Distribution Regression with Minimax-Optimal Guarantee

Zoltán Szabó (Gatsby Unit, UCL)

Joint work with

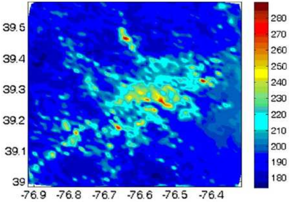
- Bharath K. Sriperumbudur (Department of Statistics, PSU),
- Barnabás Póczos (ML Department, CMU),
- Arthur Gretton (Gatsby Unit, UCL)

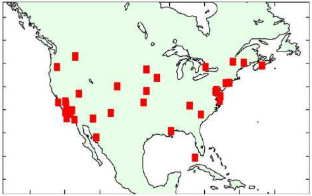
MASCOT-NUM, Toulouse
March 25, 2016

Example: sustainability

- **Goal**: aerosol prediction = air pollution $\rightarrow$ climate.

- Prediction using labelled bags:
    - bag := multi-spectral satellite measurements over an area,
    - label := local aerosol value.

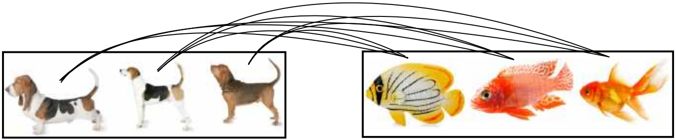




Example: existing methods

Multi-instance learning:

- [Haussler, 1999, Gärtner et al., 2002] (set kernel):



- sensible methods in regression: few,
  1. restrictive technical conditions,
  2. super-high resolution satellite image: would be needed.

One-page summary

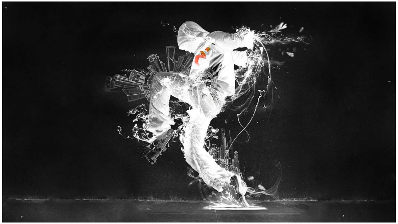
Contributions:

1. Practical: state-of-the-art accuracy (aerosol).
2. Theoretical:
   - General bags: graphs, time series, texts, ...
   - Consistency of set kernel in regression (17-year-old open problem).
   - How many samples/bag?

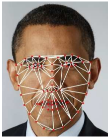
One-page summary

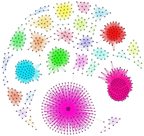
Contributions:

1. Practical: state-of-the-art accuracy (aerosol).
2. Theoretical:
   - General bags: graphs, time series, texts, ...
   - Consistency of set kernel in regression (17-year-old open problem).
   - How many samples/bag?
   - AISTATS-2015 (oral – 6.11%) → JMLR in revision.



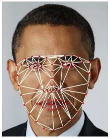
Objects in the bags

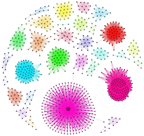




- Examples:
    - time-series modelling: user = set of time-series,
    - computer vision: image = collection of patch vectors,
    - NLP: corpus = bag of documents,
    - network analysis: group of people = bag of friendship graphs, ...

Objects in the bags





- Examples:
  - time-series modelling: user = set of time-series,
  - computer vision: image = collection of patch vectors,
  - NLP: corpus = bag of documents,
  - network analysis: group of people = bag of friendship graphs, . . .
- Wider context (statistics): point estimation tasks.

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{P}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{P}_i$: bag from P_i, $N := |\hat{P}_i|$.
  - test bag: $\hat{P}$.

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{P}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{P}_i$: bag from P_i, $N := |\hat{P}_i|$.
  - test bag: $\hat{P}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} \Big[f\big(\underbrace{\mu_{\hat{P}_i}}_{\text{feature of } \hat{P}_i} \big) - y_i \Big]^2 + \lambda \|f\|_{\mathcal{H}}^2 .$$

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{P}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{P}_i$: bag from P_i, $N := |\hat{P}_i|$.
  - test bag: $\hat{P}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}(K)} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[f\left(\mu_{\hat{P}_i}\right) - y_i \right]^2 + \lambda \|f\|_{\mathcal{H}}^2 .$$

- Prediction:

$$\hat{y}\left(\hat{P}\right) = \mathbf{g}^T (\mathbf{G} + \ell\lambda\mathbf{I})^{-1}\mathbf{y},$$
$$\mathbf{g} = \left[K\left(\mu_{\hat{P}}, \mu_{\hat{P}_i}\right)\right], \mathbf{G} = \left[K\left(\mu_{\hat{P}_i}, \mu_{\hat{P}_j}\right)\right], \mathbf{y} = [y_i].$$

Regression on labelled bags

- Given:
  - labelled bags: $\hat{\mathbf{z}} = \left\{\left(\hat{P}_i, y_i\right)\right\}_{i=1}^{\ell}$, $\hat{P}_i$: bag from P_i, $N := |\hat{P}_i|$.
  - test bag: $\hat{P}$.
- Estimator:

$$f_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{f \in \mathcal{H}(K)} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[f\left(\mu_{\hat{P}_i}\right) - y_i\right]^2 + \lambda \|f\|_{\mathcal{H}}^2 .$$

- Prediction:

$$\hat{y}\left(\hat{P}\right) = \mathbf{g}^T(\mathbf{G} + \ell\lambda\mathbf{I})^{-1}\mathbf{y},$$
$$\mathbf{g} = \left[K\left(\mu_{\hat{P}}, \mu_{\hat{P}_i}\right)\right], \mathbf{G} = \left[K\left(\mu_{\hat{P}_i}, \mu_{\hat{P}_j}\right)\right], \mathbf{y} = [y_i].$$

Challenges

1. Inner product of distributions: $K\left(\mu_{\hat{P}_i}, \mu_{\hat{P}_j}\right) = ?$
2. How many samples/bag?

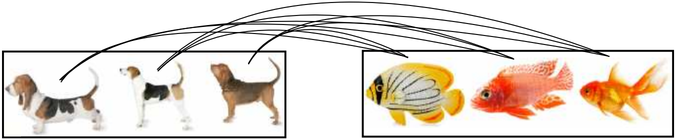
Regression on labelled bags: similarity

Let us define an inner product on distributions $[\tilde{K}(P, Q)]$:

1. Set kernel: $A = \{a_i\}_{i=1}^N$, $B = \{b_j\}_{j=1}^N$.

$$\tilde{K}(A, B) = \frac{1}{N^2} \sum_{i,j=1}^{N} k(a_i, b_j) = \left\langle \underbrace{\frac{1}{N} \sum_{i=1}^{N} \varphi(a_i)}_{\text{feature of bag } A}, \frac{1}{N} \sum_{j=1}^{N} \varphi(b_j) \right\rangle.$$

Remember:



Regression on labelled bags: similarity

Let us define an inner product on distributions [$\tilde{K}(P,Q)$]:

1. Set kernel: $A = \{a_i\}_{i=1}^N$, $B = \{b_j\}_{j=1}^N$.

$$\tilde{K}(A,B) = \frac{1}{N^2}\sum_{i,j=1}^{N} k(a_i,b_j) = \left\langle \underbrace{\frac{1}{N}\sum_{i=1}^{N}\varphi(a_i)}_{\text{feature of bag } A}, \frac{1}{N}\sum_{j=1}^{N}\varphi(b_j)\right\rangle.$$

2. Taking 'limit' [Berlinet and Thomas-Agnan, 2004, Altun and Smola, 2006, Smola et al., 2007]: $a \sim P, b \sim Q$

$$\tilde{K}(P,Q) = \mathbb{E}_{a,b}k(a,b) = \left\langle \underbrace{\mathbb{E}_a\varphi(a)}_{\text{feature of distribution } P =: \mu_P}, \mathbb{E}_b\varphi(b)\right\rangle.$$

Example (Gaussian kernel): $k(\mathbf{a},\mathbf{b}) = e^{-\|\mathbf{a}-\mathbf{b}\|_2^2/(2\sigma^2)}$.

RKHS definition(s)

Given: $\mathcal{D}$ set.

- Kernel: $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.

RKHS definition(s)

Given: $\mathcal{D}$ set.

- Kernel: $k(a, b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.
- RKHS: $H \subset \mathbb{R}^{\mathcal{D}}$ Hilbert space, $\delta_b(f) = f(b)$ is continuous ($\forall b$).

RKHS definition(s)

Given: $\mathcal{D}$ set.

- Kernel: $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.
- RKHS: $H \subset \mathbb{R}^{\mathcal{D}}$ Hilbert space, $\delta_b(f) = f(b)$ is continuous ($\forall b$).
- Reproducing kernel of an $H \subset \mathbb{R}^{\mathcal{D}}$ Hilbert space,
  1. $k(\cdot, b) \in H$,

RKHS definition(s)

Given: $\mathcal{D}$ set.

- Kernel: $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.
- RKHS: $H \subset \mathbb{R}^{\mathcal{D}}$ Hilbert space, $\delta_b(f) = f(b)$ is continuous ($\forall b$).
- Reproducing kernel of an $H \subset \mathbb{R}^{\mathcal{D}}$ Hilbert space,
  1. $k(\cdot, b) \in H$,
  2. $\langle f, k(\cdot, b) \rangle_H = f(b)$. Note: $k(a,b) = \langle k(\cdot, a), k(\cdot, b) \rangle_H$.

RKHS definition(s)

Given: $\mathcal{D}$ set.

- Kernel: $k(a,b) = \langle \varphi(a), \varphi(b) \rangle_{\mathcal{F}}$, $\mathcal{F}$: Hilbert space.
- RKHS: $H \subset \mathbb{R}^{\mathcal{D}}$ Hilbert space, $\delta_b(f) = f(b)$ is continuous $(\forall b)$.
- Reproducing kernel of an $H \subset \mathbb{R}^{\mathcal{D}}$ Hilbert space,
  1. $k(\cdot, b) \in H$,
  2. $\langle f, k(\cdot, b) \rangle_H = f(b)$. Note: $k(a,b) = \langle k(\cdot, a), k(\cdot, b) \rangle_H$.
- $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ sym. is pd. if $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \succeq 0$.

Other valid similarities

Recall: $K(P,Q) = \langle \mu_P, \mu_Q \rangle$.

K_G	K_e	K_C
$e^{-\frac{\left\|\mu_P-\mu_Q\right\|^2}{2\theta^2}}$	$e^{-\frac{\left\|\mu_P-\mu_Q\right\|}{2\theta^2}}$	$\left(1+\left\|\mu_P-\mu_Q\right\|^2/\theta^2\right)^{-1}$

K_t	K_i
$\left(1+\left\|\mu_P-\mu_Q\right\|^\theta\right)^{-1}$	$\left(\left\|\mu_P-\mu_Q\right\|^2+\theta^2\right)^{-\frac{1}{2}}$

Functions of $\|\mu_P - \mu_Q\| \Rightarrow$ computation: similar to set kernel.

Regression on labelled bags: baseline

Quality of estimator, baseline:

$$\mathcal{R}(f) = \mathbb{E}_{(\mu_P,y)\sim\rho}[f(\mu_P) - y]^2,$$
$$f_\rho = \text{best regressor}.$$

How many samples/bag to get the accuracy of f_ρ? Possible?

Assume (for a moment): $f_\rho \in \mathcal{H}(K)$.

Our result: how many samples/bag

- Known [Caponnetto and De Vito, 2007]: best/achieved rate

$$\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f_{\rho}) = \mathcal{O}\left(\ell^{-\frac{bc}{bc+1}}\right),$$

b – size of the input space, c – smoothness of f_{ρ}.

Our result: how many samples/bag

- Known [Caponnetto and De Vito, 2007]: best/achieved rate

$$\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f_{\rho}) = \mathcal{O}\left(\ell^{-\frac{bc}{bc+1}}\right),$$

  b – size of the input space, c – smoothness of f_{ρ}.

- Let $N = \tilde{\mathcal{O}}(\ell^{a})$. N: size of the bags. ℓ: number of bags.

Our result

- If $2 \leq a$, then $f_{\hat{\mathbf{z}}}^{\lambda}$ attains the best achievable rate.

Our result: how many samples/bag

- Known [Caponnetto and De Vito, 2007]: best/achieved rate

$$\mathcal{R}(f_{\mathbf{z}}^{\lambda}) - \mathcal{R}(f_{\rho}) = \mathcal{O}\left(\ell^{-\frac{bc}{bc+1}}\right),$$

  b – size of the input space, c – smoothness of f_{ρ}.
- Let $N = \tilde{\mathcal{O}}(\ell^{a})$. N: size of the bags. ℓ: number of bags.

Our result

- If $2 \leq a$, then $f_{\hat{\mathbf{z}}}^{\lambda}$ attains the best achievable rate.
- In fact, $a = \frac{b(c+1)}{bc+1} < 2$ is enough.
- Consequence: regression with set kernel is consistent.

Why can we get consistency/rates? – intuition

- Convergence of the mean embedding:

$$\|\mu_P - \mu_{\hat{P}}\|_H = \mathcal{O}\left(\frac{1}{\sqrt{N}}\right).$$

- Hölder property of K $(0 < L,\ 0 < h \leq 1)$:

$$\|K(\cdot, \mu_P) - K(\cdot, \mu_{\hat{P}})\|_{\mathcal{H}} \leq L \|\mu_P - \mu_{\hat{P}}\|_H^h.$$

- $f_{\hat{z}}^{\lambda}$ depends 'nicely' on $\mu_{\hat{P}}$. [39 pages]

Extensions

1. Misspecified setting ($f_\rho \in L^2 \backslash \mathcal{H}$):
   - Consistency: convergence to $\inf_{f \in \mathcal{H}} \|f - f_\rho\|_{L^2}$.
   - Smoothness on f_ρ: computational & statistical tradeoff.

Extensions

2. Vector-valued output:
   - Y: separable Hilbert space $\Rightarrow K(\mu_P, \mu_Q) \in \mathcal{L}(Y)$.
   - Prediction on a test bag $\hat{P}$:

$$\hat{y}(\hat{P}) = \mathbf{g}^T(\mathbf{G} + \ell\lambda\mathbf{I})^{-1}\mathbf{y},$$
$$\mathbf{g} = [K(\mu_{\hat{P}}, \mu_{\hat{P}_i})], \mathbf{G} = [K(\mu_{\hat{P}_i}, \mu_{\hat{P}_j})], \mathbf{y} = [y_i].$$

   Specifically: $Y = \mathbb{R} \Rightarrow \mathcal{L}(Y) = \mathbb{R}$; $Y = \mathbb{R}^d \Rightarrow \mathcal{L}(Y) = \mathbb{R}^{d\times d}$.

Aerosol prediction result ($100 \times RMSE$)

We perform on par with the state-of-the-art, hand-engineered method.

- Zhuang Wang, Liang Lan, Slobodan Vucetic. IEEE Transactions on Geoscience and Remote Sensing, 2012: $7.5 - 8.5$ ($\pm 0.1 - 0.6$):
  - hand-crafted features.
- Our prediction accuracy: 7.81 (± 1.64).
  - no expert knowledge.
- Code in ITE: #2 on mloss,

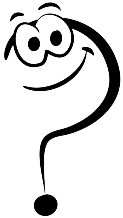
  `https://bitbucket.org/szzoli/ite/`

Summary

- Problem: distribution regression.
- Contribution:
  - computational & statistical tradeoff analysis,
  - specifically, the set kernel is consistent: 17-year-old open question,
  - minimax optimal rate is achievable: sub-quadratic bag size.
- Details (JMLR in revision):

`http://arxiv.org/abs/1411.2066`

Thank you for the attention!



Acknowledgments: This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350. A part of 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.

Altun, Y. and Smola, A. (2006).
Unifying divergence minimization and statistical inference via convex duality.
In *Conference on Learning Theory (COLT)*, pages 139–153.

Berlinet, A. and Thomas-Agnan, C. (2004).
Reproducing Kernel Hilbert Spaces in Probability and Statistics.
Kluwer.

Caponnetto, A. and De Vito, E. (2007).
Optimal rates for regularized least-squares algorithm.
Foundations of Computational Mathematics, 7:331–368.

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/convolutions.pdf).

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007).
A Hilbert space embedding for distributions.
In *Algorithmic Learning Theory (ALT)*, pages 13–31.