\vec{a}(p^I_\lambda)$ by first getting $\hat{p}_\lambda$ with kernel density estimation, and then Monte Carlo numerical integration:

$$\{u_i\}_{i=1}^{n_e} \overset{iid}{\sim} \mathrm{Unif}(\mathcal{X}); \quad \hat{a}_\alpha(\hat{p}^S_{\lambda_j}) = \frac{1}{n_e}\sum_{i=1}^{n_e} \varphi_\alpha(u_i)\,\hat{p}^S_{\lambda_j}(u_i).$$

Theory

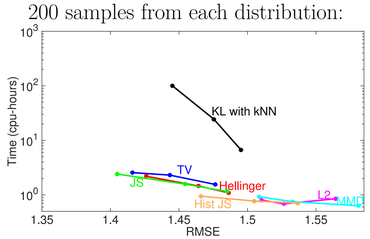
For fixed p and q, the paper bounds

$$\Pr\left(\left|K(p,q) - z(\hat{A}(\hat{p}))^\mathsf{T} z(\hat{A}(\hat{q}))\right| \geq \varepsilon\right).$$

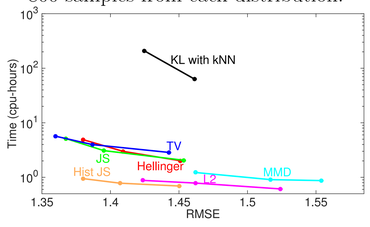
Experiment: Mixture Components

Generate Gaussian mixtures with 1 to 10 components; train on 4 000, 8 000, and 16 000 distributions, test on 2 000.

200 samples from each distribution:



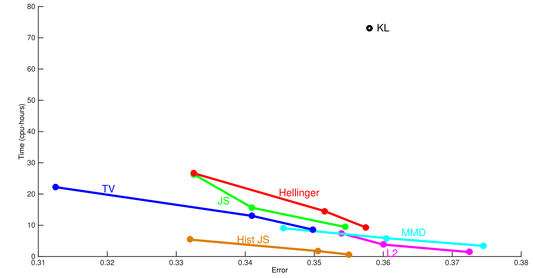
800 samples from each distribution:



AIC, BIC barely better than predicting the mean (RMSE 2.8).

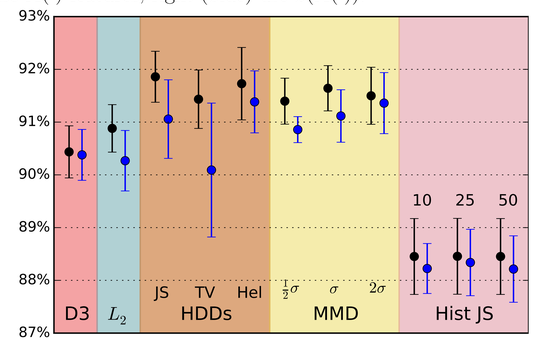
Experiment: Image Classification

CIFAR-10 cats versus dogs, pixel-level features; train on 2 500, 5 000, and 10 000 images, test on 2 000.



Experiment: Scene Classification

Scene-15 dataset, 10 random splits, features from the last convolutional layer of `imagenet-vgg-verydeep-16`. Left (black) use $\hat{A}(\cdot)$ features, right (blue) use $z(\hat{A}(\cdot))$.



[1] Fuglede. Spirals in Hilbert space: With an application in information theory. Expositiones Mathematicae, 23(1), 23–45 (2005).

[2] Jitkrittum, Gretton, Heess, et al. Kernel-Based Just-In-Time Learning for Passing Expectation Propagation Messages. UAI 2015.

[3] Flaxman, Wang, and Smola. Who Supported Obama in 2012? Ecological Inference through Distribution Regression. KDD 2015.

[4] Muandet, Fukumizu, Dinuzzo, and Schölkopf. Learning from distributions via support measure machines. NIPS 2012.

[5] Ntampaka, Trac, Sutherland, et al. A Machine Learning Approach for Dynamical Mass Measurements of Galaxy Clusters. ApJ 803(2), 50 (2015).

[6] Oliva, Neiswanger, Póczos, et al. Fast Distribution To Real Regression. AISTATS 2014.

[7] Póczos, Xiong, Sutherland, and Schneider. Nonparametric Kernel Estimators for Image Classification. CVPR 2012.

[8] Rahimi and Recht. Random Features for Large-Scale Kernel Machines. NIPS 2007.