Distribution-to-Anything Regression

Zoltán Szabó

Joint work with Arthur Gretton, Barnabás Póczos (CMU), Bharath K. Sriperumbudur (PSU)

Gatsby Unit, Research Talk
September 8, 2014
Outline

- Intuitive problem definition, motivation.
- Previous methods.
- The problem.
- Algorithm, consistency.
Problem: regression from distributions

- Given: $\{(x_i, y_i)\}_{i=1}^l$ samples $\mathcal{H} \ni f = ?$ such that $f(x_i) \approx y_i$.

- Our interest:
  - $x_i$-s are distributions, but (challenge!)
  - only samples are given from $x_i$-s: $\{x_{i,n}\}_{n=1}^N$.
  - $y_i$: could be ’anything’ (scalar, vector, function, ...).
Examples:

- image = set of patches/visual descriptors,
- document = bag of words/sentences/paragraphs,
- molecule = different configurations/shapes,
- group of people on a social network: bag of friendship graphs,
- customer = his/her shopping records.
user = set of trial time-series,
tissue = collection of cells,
web page = its links,
hard-drive = attribute patterns (temperature, . . . ),
video = collection of images.
Several problems are covered in machine learning and statistics:

- multi-instance learning,
- point estimation tasks without analytical formula.
Existing methods: parametric approaches

Idea:
1. compute similarity of distributions or bags of samples,
2. apply the estimated similarities in a learning algorithm.

First approach (parametric):
1. Fit a parametric model to bags.
2. Similarity of bags = that of the estimated parameters.
Typical examples with analytical similarities:

- Gaussians,
- finite mixtures of Gaussians,
- certain members of the exponential family (known log-normalizer, zero carrier measure).

Ref.:
[Jebara et al., 2004, Wang et al., 2009, Nielsen and Nock, 2012].
Existing methods: kernelized Gaussian measures

- Assumption: training distributions are Gaussians in a RKHS.
- Algorithmically appealing:
  - often divergences = function(\(\leq 2\)-order moments) \(\Rightarrow\)
  - easy to kernelize.
- Ref.: [Jebara et al., 2004, Zhou and Chellappa, 2006].
Existing methods: (positive definite) kernels

Include:

- semigroup kernels [Cuturi et al., 2005],
- nonextensive information theoretical kernel constructions [Martins et al., 2009],
- kernels based on special metrics of $\mathbb{R}^\geq 0$ [Hein and Bousquet, 2005].

Intuition (semigroup kernel):

- sum of 2 measures: more concentrated if they overlap.
- value of dispersion: entropy, inverse generalized variance.
Several divergence measures (KL, Rényi, Tsallis, ...) can be written in terms of
\[ D(a, b) = \int p^a(x)q^b(x)dx. \]  
(1)

\[ D(a, b) \] can be consistently estimated (using e.g. kNN-s) [Póczos et al., 2011].

\textit{Not} kernels.
Existing methods: set metric based algorithms

- Hausdorff metric [Edgar, 1995]:

\[
\label{eq:2}
d_H(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\}.
\]

- Metric on compact sets of metric spaces \([(M, d); X, Y \subseteq M]\).
- 'Slight' problem: highly sensitive to outliers.
Hausdorff metric variations:

- ranked- specifically maximal-, minimal Hausdorff metrics [Wang and Zucker, 2000, Wu et al., 2010],
- average Hausdorff metric [Zhang and Zhou, 2009],
- contextual Hausdorff dissimilarity [Chen and Wu, 2012].
Mini summary of the existing methods

- Dates back to [Haussler, 1999, Gärtner et al., 2002].
- There are several multi-instance methods, applications.
Mini summary of the existing methods

- Dates back to [Haussler, 1999, Gärtner et al., 2002].
- There are several multi-instance methods, applications.

One ’small’ open question:

Do any of these techniques make sense?
Existing methods: 'exceptions'


\[ y_i = \max(\mathbb{I}_R(x_{i,1}), \ldots, \mathbb{I}_R(x_{i,N})) \in \{0, 1\}, \]  

(3)

where \( R = \) unknown rectangle.
Existing methods: 'exceptions'


\[ y_i = \max(\mathbb{I}_R(x_{i,1}), \ldots, \mathbb{I}_R(x_{i,N})) \in \{0, 1\}, \quad (3) \]

where \( R = \) unknown rectangle.

- Density based approaches, regression [Póczos et al., 2013, Oliva et al., 2014]:
  - densities live on compact Euclidean domain,
  - density estimation: nuisance step.
Distribution regression: idea

- \( z = \{(x_i, y_i)\}_{i=1}^l : x_i \in M_1^+ (\mathcal{D}), y_i \in Y. \)
- \( \hat{z} = \left\{ \left( \{x_{i,n}\}_{n=1}^N, y_i \right) \right\}_{i=1}^l : x_{i,1}, \ldots, x_{i,N} \sim \text{i.i.d.} x_i. \)
- Goal: learn the relation between \( x \) and \( y \) based on \( \hat{z} \).
- Idea: embed the distributions (\( \mu \)) + apply ridge regression

\[
M_1^+ (\mathcal{D}) \xrightarrow{\mu} X(\subseteq H = H(k)) \xrightarrow{f \in \mathcal{H} = \mathcal{H}(K)} Y.
\]
Given: kernel \( k : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R} \).

Mean embedding of a distribution \( x \in \mathcal{M}_1^+ (\mathcal{D}) \):

\[
\mu_x = \int_{\mathcal{D}} k(\cdot, u) dx(u) \in H(k).
\]  
\( (4) \)

Mean embedding of the empirical distribution \( \tilde{x}_i = \frac{1}{N} \sum_{n=1}^{N} \delta_{x_{i,n}} \in \mathcal{M}_1^+ (\mathcal{D}) \):

\[
\mu_{\tilde{x}_i} = \int_{\mathcal{D}} k(\cdot, u) d\tilde{x}_i(u) = \frac{1}{N} \sum_{n=1}^{N} k(\cdot, x_{i,n}) \in H(k).
\]  
\( (5) \)
Goal: learn an $X = \mu(\mathcal{M}^+_1(\mathcal{D})) \rightarrow Y$ function.

If $Y = \mathbb{R}$:

- We take a $K : X \times X \rightarrow \mathbb{R}$ kernel.
- Example: linear $K$ gives rise to the set kernel

$$K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j}) = \langle \mu_{\hat{x}_i}, \mu_{\hat{x}_j} \rangle_{H(k)} = \frac{1}{N^2} \sum_{n,m=1}^{N} k(x_{i,n}, x_{j,m}).$$  (6)
Ridge regression step

- Goal: learn an \( X = \mu (\mathcal{M}_1^+ (\mathcal{D})) \rightarrow Y \) function.
- If \( Y = \mathbb{R} \):
  - We take a \( K : X \times X \rightarrow \mathbb{R} \) kernel.
  - Example: linear \( K \) gives rise to the set kernel

\[
K (\mu_{\hat{x}_i}, \mu_{\hat{x}_j}) = \langle \mu_{\hat{x}_i}, \mu_{\hat{x}_j} \rangle_{H(k)} = \frac{1}{N^2} \sum_{n,m=1}^{N} k(x_i,n, x_j,m). \tag{6}
\]

- If \( Y \) is separable Hilbert:
  - We consider a \( K : X \times X \rightarrow \mathcal{L}(Y) \) operator-valued kernel.
  - \( K \) uniquely determines an RKHS(\( K \)).
Vector-valued RKHS

Definition:

- A $\mathcal{H} \subseteq Y^X$ Hilbert space of functions is RKHS if

  $$A_{\mu_x,y} : f \mapsto \langle y, f(\mu_x) \rangle_Y \quad (7)$$

  is continuous for $\forall \mu_x \in X, y \in Y$.

- The evaluation functional is continuous in every direction.

Riesz representation theorem $\Rightarrow$

- $\exists K_{\mu_t} \in \mathcal{L}(Y, \mathcal{H})$:

  $$K(\mu_x, \mu_t)(y) = (K_{\mu_t}y)(\mu_x), \quad (\forall \mu_x, \mu_t \in X), \text{ or shortly}$$

  $$K(\cdot, \mu_t)(y) = K_{\mu_t}y, \quad (8)$$

  $\mathcal{H}(K) = \overline{\text{span}}\{K_{\mu_t}y : \mu_t \in X, y \in Y\}. \quad (9)$
Examples ($Y = \mathbb{R}^d$):

1. $K_i : X \times X \to \mathbb{R}$ kernels ($i = 1, \ldots, d$). Diagonal kernel:

$$K(\mu_a, \mu_b) = \text{diag}(K_1(\mu_a, \mu_b), \ldots, K_d(\mu_a, \mu_b)). \quad (10)$$

2. Combination of $D_j$ diagonal kernels [$D_j(\mu_a, \mu_b) \in \mathbb{R}^{r \times r}$, $A_j \in \mathbb{R}^{r \times d}$]:

$$K(\mu_a, \mu_b) = \sum_{j=1}^{m} A_j^* D_j(\mu_a, \mu_b) A_j. \quad (11)$$
Objective function

- \( f_{\mathcal{H}} \in \mathcal{H} = \mathcal{H}(K) \): ideal/optimal in expected risk sense (\( \mathcal{E} \)):

\[
\mathcal{E} [f_{\mathcal{H}}] = \inf_{f \in \mathcal{H}} \mathcal{E}[f] = \inf_{f \in \mathcal{H}} \int_{X \times Y} \| f(\mu_a) - y \|^2_Y \, d\rho(\mu_a, y). \tag{12}
\]

- One-stage difficulty (\( \int \to z \)):

\[
f^\lambda_z = \arg \min_{f \in \mathcal{H}} \left( \frac{1}{l} \sum_{i=1}^{l} \| f(\mu_{x_i}) - y_i \|^2_Y + \lambda \| f \|^2_{\mathcal{H}} \right). \tag{13}
\]

- Two-stage difficulty (\( z \to \hat{z} \)):

\[
f^\lambda_{\hat{z}} = \arg \min_{f \in \mathcal{H}} \left( \frac{1}{l} \sum_{i=1}^{l} \| f(\mu_{\hat{x}_i}) - y_i \|^2_Y + \lambda \| f \|^2_{\mathcal{H}} \right). \tag{14}
\]
Algorithmically: ridge regression $\Rightarrow$ analytical solution

- Given:
  - training sample: $\hat{z}$,
  - test distribution: $t$.

- Prediction:

\[
(f_2^\lambda \circ \mu)(t) = [y_1, \ldots, y_l](K + l\lambda I_l)^{-1}k, \quad (15)
\]

\[
K = [K_{ij}] = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathcal{L}(Y)^{l \times l}, \quad (16)
\]

\[
k = \begin{bmatrix}
K(\mu_{\hat{x}_1}, \mu_t) \\
\vdots \\
K(\mu_{\hat{x}_l}, \mu_t)
\end{bmatrix} \in \mathcal{L}(Y)^l. \quad (17)
\]
We studied

- the excess error: \( \mathcal{E} [f_\hat{\lambda}] - \mathcal{E} [f_{\mathcal{H}}] \), i.e,
- the goodness compared to the best function from \( \mathcal{H} \).

Result: if \( l \geq \lambda^{-\frac{1}{b}-1} \), then with high probability

\[
\mathcal{E} [f_\hat{\lambda}] - \mathcal{E} [f_{\mathcal{H}}] \lesssim \frac{\log^h(l)}{N^h \lambda^3} + \lambda^c + \frac{1}{l^2 \lambda} + \frac{1}{l \lambda^\frac{1}{b}}.
\] (18)

\[ \Rightarrow \text{Consistency for suitable } (l, N, \lambda) \text{ choices.} \]
• $\mathcal{D}$: separable, topological.
• $Y$: separable Hilbert.
• $k$:
  • bounded: $\sup_{u \in \mathcal{D}} k(u, u) \leq B_k \in (0, \infty)$,
  • continuous.
• $\mu : (\mathcal{M}_1^+ (\mathcal{D}), \sigma(\text{weak})) \rightarrow (H, \mathcal{B}(H))$ is measurable.
Assumptions-2

- **K:**
  1. bounded:
    \[
    \|K_{\mu_a}\|_{\text{HS}}^2 = \text{Tr} \left( K_{\mu_a}^* K_{\mu_a} \right) \leq B_K \in (0, \infty), \quad (\forall \mu_a \in X). \tag{19}
    \]
  2. Hölder continuous: \( \exists L > 0, \ h \in (0, 1] \) such that
    \[
    \|K_{\mu_a} - K_{\mu_b}\|_{L(\mathcal{Y},\mathcal{H})} \leq L \|\mu_a - \mu_b\|_H^h, \quad \forall (\mu_a, \mu_b) \in X \times X.
    \]
- \( y \) is bounded: \( \exists C < \infty \) such that \( \|y\|_\mathcal{Y} \leq C \) almost surely.
Let the $T : \mathcal{H} \rightarrow \mathcal{H}$ operator be

$$T = \int_X K(\cdot, \mu_a)K^*(\cdot, \mu_a) d\rho_X(\mu_a)$$

with eigenvalues $t_n$ ($n = 1, 2, \ldots$).

Assumption: $\rho \in \mathcal{P}(b, c) = \text{set of distributions on } X \times Y$

- $\alpha \leq n^b t_n \leq \beta$ ($\forall n \geq 1; \alpha > 0, \beta > 0$),
- $\exists g \in \mathcal{H}$ such that $f_{\mathcal{H}} = T^{\frac{c-1}{2}} g$ with $\|g\|_{\mathcal{H}}^2 \leq R$ ($R > 0$),

where $b \in (1, \infty)$, $c \in [1, 2]$. 

Zoltán Szabó  
Distribution-to-Anything Regression
Notes on the assumptions, examples

(*) := If $\mathcal{D}$ is compact metric, $k$ is universal: $\mu$ continuous.

$Y = \mathbb{R}$: the $K$ requirements simplify to

- $K(\mu_a, \mu_a) \leq B_K$.
- $\|K(\cdot, \mu_a) - K(\cdot, \mu_b)\|_{\mathcal{H}(K)} \leq L \|\mu_a - \mu_b\|_{H(k)}^h$.

Linear $K$: $K(\mu_a, \mu_b) = \langle \mu_a, \mu_b \rangle_H \Rightarrow L = 1, h = 1, B_K = B_k$. 
In case of (*) and $Y = \mathbb{R}$: Hölder $K$-s

<table>
<thead>
<tr>
<th></th>
<th>$K_G$</th>
<th>$K_e$</th>
<th>$K_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$e^{-\frac{</td>
<td></td>
<td>\mu_a - \mu_b</td>
</tr>
<tr>
<td>$h$</td>
<td>$h = 1$</td>
<td>$h = \frac{1}{2}$</td>
<td>$h = 1$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$K_t$</th>
<th>$K_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\left(1 + \frac{</td>
<td></td>
</tr>
<tr>
<td>$h$</td>
<td>$h = \frac{\theta}{2}$ ($\theta \leq 2$)</td>
<td>$h = 1$</td>
</tr>
</tbody>
</table>
Numerical experiences \((Y = \mathbb{R})\)

- **Supervised entropy learning:**
  - more precise than the only theoretically justified method,
  - by avoiding density estimation.
- **Aerosol prediction from satellite images:**
  - \(\approx\) domain-specific, engineered methods,
  - beating state-of-the-art MI techniques.
5 Mar.: Consistent distribution regression via mean embedding. University of Hertfordshire.
5 Mar.: Consistent distribution regression via mean embedding. University of Hertfordshire.

27 Mar.: MERR code made publicly available in ITE (https://bitbucket.org/szzoli/ite/).
5 Mar.: Consistent distribution regression via mean embedding. University of Hertfordshire.
27 Mar.: MERR code made publicly available in ITE (https://bitbucket.org/szzoli/ite/).
2 Apr.: Learning on distributions. Kernel methods for big data workshop (Lille).
5 Mar.: Consistent distribution regression via mean embedding. University of Hertfordshire.

27 Mar.: MERR code made publicly available in ITE (https://bitbucket.org/szzoli/ite/).

2 Apr.: Learning on distributions. Kernel methods for big data workshop (Lille).

2 May: Distribution regression - the set kernel heuristic is consistent. CSML Lunch Talk Series.
5 Mar.: Consistent distribution regression via mean embedding. University of Hertfordshire.

27 Mar.: MERR code made publicly available in ITE (https://bitbucket.org/szzoli/ite/).

2 Apr.: Learning on distributions. Kernel methods for big data workshop (Lille).

2 May: Distribution regression - the set kernel heuristic is consistent. CSML Lunch Talk Series.

4-5 Sept.: Simple consistent distribution regression on compact metric domains. SAHD, London, UK.
Submitted (NIPS): Two-stage Sampled Learning Theory on Distributions.
Submitted (NIPS): Two-stage Sampled Learning Theory on Distributions.

Submitted (NIPS): Two-stage Sampled Learning Theory on Distributions.


Submitted (NIPS): Two-stage Sampled Learning Theory on Distributions.


Problem: two-stage sampled distribution regression.
There exist a large number of heuristics.
Studied algorithm:
  - ridge regression,
  - simple, analytical solution.
Contribution:
  - consistency under mild conditions.
  - specially, set kernel is consistent in regr. (15-year-old open question).
Open questions

- Theoretical perspective:
  - Hölder $K$ constructions for $Y \neq \mathbb{R}$.
  - Equivalent/sufficient $\mathcal{P}(b, c)$ characterizations.
  - Alternative priors ($\rho$), discrepancy criteria ($\mathcal{E}$).

- Algorithmic question:
  - $\text{dim}(Y) < \infty$: large-scale solvers (Dino),
  - $\text{dim}(Y) = \infty$: op-MKL?

- Applications: functional outputs (H. Kadri).
Thank you for the attention!
Given: $\mathcal{D} \neq \emptyset$ set.

$\tau \subseteq 2^{\mathcal{D}}$ is called a *topology* on $\mathcal{D}$ if:

1. $\emptyset \in \tau$, $\mathcal{D} \in \tau$.
2. Finite intersection: $O_1 \in \tau$, $O_2 \in \tau \Rightarrow O_1 \cap O_2 \in \tau$.
3. Arbitrary union: $O_i \in \tau$ ($i \in I$) $\Rightarrow \bigcup_{i \in I} O_i \in \tau$.

Then, $(\mathcal{D}, \tau)$ is called a *topological space*; $O \in \tau$: *open sets*. 
Topology: examples

- \( \tau = \{\emptyset, \mathcal{D}\} \): indiscrete topology.
- \( \tau = 2^\mathcal{D} \): discrete topology.
- \((\mathcal{D}, d)\) metric space:
  - Open ball: \( B_\varepsilon(x) = \{y \in \mathcal{D} : d(x, y) < \varepsilon\} \).
  - \( O \subseteq \mathcal{D} \) is open if for \( \forall x \in O \ \exists \varepsilon > 0 \) such that \( B_\varepsilon(x) \subseteq O \).
  - \( \tau := \{O \subseteq \mathcal{D} : O \) is an open subset of \( \mathcal{D}\}\).
Given: \((\mathcal{D}, \tau)\). \(A \subseteq \mathcal{D}\) is

- **closed** if \(\mathcal{D}\setminus A \in \tau\) (i.e., its complement is open),
- **compact** if for any family \((O_i)_{i \in I}\) of open sets with \(A \subseteq \bigcup_{i \in I} O_i\), \(\exists i_1, \ldots, i_n \in I\) with \(A \subseteq \bigcup_{j=1}^n O_{i_j}\).

**Closure of** \(A \subseteq \mathcal{D}\):

\[
\bar{A} := \bigcap_{A \subseteq C \text{ closed in } \mathcal{D}} C.
\]  \hspace{1cm} (20)

- \(A \subseteq \mathcal{D}\) is **dense** if \(\bar{A} = \mathcal{D}\).
- \((\mathcal{D}, \tau)\) is **separable** if \(\exists\) countable, dense subset of \(\mathcal{D}\).

Counterexample: \(l^\infty/L^\infty\).
The discrete space

- \((\mathcal{D}, 2^{\mathcal{D}})\): complete metric space.
- Discrete metric (inducing the discrete topology):
  
  $$d(x, y) = \begin{cases} 
  0, & \text{if } x = y \\
  1, & \text{if } x \neq y 
  \end{cases}.$$  

  (21)

- Discrete space: separable \(\iff\) \(|\mathcal{D}|\) is countable.


*International Conference on Artificial Intelligence and Statistics (AISTATS; JMLR W&CP)*, 33:706–714.

*International Conference on Artificial Intelligence and Statistics (AISTATS; JMLR W&CP)*, 31:507–515.


In *SIAM International Conference on Data Mining (SDM)*, pages 430–441.

