Kernel-based learning on probability distributions

Zoltán Szabó (Gatsby Unit, UCL)

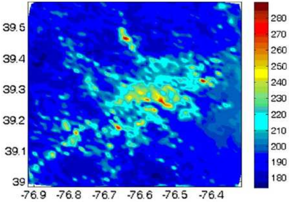
University of California, San Diego

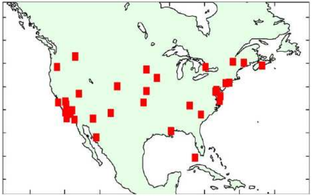
March 16, 2016

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.

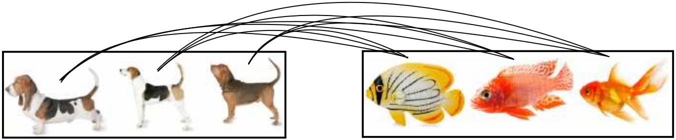




Example: existing methods

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

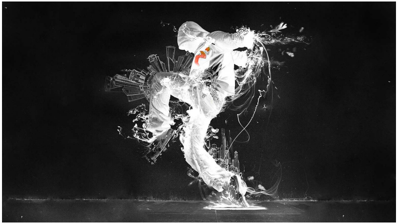
Contributions:

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?

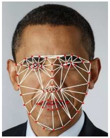
One-page summary

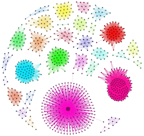
Contributions:

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%) $\rightarrow$ JMLR in revision.

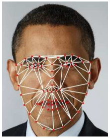


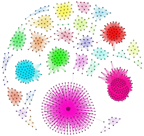
Objects in the bags





- 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, ...





- 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:

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

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:

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

- 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:

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

- 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?

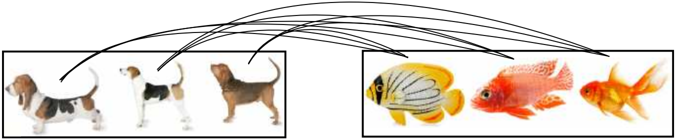
Regression on labelled bags: similarity

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

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.

Aerosol prediction result ($100 \times RMSE$)

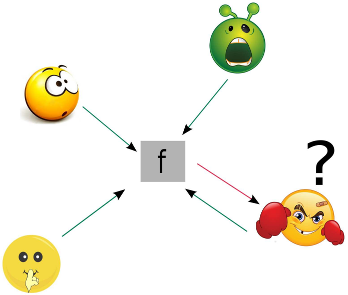
We perform on par with the state-of-the-art, hand-engineered method.

- 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/
```

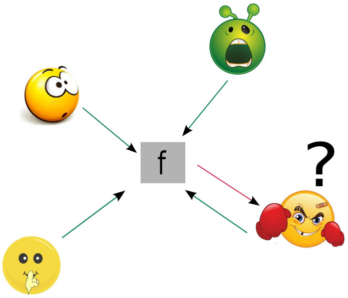
Related results

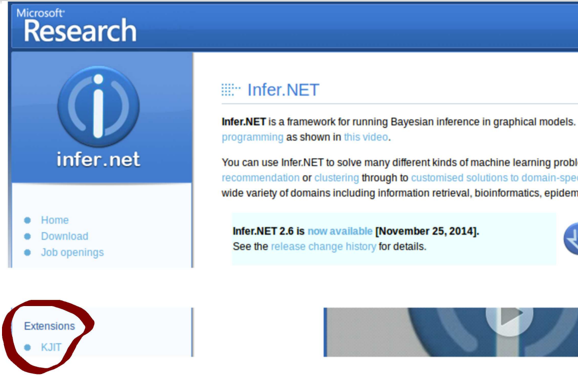
Distribution regression with random Fourier features



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

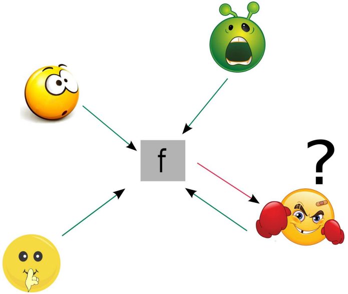
Distribution regression with random Fourier features

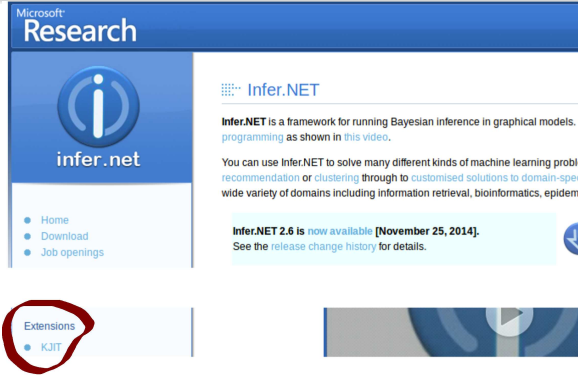




- Kernel EP [UAI-2015]:
  - distribution regression phrasing,
  - learn the message-passing operator for 'tricky' factors.
  - extends Infer.NET; speed $\Leftarrow$ RFF.

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 [→NIPS-2016]:
  - 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.

Why can we get consistency/rates? – intuition

- Convergence of the mean embedding:

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

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

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

- $f_{\hat{z}}^{\lambda}$ depends 'nicely' on $K(\mu_{\hat{P}}, \mu_{\hat{Q}}) = \left\langle K(\cdot, \mu_{\hat{P}}), K(\cdot, \mu_{\hat{Q}}) \right\rangle_{\mathcal{H}}$. [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: $Y = \mathbb{R} \Rightarrow \mathcal{L}(Y) = \mathbb{R}$; $Y = \mathbb{R}^d \Rightarrow \mathcal{L}(Y) = \mathbb{R}^{d\times d}$.

Other valid similarities

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

$\tilde{K}_G$	$\tilde{K}_e$	$\tilde{K}_C$
$e^{-\frac{\lVert \mu_P - \mu_Q \rVert^2}{2\theta^2}}$	$e^{-\frac{\lVert \mu_P - \mu_Q \rVert}{2\theta^2}}$	$\left(1 + \lVert \mu_P - \mu_Q \rVert^2 / \theta^2\right)^{-1}$

$\tilde{K}_t$	$\tilde{K}_i$
$\left(1 + \lVert \mu_P - \mu_Q \rVert^\theta\right)^{-1}$	$\left(\lVert \mu_P - \mu_Q \rVert^2 + \theta^2\right)^{-\frac{1}{2}}$

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

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.