Performance guarantees for kernel-based learning on probability distributions

Zoltán Szabó (Gatsby Unit, UCL)

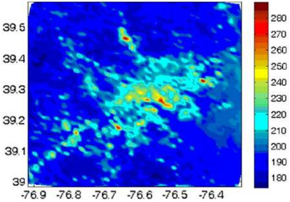
Max Planck Institute for Intelligent Systems
Tübingen

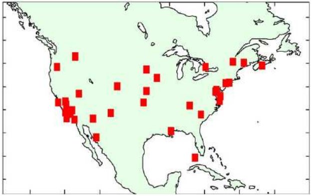
March 16, 2016

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.

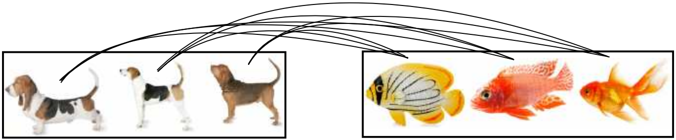




Example: existing methods

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

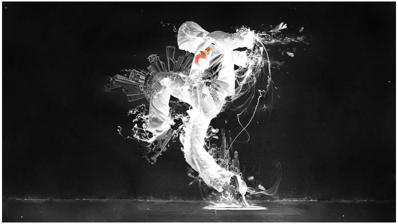
Contributions:

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?

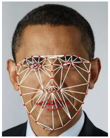
One-page summary

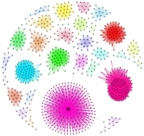
Contributions:

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.



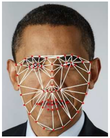
Objects in the bags

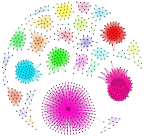




- Examples:
  - time-series modelling: user = set of time-series,
  - computer vision: image = collection of patch vectors,
  - NLP: corpus = bag of documents,
  - network analysis: group of people = bag of friendship graphs, ...

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:

$$w_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{w} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[\left\langle w, \underbrace{\psi(\hat{P}_i)}_{\text{feature of } \hat{P}_i} \right\rangle - y_i\right]^2 + \lambda \|w\|^2 .$$

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:

$$w_{\hat{\mathbf{z}}}^{\lambda} = \arg\min_{w} \frac{1}{\ell} \sum_{i=1}^{\ell} \left[\left\langle w, \psi(\hat{P}_i)\right\rangle - y_i\right]^2 + \lambda \left\|w\right\|^2.$$

- Prediction:

$$\hat{y}(\hat{P}) = \mathbf{g}^T(\mathbf{K} + \ell\lambda\mathbf{I})^{-1}\mathbf{y},$$
$$\mathbf{g} = [K(\hat{P}_i, \hat{P})], \mathbf{K} = [\underbrace{K(\hat{P}_i, \hat{P}_j)}_{:=\left\langle\psi(\hat{P}_i),\psi(\hat{P}_j)\right\rangle}], \mathbf{y} = [y_i].$$

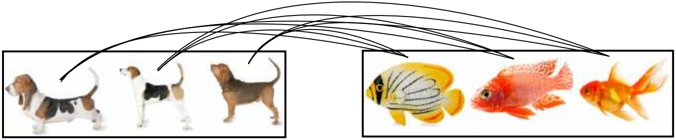
Regression on labelled bags: similarity

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

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

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

Remember:



Regression on labelled bags: similarity

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

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

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

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

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

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

Regression on labelled bags: baseline

Quality of estimator, baseline:

$$\mathcal{R}(w) = \mathbb{E}_{(\psi(Q),y)\sim\rho}[\langle w, \psi(Q)\rangle - y]^2,$$
$$w_\rho = \text{best regressor}.$$

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

Our result: how many samples/bag

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

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

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

Our result: how many samples/bag

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

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

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

Our result

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

Our result: how many samples/bag

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

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

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

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

Our result

- If $2 \leq a$, then $w_{\hat{\mathbf{z}}}^{\lambda}$ attains the best achievable rate.
- In fact, $a = \frac{b(c+1)}{bc+1} < 2$ is enough.
- Consequence: regression with set kernel is consistent.
- The same result holds for Hölder K-s: Gaussian [Christmann and Steinwart, 2010], ...

Aerosol prediction result ($100 \times RMSE$)

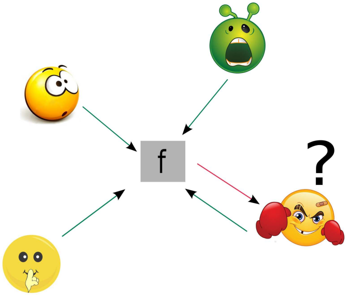
We perform on par with the state-of-the-art, hand-engineered method.

- 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/`

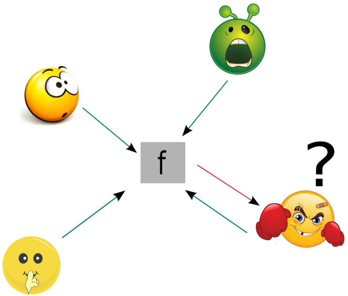
Related results

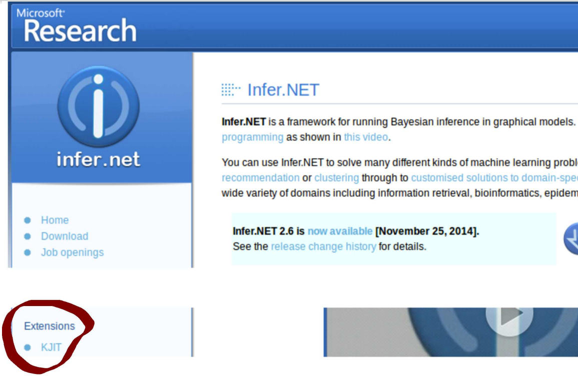
Distribution regression with random Fourier features



- Kernel EP [UAI-2015]:
    - distribution regression phrasing,
    - learn the message-passing operator for 'tricky' factors.

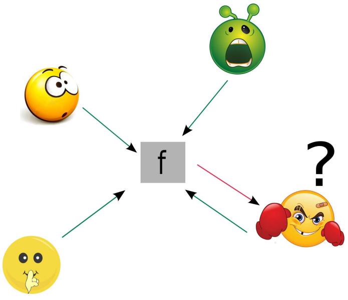
Distribution regression with random Fourier features

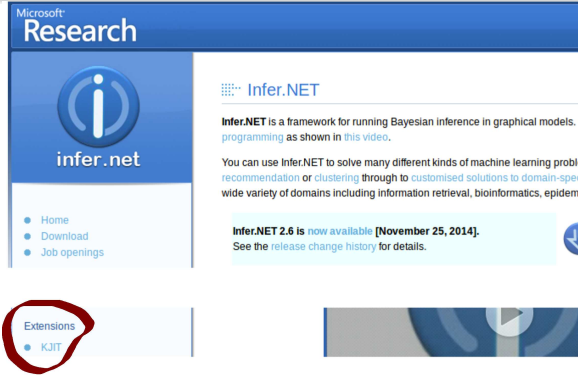




- Kernel EP [UAI-2015]:
  - distribution regression phrasing,
  - learn the message-passing operator for 'tricky' factors.
  - extends Infer.NET; speed $\Leftarrow$ RFF.

Distribution regression with random Fourier features





- Kernel EP [UAI-2015]:
  - distribution regression phrasing,
  - learn the message-passing operator for 'tricky' factors.
  - extends Infer.NET; speed $\Leftarrow$ RFF.
- Random Fourier features [NIPS-2015 (spotlight - 3.65%)]:
  - exponentially tighter guarantee.

+Applications, with Gatsby students

- Bayesian manifold learning [NIPS-2015]:
  - App.: climate data $\rightarrow$ weather station location.
- Fast, adaptive sampling method based on RFF [NIPS-2015]:
  - App.: approximate Bayesian computation, hyperparameter inference.

+Applications, with Gatsby students

- Bayesian manifold learning [NIPS-2015]:
  - App.: climate data → weather station location.
- Fast, adaptive sampling method based on RFF [NIPS-2015]:
  - App.: approximate Bayesian computation, hyperparameter inference.
- Interpretable 2-sample testing [ICML-2016 submission]:
  - App.:
    - random → smart features,
    - discriminative for doc. categories, emotions.
  - empirical process theory (VC subgraphs).

Summary

Regression on

- bags/distributions:
  - minimax optimality,
  - set kernel is consistent.
- random Fourier features: exponentially tighter bounds.

Several applications (with open source code).

Acknowledgments: This work was supported by the Gatsby Charitable Foundation.

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.

Christmann, A. and Steinwart, I. (2010).
Universal kernels on non-standard input spaces.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 406–414.

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.