Simple Consistent Distribution Regression on Compact Metric Domains

Zoltán Szabó[1], Arthur Gretton[1], Barnabás Póczos[2], Bharath K. Sriperumbudur[3]

[1]Gatsby Unit, University College London [2]Machine Learning Department, Carnegie Mellon University
[3]Department of Pure Mathematics and Mathematical Statistics, University of Cambridge

Problem

- Distribution regression, with two-stage sampling [1]:
  - Input = probability measure, output = real number, but
  - we only have samples from the input distributions.
- Covered machine learning tasks include:
  - multiple instance (MI) learning (MIL),
  - point estimates of statistics (e.g., entropy or a hyperparameter).
- Existing methods: heuristics, or require density estimation (which typically scale poorly in dimension).

Contribution

- We study an alternative, simple method: embed the distributions to a RKHS, then apply ridge regression.
- Results:
  - Consistency, convergence rate $\xrightarrow{\text{specially}}$
  - Set kernels [2, 3] are consistent in regression (15-year-old open problem).

Introduction

Existing heuristics:

- parametric model fitting, kernelized Gaussian divergences, kernels on distributions, Rényi-, Tsallis divergence, MIL (classification: PAC; set (semi)metric).
- issues: parameterization may fail to hold; metric/kernel? consistent estimation? consistency in learning tasks?

Theoretically justified methods [1, 4]:

- require density estimation (often poor scaling).
- assume density, compact Euclidean domain.

Distribution Regression

$\mathcal{M}_1^+(\mathfrak{X})$ (Borel) probability measures on domain $\mathfrak{X}$.

- $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^l$: $(x_i, y_i) \in \mathcal{M}_1^+(\mathfrak{X}) \times \mathbb{R}$, i.i.d.
- Given: $\hat{\mathbf{z}} = \{(\{x_{i,n}\}_{n=1}^N, y_i)\}_{i=1}^l$, where $x_{i,1}, \ldots, x_{i,N} \overset{i.i.d.}{\sim} x_i$.

- **Goal**: learn the relation between (x, y) given $\hat{\mathbf{z}}$.
- **Idea**:

$$\mathcal{M}_1^+(\mathfrak{X}) \xrightarrow{\mu} X (\subseteq H) \xrightarrow{f \in \mathcal{H} = \mathcal{H}(K)} \mathbb{R},$$

  i.e., embed the distributions to a $H = H(k)$ =RKHS(k) on $\mathfrak{X}$, then $H \to \mathbb{R}$ ridge regression.
- **Notations**: k is a kernel on $\mathfrak{X}$, mean embedding

$$\mu_x = \int_{\mathfrak{X}} k(\cdot, u) \mathrm{d}x(u) = \mathbb{E}_{u \sim x}[k(\cdot, u)], \quad X = \mu\left(\mathcal{M}_1^+(\mathfrak{X})\right).$$

  $\rho(\mu_x, y) = \rho(y|\mu_x)\rho_X(\mu_x)$, regression function of ρ, $\|\cdot\|_\rho$, T:

$$f_\rho(\mu_a) = \int_{\mathbb{R}} y \mathrm{d}\rho(y|\mu_a), \|f\|_\rho^2 = \int_X |f(\mu_a)|^2 \mathrm{d}\rho_X(\mu_a),$$
$$(Tg)(\mu_a) = \int_X K(\mu_a, \mu_b) g(\mu_b) \mathrm{d}\rho_X(\mu_b), \quad T: L^2_{\rho_X} \to L^2_{\rho_X}.$$

Objective Function, Algorithm

- **Cost function** (of MERR):

$$f_{\hat{\mathbf{z}}}^\lambda = \arg\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^l [f(\mu_{\hat{x}_i}) - y_i]^2 + \lambda \|f\|_{\mathcal{H}}^2 \quad (\lambda > 0),$$

  where $\hat{x}_i = \frac{1}{N}\sum_{n=1}^N \delta_{x_{i,n}}$ is the i^{th} empirical distribution.
- Analytical **solution**: prediction on a new distribution t

$$(f_{\hat{\mathbf{z}}}^\lambda \circ \mu)(t) = [y_1, \ldots, y_l](\mathbf{K} + l\lambda \mathbf{I}_l)^{-1}\mathbf{k},$$
$$\mathbf{K} = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})], \ \mathbf{k} = [K(\mu_{\hat{x}_1}, \mu_t); \ldots; K(\mu_{\hat{x}_l}, \mu_t)].$$

Consistency Theorem

If (i) $\mathfrak{X}$: compact metric, (ii) k: continuous, (iii), $\mu: \mathcal{M}_1^+(\mathfrak{X}) \to H(k)$ measurable ($\Leftarrow k$: universal), (iv) $\Psi_K(\mu_c) := K(\cdot, \mu_c): X \to \mathcal{H}$ is Hölder continuous, (v) $supp(\rho_X) = X$, (vi) $f_\rho \in Im(T^s)$; $\frac{1}{2} < s \leq \frac{3}{2}$, then with high probability

$$\left\|f_{\hat{\mathbf{z}}}^\lambda - f_\rho\right\|_{\mathcal{H}} \precsim \frac{\log^{\frac{h}{2}}(l)}{N^{\frac{h}{2}}\lambda^2} + \frac{1}{\sqrt{l}\lambda} + \lambda^{s - \frac{1}{2}}.$$

For suitable (l, N, λ) choices r.h.s $\to 0$. Specially, for linear $K(\mu_a, \mu_b) = \langle \mu_a, \mu_b \rangle_H$ we get the consistency of the set kernel:

$$K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j}) = \frac{1}{N^2} \sum_{n,m=1}^N k(x_{i,n}, x_{j,m}).$$

Applications

- Supervised entropy learning: MERR is more precise than the only theoretically justified method [1] (by avoiding density estimation).
- Aerosol prediction based on multispectral satellite images: compares favourably to domain-specific, engineered methods (beating state-of-the art MI techniques).

References

[1] Barnabás Póczos, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman. Distribution-free distribution regression. *AISTATS; JMLR W&CP*, 31:507–515, 2013.

[2] David Haussler. Convolution kernels on discrete structures. Technical report, Department of Computer Science, University of California at Santa Cruz, 1999.

[3] Thomas Gärtner, Peter A. Flach, Adam Kowalczyk, and Alexander Smola. Multi-instance kernels. In *ICML*, pages 179–186, 2002.

[4] Junier B. Oliva, Willie Neiswanger, Barnabás Póczos, Jeff Schneider, and Eric Xing. Fast distribution to real regression. *AISTATS; JMLR W&CP*, 33:706–714, 2014.

SAHD-2014, London, UK. Acknowledgements: This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350.