Distribution Regression – Make It Simple and Consistent

Zoltán Szabó¹, Bharath K. Sriperumbudur², Barnabás Póczos³, Arthur Gretton¹
¹Gatsby Unit, UCL ²Department of Statistics, PSU ³Machine Learning Department, CMU

Problem

- Distribution regression:
  - Input: distribution, output $\in \mathbb{R}/\mathbb{R}^d$/separable Hilbert space.
  - Challenge: sampled input distributions.
- Examples:
  - multiple instance learning (MIL),
  - point estimates of statistics (entropy/hyperparameter/...).
- Existing methods: heuristics, or require density estimation (which typically scale poorly in dimension).

Distribution Regression

- $D(\mathcal{X})$ distributions on domain $(\mathcal{X}, k)$.
- $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m \stackrel{i.i.d.}{\sim} \mathcal{M}$: $(x_i, y_i) \in D(\mathcal{X}) \times Y$.
- Given: $\hat{\mathbf{z}} = \{(x_i,n)_{n=1}^N, y_i\}_{i=1}^m$, where $\{(x_i,n)_{n=1}^N, i.i.d., x_i\}$.

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

Idea: $D(\mathcal{X}) \Rightarrow X \subseteq H(k) \stackrel{f \in \mathcal{K}(k)}{\Rightarrow} Y$.

Mean embedding: $\mu_x = \int_X k(., u) d\pi(x)$.

Objective Function, Algorithm

- Cost function (of MERR): $f^*_2 = \arg\min_{f \in \mathcal{K}} \frac{1}{m} \sum_{i=1}^m \|f(\hat{\mu}_x) - y_i\|^2_2 + \lambda \|f\|^2_{\mathcal{H}}$ ($\lambda > 0$).
- Analytical solution: prediction on a new distribution $t$ ($f^*_2 \circ \mu)(t) = k(K + l\mathbf{I})^{-1}(y_1, \ldots, y_m)$.
- Well-specified case ($f_p \in \mathcal{K}$): $f_p$ is ’$c$-smooth’ with ’$b$-decaying covariance operator’ and $l \geq \lambda^{-\frac{c}{2}-1}$, then $\hat{\mathcal{E}}(f^*_2, f_p) \leq \frac{\log^2(b)}{N^2\lambda^2} + \frac{1}{\sqrt{b\lambda}} + \frac{1}{l\lambda^2}$.
- Misspecified case ($f_p \in L^2_p \setminus \mathcal{K}$): $f_p$ is ’$s$-smooth’, $L^2_p$ is separable, and $\|f\| \leq l$, then $\hat{\mathcal{E}}(f^*_2, f_p) \leq \frac{\log^2(l)}{N^2\lambda^2} + \frac{1}{\sqrt{b\lambda}} + \sqrt{\frac{\lambda_{\min}(1, s)}{\lambda}} + \lambda_{\min}(1, s)$.

Distribution Regression – Make It Simple and Consistent

- Trend: $\mathcal{E}_\lambda = \mathcal{E}^0 + \frac{1}{\sqrt{\lambda}}$, where $\mathcal{E}^0 = \frac{1}{\sqrt{\lambda}}$.

Applications

Supervised entropy learning:

- Label = entropy of the distribution represented by a bag.

Aerosol prediction:

- Bag = multispectral satellite image ’pixels’ over an area.
- Label = aerosol value (highly accurate, expensive ground-based instrument).

Performance Guarantees

- Problem: $K(\mu_x, \mu_y) = \langle \mu_x, \mu_y \rangle_H \Rightarrow$ we get the set kernel $K(\mu_x, \mu_y) = \frac{1}{N^2} \sum_{n,m=1}^N k(x_i, n, x_{i,m})$.

Acceptances

- Code: in ITE (https://bitbucket.org/zszoli/ite/).

Acknowledgements

This work was supported by the Gatesy Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350. The work was carried out while Bharath K. Sriperumbudur was a research fellow in the Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge, UK.