Distribution-to-Anything Regression

Zoltán Szabó

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

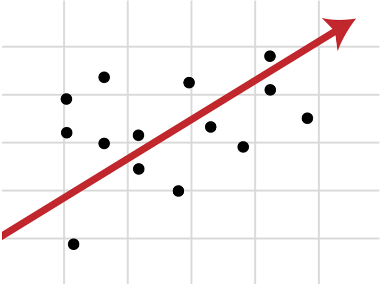
Gatsby Unit, Research Talk
September 8, 2014

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, ...).

Two-stage sampled setting = bag-of-features

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.

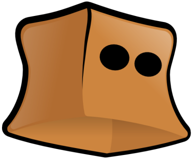
Bag-of-feature representation: further examples

- user $=$ set of trial time-series,
- tissue $=$ collection of cells,
- web page $=$ its links,
- hard-drive $=$ attribute patterns (temperature, $\dots$),
- video $=$ collection of images.

Distribution regression: wider context

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.

Existing methods: parametric approaches-2

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.

Existing methods: consistent divergence estimates

- Several divergence measures (KL, Rényi, Tsallis, ...) can be written in terms of

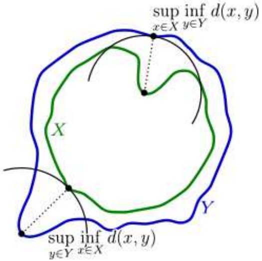
$$D(a,b)=\int p^a(x)q^b(x)\mathrm{d}x. \tag{1}$$

- $D(a,b)$ can be consistently estimated (using e.g. kNN-s) [Póczos et al., 2011].
- *Not* kernels.

Existing methods: set metric based algorithms

- Hausdorff metric [Edgar, 1995]:

$$d_H(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\}. \quad (2)$$



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

Existing methods: set metric based algorithms-2

Hausdorff metric variations:

- ranked- $\xrightarrow{\text{specially}}$ 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:

$\boxed{\text{Do any of these techniques make sense?}}$

Existing methods: 'exceptions'

- APR (axis-parallel rectangles) and its variants, classification [Auer, 1998, Long and Tan, 1998, Blum and Kalai, 1998, Babenko et al., 2011, Zhang et al., 2013, Sabato and Tishby, 2012]:

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

where R = unknown rectangle.

Existing methods: 'exceptions'

- APR (axis-parallel rectangles) and its variants, classification [Auer, 1998, Long and Tan, 1998, Blum and Kalai, 1998, Babenko et al., 2011, Zhang et al., 2013, Sabato and Tishby, 2012]:

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

  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

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

$$M_1^+(\mathcal{D}) \xrightarrow{\mu} X (\subseteq H = H(k)) \xrightarrow{f \in \mathcal{H} = \mathcal{H}(K)} Y.$$

Embedding step: $M_1^+(\mathcal{D}) \xrightarrow{\mu} X \subseteq H(k)$

- Given: kernel $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$.
- Mean embedding of a distribution $x \in \mathcal{M}_1^+(\mathcal{D})$:

$$\mu_x = \int_{\mathcal{D}} k(\cdot, u)\mathrm{d}x(u) \in H(k). \tag{4}$$

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

$$\mu_{\hat{x}_i} = \int_{\mathcal{D}} k(\cdot, u)\mathrm{d}\hat{x}_i(u) = \frac{1}{N}\sum_{n=1}^{N} k(\cdot, x_{i,n}) \in H(k). \tag{5}$$

Ridge regression step

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

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

Ridge regression step

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

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

- If Y is separable Hilbert:
  - We consider a $K : X \times X \to \mathcal{L}(Y)$ 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

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

  is *continuous* for $\forall \mu_x \in X, y \in Y$.
- $=$ The evaluation functional is continuous in every direction.

Riesz representation theorem $\Rightarrow$

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

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

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

$$\mathcal{H}(K) = \overline{span}\{K_{\mu_t} y : \mu_t \in X, y \in Y\}. \tag{9}$$

Vector-valued RKHS – continued

Examples ($Y = \mathbb{R}^d$):

1. $K_i : X \times X \to \mathbb{R}$ kernels ($i = 1, \ldots, d$). Diagonal kernel:
$$K(\mu_a, \mu_b) = diag(K_1(\mu_a, \mu_b), \ldots, K_d(\mu_a, \mu_b)). \tag{10}$$
2. Combination of D_j diagonal kernels [$D_j(\mu_a, \mu_b) \in \mathbb{R}^{r \times r}$, $A_j \in \mathbb{R}^{r \times d}$]:
$$K(\mu_a, \mu_b) = \sum_{j=1}^{m} A_j^* D_j(\mu_a, \mu_b) A_j. \tag{11}$$

Objective function

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

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

- One-stage difficulty ($\int \to \mathbf{z}$):

$$f_{\mathbf{z}}^{\lambda} = \arg\min_{f \in \mathcal{H}} \left(\frac{1}{l} \sum_{i=1}^{l} \|f(\mu_{x_i}) - y_i\|_Y^2 + \lambda \|f\|_{\mathcal{H}}^2 \right). \quad (13)$$

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

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}} \left(\frac{1}{l} \sum_{i=1}^{l} \|f(\mu_{\hat{x}_i}) - y_i\|_Y^2 + \lambda \|f\|_{\mathcal{H}}^2 \right). \quad (14)$$

Algorithmically: ridge regression $\Rightarrow$ analytical solution

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

$$(f_{\hat{\mathbf{z}}}^{\lambda} \circ \mu)(t) = [y_1, \ldots, y_l](\mathbf{K} + l\lambda \mathbf{I}_l)^{-1}\mathbf{k}, \tag{15}$$

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

$$\mathbf{k} = \begin{bmatrix} K(\mu_{\hat{x}_1}, \mu_t) \\ \vdots \\ K(\mu_{\hat{x}_l}, \mu_t) \end{bmatrix} \in \mathcal{L}(Y)^l. \tag{17}$$

Consistency

- We studied
  - the excess error: $\mathcal{E}\left[f_{\hat{\mathbf{z}}}^{\lambda}\right] - \mathcal{E}\left[f_{\mathcal{H}}\right]$, 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

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

- $\Rightarrow$ Consistency for suitable (l, N, λ) choices.

Assumptions-1

- $\mathcal{D}$: separable, topological.
- Y: separable Hilbert.
- k:
  - bounded: $\sup_{u\in\mathcal{D}} k(u,u) \leq B_k \in (0,\infty)$,
  - continuous.
- $\mu : (\mathcal{M}_1^+(\mathcal{D}), \sigma(weak)) \to (H, \mathcal{B}(H))$ is measurable.

Assumptions-2

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

Assumptions-3

- Let the $T : \mathcal{H} \to \mathcal{H}$ operator be

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

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

- Assumption: $\rho \in \mathcal{P}(b, c)$ = set of distributions on $X \times Y$
  - $\alpha \leq n^b t_n \leq \beta \quad (\forall n \geq 1; \alpha > 0, \beta > 0)$,
  - $\exists g \in \mathcal{H}$ such that $f_{\mathcal{H}} = T^{\frac{c-1}{2}} g$ with $\|g\|_{\mathcal{H}}^2 \leq R$ $(R > 0)$,

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

Notes on the assumptions, examples

- (*) $:=$ If $\mathcal{D}$ is compact metric, k is universal: μ 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

K_G	K_e	K_C
$e^{-\frac{\Vert\mu_a-\mu_b\Vert_H^2}{2\theta^2}}$ $h=1$	$e^{-\frac{\Vert\mu_a-\mu_b\Vert_H}{2\theta^2}}$ $h=\frac{1}{2}$	$\left(1+\Vert\mu_a-\mu_b\Vert_H^2/\theta^2\right)^{-1}$ $h=1$

K_t	K_i
$\left(1+\Vert\mu_a-\mu_b\Vert_H^{\theta}\right)^{-1}$ $h=\frac{\theta}{2}$ $(\theta\le 2)$	$\left(\Vert\mu_a-\mu_b\Vert_H^2+\theta^2\right)^{-\frac{1}{2}}$ $h=1$

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.

Dissemination (2014)

- 5 Mar.: Consistent distribution regression via mean embedding. University of Hertfordshire.

Dissemination (2014)

- 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/`).

Dissemination (2014)

- 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).

Dissemination (2014)

- 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.

Dissemination (2014)

- 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.

Dissemination (2014) – continued

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

Dissemination (2014) – continued

- Submitted (NIPS): Two-stage Sampled Learning Theory on Distributions.
- Submitted (UCL Workshop on the Theory of Big Data): Consistent Vector-valued Distribution Regression.

Dissemination (2014) – continued

- Submitted (NIPS): Two-stage Sampled Learning Theory on Distributions.
- Submitted (UCL Workshop on the Theory of Big Data): Consistent Vector-valued Distribution Regression.
- Invited talk: Statistical Science Seminars, Oct 9.

Dissemination (2014) – continued

- Submitted (NIPS): Two-stage Sampled Learning Theory on Distributions.
- Submitted (UCL Workshop on the Theory of Big Data): Consistent Vector-valued Distribution Regression.
- Invited talk: Statistical Science Seminars, Oct 9.
- In preparation (JMLR): Two-Stage Sampled Distribution Regression on Separable Topological Domains: A Simple and Consistent Approach.

Summary

- 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 (ρ), discrepancy criteria ($\mathcal{E}$).
- Algorithmic question:
  - $dim(Y) < \infty$: large-scale solvers (Dino),
  - $dim(Y) = \infty$: op-MKL?
- Applications: functional outputs (H. Kadri).

Thank you for the attention!

Topological space, open sets

- 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: $O_i \in \tau$ $(i \in I) \Rightarrow \cup_{i \in I} O_i \in \tau$.

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: $B_\epsilon(x) = \{y \in \mathcal{D} : d(x,y) < \epsilon\}$.
  - $O \subseteq \mathcal{D}$ is open if for $\forall x \in O$ $\exists \epsilon > 0$ such that $B_\epsilon(x) \subseteq O$.
  - $\tau := \{O \subseteq \mathcal{D} : O \text{ is an open subset of } \mathcal{D}\}$.

Given: $(\mathcal{D},\tau)$. $A \subseteq \mathcal{D}$ is

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

Closure of $A \subseteq \mathcal{D}$:

$$\bar{A} := \bigcap_{A\subseteq C \text{ closed in } \mathcal{D}} C. \tag{20}$$

- $A \subseteq \mathcal{D}$ is *dense* if $\bar{A} = \mathcal{D}$.
- $(\mathcal{D},\tau)$ is *separable* if $\exists$ countable, dense subset of $\mathcal{D}$. Counterexample: l^∞ / L^∞.

The discrete space

- $(\mathcal{D}, 2^{\mathcal{D}})$: complete metric space.
- Discrete metric (inducing the discrete topology):

$$d(x,y) = \begin{Bmatrix} 0, \text{ if } x = y \\ 1, \text{ if } x \neq y \end{Bmatrix}. \tag{21}$$

- Discrete space: separable $\Leftrightarrow$ $|\mathcal{D}|$ is countable.

Auer, P. (1998).
Approximating hyper-rectangles: Learning and pseudorandom sets.
Journal of Computer and System Sciences, 57:376–388.

Babenko, B., Verma, N., Dollár, P., and Belongie, S. (2011).
Multiple instance learning with manifold bags.
In *International Conference on Machine Learning (ICML)*, pages 81–88.

Blum, A. and Kalai, A. (1998).
A note on learning from multiple-instance examples.
Machine Learning, 30:23–29.

Chen, Y. and Wu, O. (2012).
Contextual Hausdorff dissimilarity for multi-instance clustering.
In *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 870–873.

Cuturi, M., Fukumizu, K., and Vert, J.-P. (2005).
Semigroup kernels on measures.
Journal of Machine Learning Research, 6:11691198.

Edgar, G. (1995).
Measure, Topology and Fractal Geometry.
Springer-Verlag.

Gärtner, T., Flach, P. A., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.
In *International Conference on Machine Learning (ICML)*, pages 179–186.

Haussler, D. (1999).
Convolution kernels on discrete structures.
Technical report, Department of Computer Science, University of California at Santa Cruz.
(http://cbse.soe.ucsc.edu/sites/default/files/convoluti

Hein, M. and Bousquet, O. (2005).
Hilbertian metrics and positive definite kernels on probability measures.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 136–143.

Jebara, T., Kondor, R., and Howard, A. (2004).
Probability product kernels.
Journal of Machine Learning Research, 5:819–844.

Long, P. M. and Tan, L. (1998).
PAC learning of axis-aligned rectangles with respect to product distributions from multiple-instance examples.
Machine Learning, 30:7–21.

Martins, A. F. T., Smith, N. A., Xing, E. P., Aguiar, P. M. Q., and Figueiredo, M. A. T. (2009).
Nonextensive information theoretical kernels on measures.
Journal of Machine Learning Research, 10:935–975.

Nielsen, F. and Nock, R. (2012).
A closed-form expression for the Sharma-Mittal entropy of exponential families.
Journal of Physics A: Mathematical and Theoretical, 45:032003.

Oliva, J. B., Neiswanger, W., Póczos, B., Schneider, J., and Xing, E. (2014).
Fast distribution to real regression.
International Conference on Artificial Intelligence and Statistics (AISTATS; JMLR W&CP), 33:706–714.

Póczos, B., Rinaldo, A., Singh, A., and Wasserman, L. (2013).
Distribution-free distribution regression.
International Conference on Artificial Intelligence and Statistics (AISTATS; JMLR W&CP), 31:507–515.

Póczos, B., Xiong, L., and Schneider, J. (2011).
Nonparametric divergence estimation with applications to machine learning on distributions.

In *Uncertainty in Artificial Intelligence (UAI)*, pages 599–608.

Sabato, S. and Tishby, N. (2012).
Multi-instance learning with any hypothesis class.
Journal of Machine Learning Research, 13:2999–3039.

Wang, F., Syeda-Mahmood, T., Vemuri, B. C., Beymer, D., and Rangarajan, A. (2009).
Closed-form Jensen-Rényi divergence for mixture of Gaussians and applications to group-wise shape registration.
Medical Image Computing and Computer-Assisted Intervention, 12:648–655.

Wang, J. and Zucker, J.-D. (2000).
Solving the multiple-instance problem: A lazy learning approach.
In *International Conference on Machine Learning (ICML)*, pages 1119–1126.

Wu, O., Gao, J., Hu, W., Li, B., and Zhu, M. (2010).
Identifying multi-instance outliers.

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

Zhang, D., He, J., Si, L., and Lawrence, R. D. (2013).
MILEAGE: Multiple Instance LEArning with Global Embedding.
International Conference on Machine Learning (ICML; JMLR W&CP), 28:82–90.

Zhang, M.-L. and Zhou, Z.-H. (2009).
Multi-instance clustering with applications to multi-instance prediction.
Applied Intelligence, 31:47–68.

Zhou, S. K. and Chellappa, R. (2006).
From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 28:917–929.