Infinite Task Learning with Vector-Valued RKHSs

Alex Lambert
Joint work with R. Brault, Z. Szabo, M. Sangnier, F.d'Alché-Buc.

September 13, 2018

Motivation

An example of task : Quantile Regression

- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables
- $\theta \in (0, 1)$

An example of task : Quantile Regression

- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables
- $\theta \in (0, 1)$

Learn

$$q(x) = \inf \ \{y \in \mathbb{R} \ , \ P(Y \leqslant y \mid X = x) = \theta\}$$

from *iid* copies $(x_i, y_i)_{i=1}^n$

An example of task : Quantile Regression

- $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ random variables
- $\theta \in (0, 1)$

Learn

$$q(x) = \inf \ \{y \in \mathbb{R} \ , \ P(Y \leqslant y \mid X = x) = \theta\}$$

from *iid* copies $(x_i, y_i)_{i=1}^n$

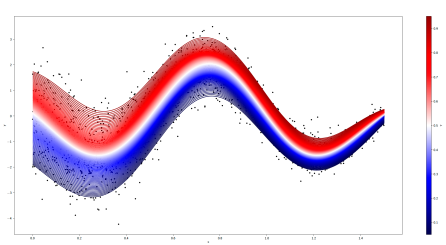

Figure 1: Example of several quantile functions (toy dataset).

An example of task : Quantile Regression

Minimize in h

$$\mathbb{E}_{X,Y}[\max(\theta(Y - h(X)), (\theta - 1)(Y - h(X)))]$$

An example of task : Quantile Regression

Minimize in h

$$\mathrm{E}_{X,Y}[\max(\theta(Y-h(X)),(\theta-1)(Y-h(X)))]$$

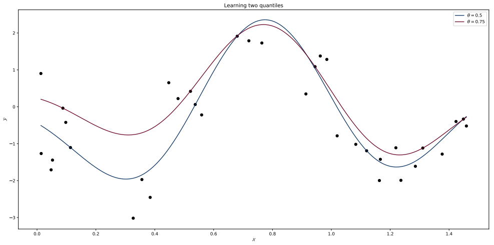

Figure 2: Two independently learnt quantile estimations.

Minimize in h

$$E_{X,Y}[\max(\theta(Y-h(X)),(\theta-1)(Y-h(X)))]$$

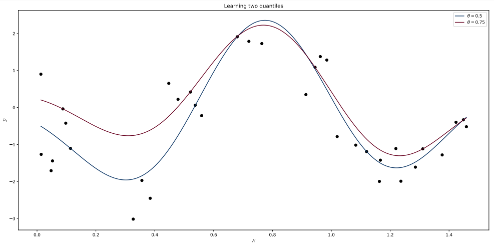

Figure 2: Two independently learnt quantile estimations.

- Not adapted to the structure of the problem

An example of task : Quantile Regression

Minimize in h

$$\mathbb{E}_{X,Y}[\max(\theta(Y-h(X)),(\theta-1)(Y-h(X)))]$$

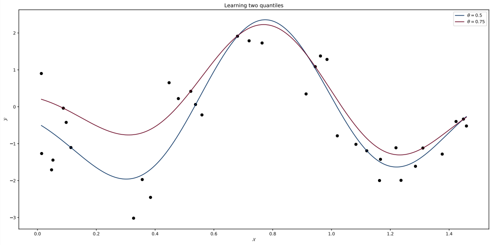

Figure 2: Two independently learnt quantile estimations.

- Not adapted to the structure of the problem
- No way to recover other quantiles

An example of task : Cost-Sensitive Classification

- Binary classification with asymetric loss function. Minimize

$$\mathsf{E}_{X,Y}\left[\left|\frac{\theta+1}{2}-\mathbb{1}_{\{-1\}}(Y)\right||1-Yh(X)|_+\right]$$

An example of task : Cost-Sensitive Classification

- Binary classification with asymetric loss function. Minimize

$$\mathbb{E}_{X,Y}\left[\left|\frac{\theta+1}{2}-\mathbb{1}_{\{-1\}}(Y)\right||1-Yh(X)|_{+}\right]$$

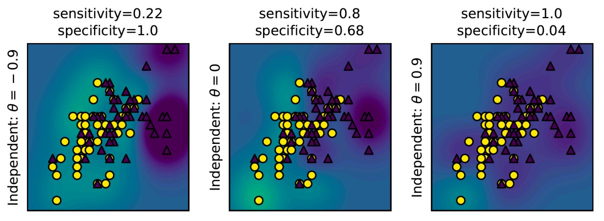

Figure 3: Independent cost-sensitive classification.

An example of task : Cost-Sensitive Classification

- Binary classification with asymetric loss function. Minimize

$$\mathrm{E}_{X,Y}\left[\left|\frac{\theta+1}{2}-\mathbb{1}_{\{-1\}}(Y)\right||1-Yh(X)|_+\right]$$

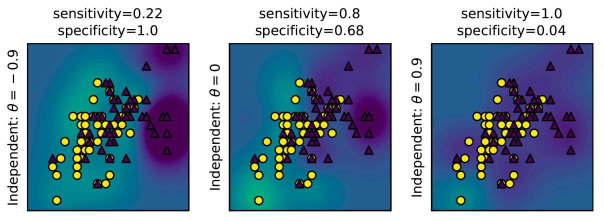

Figure 3: Independent cost-sensitive classification.

- No structure, No interpolation

An example of task : Density Level Set Estimation

(Schölkopf et al., 2000) Given $(x_i)_{i=1}^n$ *iid and* $\theta \in (0,1)$*, minimize for* $(h,t) \in \mathcal{H}_k \times \mathbb{R}$

$$J(h,t) = \frac{1}{\theta n} \sum_{i=1}^{n} \max\left(0, t - h(x_i)\right) - t + \frac{1}{2}\|h\|_{\mathcal{H}_k}^2$$

An example of task : Density Level Set Estimation

(Schölkopf et al., 2000) Given $(x_i)_{i=1}^n$ *iid and* $\theta \in (0,1)$*, minimize for* $(h,t) \in \mathcal{H}_k \times \mathbb{R}$

$$J(h,t) = \frac{1}{\theta n} \sum_{i=1}^{n} \max\left(0, t - h(x_i)\right) - t + \frac{1}{2}\|h\|^2_{\mathcal{H}_k}$$

Decision function

$$d(x) = \mathbb{1}_{\mathbb{R}_+}(h(x) - t)$$

An example of task : Density Level Set Estimation

(Schölkopf et al., 2000) Given $(x_i)_{i=1}^n$ *iid and* $\theta \in (0,1)$*, minimize for* $(h,t) \in \mathcal{H}_k \times \mathbb{R}$

$$J(h,t) = \frac{1}{\theta n} \sum_{i=1}^{n} \max\left(0, t - h(x_i)\right) - t + \frac{1}{2}\|h\|^2_{\mathcal{H}_k}$$

Decision function

$$d(x) = \mathbb{1}_{\mathbb{R}_+}(h(x) - t)$$

θ-property of the decision function

The decision function should separate new data into two separate subsets with proportion θ *of outliers.*

Multi-Task Learning

Given a problem indexed by some hyperparameter θ, solve concomitantly a finite number of tasks given by $(\theta_1, \ldots, \theta_p)$.

Multi-Task Learning

Given a problem indexed by some hyperparameter θ, solve concomitantly a finite number of tasks given by $(\theta_1, \ldots, \theta_p)$.

- Output in $\mathbb{R}^p$

Multi-Task Learning

Given a problem indexed by some hyperparameter θ, solve concomitantly a finite number of tasks given by $(\theta_1, \ldots, \theta_p)$.

- Output in $\mathbb{R}^p$
- Sum the loss functions associated to each $(\theta_i)_{i=1}^p$

Multi-Task Learning

Given a problem indexed by some hyperparameter θ, solve concomitantly a finite number of tasks given by $(\theta_1, \ldots, \theta_p)$.

- Output in $\mathbb{R}^p$
- Sum the loss functions associated to each $(\theta_i)_{i=1}^p$
- Add regularization to benefit from similarity of tasks

Multi-Task Learning

Given a problem indexed by some hyperparameter θ, solve concomitantly a finite number of tasks given by $(\theta_1, \ldots, \theta_p)$.

- Output in $\mathbb{R}^p$
- Sum the loss functions associated to each $(\theta_i)_{i=1}^p$
- Add regularization to benefit from similarity of tasks
- Create specific model constraints with prior knowledge of tasks

How to extend this to a continuum of tasks ?

A functional approach

Proposed framework : learn function-valued functions

$$\text{`input} \mapsto (\text{hyperparameter} \mapsto \text{output})\text{'}$$

$$\text{`}x \mapsto (\theta \mapsto y)\text{'}$$

A functional approach

Proposed framework : learn function-valued functions

$$\text{`input} \mapsto (\text{hyperparameter} \mapsto \text{output})\text{'}$$

$$\text{`}x \mapsto (\theta \mapsto y)\text{'}$$

Goal : Learn a global function while preserving desired properties of the output function for each hyperparameter θ.

Supervised Learning Framework

Parametrized Task

ERM setting: minimize in $h \in \mathcal{H} \subset \mathcal{F}(\mathcal{X}; \mathcal{F}(\Theta; \mathbb{R}))$ for a training set $\mathcal{S} = (x_i, y_i)_{i=1}^n$ and $\lambda > 0$

$$R_{\mathcal{S}}(h) = \sum_{i=1}^{n} V(y_i, h(x_i)) + \lambda \Omega(h)$$

where

$$V(y, h(x)) := \int_{\Theta} v(\theta, y, h(x)(\theta)) d\mu(\theta),$$

and $\Omega(h)$ is a regularization term.

Sampled Empirical Risk

Estimating the integral: Quadrature, Monte-Carlo or Quasi-Monte-Carlo.

$$\widetilde{V}(y, h(x)) := \sum_{j=1}^{m} w_j v(\theta_j, y, h(x)(\theta_j))$$

Sampled Empirical Risk

Estimating the integral: Quadrature, Monte-Carlo or Quasi-Monte-Carlo.

$$\widetilde{V}(y, h(x)) := \sum_{j=1}^{m} w_j v(\theta_j, y, h(x)(\theta_j))$$

- w_j can't depend on h

Sampled Empirical Risk

Estimating the integral: Quadrature, Monte-Carlo or Quasi-Monte-Carlo.

$$\widetilde{V}(y, h(x)) := \sum_{j=1}^{m} w_j v(\theta_j, y, h(x)(\theta_j))$$

- w_j can't depend on h
- QMC: low discrepancy sequences (Sobol) lead to error rates $\mathcal{O}(\frac{\log(m)}{m})$

Sampled Empirical Risk

Estimating the integral: Quadrature, Monte-Carlo or Quasi-Monte-Carlo.

$$\widetilde{V}(y, h(x)) := \sum_{j=1}^{m} w_j v(\theta_j, y, h(x)(\theta_j))$$

- w_j can't depend on h
- QMC: low discrepancy sequences (Sobol) lead to error rates $\mathcal{O}(\frac{\log(m)}{m})$
- No need to approximate too precisely

Functional space $\mathcal{H}$

vv-RKHS framework (Carmeli et al., 2006):

- Hilbert space of functions with values in a Hilbert space
- Regularity properties (bounded functional evaluation)

Functional space $\mathcal{H}$

vv-RKHS framework (Carmeli et al., 2006):

- Hilbert space of functions with values in a Hilbert space
- Regularity properties (bounded functional evaluation)

Take two scalar kernels $k_{\mathcal{X}}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_{\Theta}\colon \Theta \times \Theta \to \mathbb{R}$, construct

$$K\colon \begin{cases} \mathcal{X} \times \mathcal{X} & \to \mathcal{L}(\mathcal{H}_{k_\Theta}) \\ x, z & \mapsto k_{\mathcal{X}}(x, z) I_{\mathcal{H}_{k_\Theta}} \end{cases}$$

Functional space $\mathcal{H}$

vv-RKHS framework (Carmeli et al., 2006):

- Hilbert space of functions with values in a Hilbert space
- Regularity properties (bounded functional evaluation)

Take two scalar kernels $k_{\mathcal{X}}: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_{\Theta}: \Theta \times \Theta \to \mathbb{R}$, construct

$$K: \begin{cases} \mathcal{X} \times \mathcal{X} & \to \mathcal{L}(\mathcal{H}_{k_\Theta}) \\ x, z & \mapsto k_{\mathcal{X}}(x, z) I_{\mathcal{H}_{k_\Theta}} \end{cases}$$

Structure: $\mathcal{H}_K \simeq \mathcal{H}_{k_{\mathcal{X}}} \otimes \mathcal{H}_{k_\Theta}$ *i.e*

$$\mathcal{H}_K = \overline{\mathrm{span}} \; \{k_{\mathcal{X}}(\cdot, x) \cdot k_{\Theta}(\cdot, \theta), (x, \theta) \in \mathcal{X} \times \Theta\}$$

Optimization

Optimization problem:

$$\underset{h \in \mathcal{H}_K}{\arg\min}\ \widetilde{R}_{\mathcal{S}}(h) + \lambda \|h\|^2_{\mathcal{H}_K}, \quad \lambda > 0 \tag{1}$$

Optimization

Optimization problem:

$$\underset{h \in \mathcal{H}_K}{\arg\min} \; \widetilde{R}_{\mathcal{S}}(h) + \lambda \|h\|^2_{\mathcal{H}_K}, \quad \lambda > 0 \tag{1}$$

Representer Theorem

Assume that the local loss function is a proper *l.s.c* function. Then, the solution h^* to the problem (1) is unique and verifies $\forall (x, \theta) \in \mathcal{X} \times \Theta$

$$h^*(x)(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} k_{\mathcal{X}}(x, x_i) k_{\Theta}(\theta, \theta_j)$$

for some $(\alpha_{ij})_{i,j=1}^{n,m} \in \mathbb{R}^{n \times m}$.

Optimization

Optimization problem:

$$\underset{h\in\mathcal{H}_K}{\arg\min}\ \widetilde{R}_{\mathcal{S}}(h) + \lambda \|h\|^2_{\mathcal{H}_K}, \quad \lambda > 0 \tag{1}$$

Representer Theorem

Assume that the local loss function is a proper *l.s.c* function. Then, the solution h^* to the problem (1) is unique and verifies $\forall (x, \theta) \in \mathcal{X} \times \Theta$

$$h^*(x)(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} k_{\mathcal{X}}(x, x_i) k_{\Theta}(\theta, \theta_j)$$

for some $\left(\alpha_{ij}\right)_{i,j=1}^{n,m} \in \mathbb{R}^{n\times m}$.

Optimization

Solved by L-BFGS-B + smoothing of the local loss.

- Complexity in $\mathcal{O}(\sharp\text{iterations} \cdot (n^2 m + nm^2))$
- Smoothing à la Huber: infimal convolution with $\|\cdot\|^2$

Statistical Guarantees

Context of uniform stability in vv-RKHS (Kadri et al., 2015)

Generalization bound

Let $h^* \in \mathcal{H}_K$ be the solution of the problem above for the QR or CSC problem with QMC approximation. For a large class of kernels,

$$R(h^*) \leqslant \widetilde{R}_{\mathcal{S}}(h^*) + \mathcal{O}_{\mathbf{P}_{X,Y}}\left(\frac{1}{\sqrt{\lambda n}}\right) + \mathcal{O}\left(\frac{\log(m)}{\sqrt{\lambda} m}\right)$$

Statistical Guarantees

Context of uniform stability in vv-RKHS (Kadri et al., 2015)

Generalization bound

Let $h^* \in \mathcal{H}_K$ be the solution of the problem above for the QR or CSC problem with QMC approximation. For a large class of kernels,

$$R(h^*) \leqslant \widetilde{R}_{\mathcal{S}}(h^*) + \mathcal{O}_{\mathbf{P}_{X,Y}}\left(\frac{1}{\sqrt{\lambda n}}\right) + \mathcal{O}\left(\frac{\log(m)}{\sqrt{\lambda m}}\right)$$

- Requires bounded random variables in QR

Statistical Guarantees

Context of uniform stability in vv-RKHS (Kadri et al., 2015)

Generalization bound

Let $h^* \in \mathcal{H}_K$ be the solution of the problem above for the QR or CSC problem with QMC approximation. For a large class of kernels,

$$R(h^*) \leqslant \widetilde{R}_{\mathcal{S}}(h^*) + \mathcal{O}_{\mathbf{P}_{X,Y}}\left(\frac{1}{\sqrt{\lambda n}}\right) + \mathcal{O}\left(\frac{\log(m)}{\sqrt{\lambda m}}\right)$$

- Requires bounded random variables in QR
- Tradeoff between n and m

Statistical Guarantees

Context of uniform stability in vv-RKHS (Kadri et al., 2015)

Generalization bound

Let $h^* \in \mathcal{H}_K$ be the solution of the problem above for the QR or CSC problem with QMC approximation. For a large class of kernels,

$$R(h^*) \leqslant \widetilde{R}_{\mathcal{S}}(h^*) + \mathcal{O}_{\mathbf{P}_{X,Y}}\left(\frac{1}{\sqrt{\lambda n}}\right) + \mathcal{O}\left(\frac{\log(m)}{\sqrt{\lambda m}}\right)$$

- Requires bounded random variables in QR
- Tradeoff between n and m
- Mild hypothesis on the kernels

Numerical experiments: Infinite Quantile Regression

Crossing penalty: hard or soft constraints.

$$\Omega_{\mathrm{nc}}(h) := \lambda_{nc} \int_{\mathcal{X}} \int_{\Theta} \left| -\frac{\partial h}{\partial \theta}(x)(\theta) \right|_{+} d\mu(\theta) d\mathbf{P}(x)$$

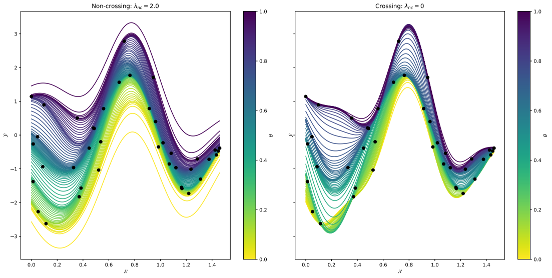

Figure 4: Comparison w/o crossing penalty for IQR.

Numerical experiments: Infinite Quantile Regression

Crossing penalty: hard or soft constraints.

$$\Omega_{\mathrm{nc}}(h) := \lambda_{nc} \int_{\mathcal{X}} \int_{\Theta} \left| -\frac{\partial h}{\partial \theta}(x)(\theta) \right|_{+} d\mu(\theta) d\mathbf{P}(x)$$

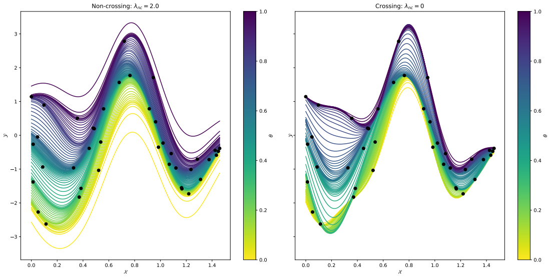

Figure 4: Comparison w/o crossing penalty for IQR.

- Matches state of the art on 20 UCI datasets. (Sangnier et al., 2016)

Numerical experiments: Infinite Cost-Sensitive Classification

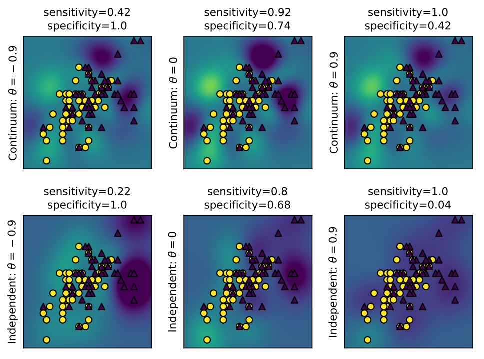

Figure 5: ICSC vs Independent learning

Numerical experiments: Infinite Cost-Sensitive Classification

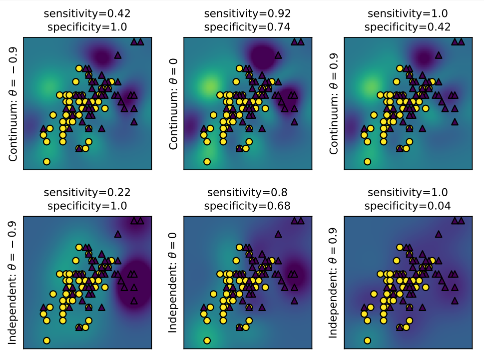

Figure 5: ICSC vs Independent learning

- Improves performances

Numerical experiments: Infinite Cost-Sensitive Classification

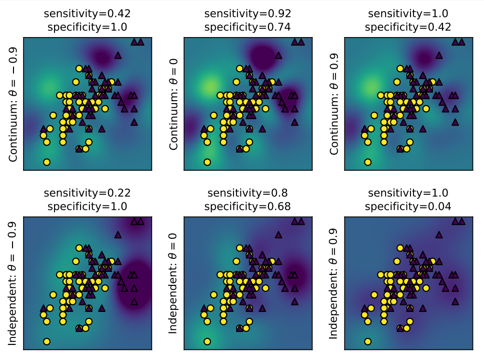

Figure 5: ICSC vs Independent learning

- Improves performances
- Hard to tune the kernels

An unsupervised task : Density level set estimation

Functional learning

Integrated problem: minimize in $h, t \in \mathcal{H}_K \times \mathcal{H}_{k_b}$

$$\int_0^1 \frac{1}{\theta n} \sum_{i=1}^{n} \max\left(0, t(\theta) - h(x_i)(\theta)\right) - t(\theta) + \frac{1}{2} \|h(\cdot)(\theta)\|^2_{\mathcal{H}_{k_{\mathcal{X}}}} \mathrm{d}\mu(\theta)$$

Functional learning

Integrated problem: minimize in $h, t \in \mathcal{H}_K \times \mathcal{H}_{k_b}$

$$\int_0^1 \frac{1}{\theta n} \sum_{i=1}^{n} \max\left(0, t(\theta) - h(x_i)(\theta)\right) - t(\theta) + \frac{1}{2} \|h(\cdot)(\theta)\|^2_{\mathcal{H}_{k_{\mathcal{X}}}} \mathrm{d}\mu(\theta)$$

Take $(\theta_j)_{j=1}^m \in (0, 1)$ a QMC sequence, minimize

$$\begin{aligned} J(h, t) = \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{\theta_j} \max\left(0, t(\theta_j) - h(x_i)(\theta_j)\right) \\ - t(\theta_j) + \left\|h(\cdot)(\theta_j)\right\|^2_{\mathcal{H}_{k_{\mathcal{X}}}} + \frac{\lambda}{2} \|t\|^2_{\mathcal{H}_{k_b}} \end{aligned}$$

Solving in Vector-Valued RKHSs

Finite expansion of the solution

There exist $\left(\alpha_{ij}\right)_{i,j=1}^{n,m} \in \mathbb{R}^{n\times m}$ and $\left(\beta_j\right)_{j=1}^{m} \in \mathbb{R}^m$ such that for $\forall (x, \nu) \in \mathcal{X} \times (0, 1)$,

$$h^*(x)(\nu) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} k_{\mathcal{X}}(x, x_i) k_{\nu}(\nu, \nu_j)$$

$$t^*(\nu) = \sum_{j=1}^{m} \beta_j k_b(\nu, \nu_j)$$

Solving in Vector-Valued RKHSs

Finite expansion of the solution

There exist $(\alpha_{ij})_{i,j=1}^{n,m} \in \mathbb{R}^{n\times m}$ and $(\beta_j)_{j=1}^{m} \in \mathbb{R}^m$ such that for $\forall (x, \nu) \in \mathcal{X} \times (0, 1)$,

$$h^*(x)(\nu) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} k_{\mathcal{X}}(x, x_i) k_{\nu}(\nu, \nu_j)$$

$$t^*(\nu) = \sum_{j=1}^{m} \beta_j k_b(\nu, \nu_j)$$

- Weak regularizer but still representer

Solving in Vector-Valued RKHSs

Finite expansion of the solution

There exist $(\alpha_{ij})_{i,j=1}^{n,m} \in \mathbb{R}^{n\times m}$ and $(\beta_j)_{j=1}^{m} \in \mathbb{R}^m$ such that for $\forall(x,\nu) \in \mathcal{X} \times (0,1)$,

$$h^*(x)(\nu) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} k_{\mathcal{X}}(x, x_i) k_{\nu}(\nu, \nu_j)$$

$$t^*(\nu) = \sum_{j=1}^{m} \beta_j k_b(\nu, \nu_j)$$

- Weak regularizer but still representer
- Classical convex problem in $\mathbb{R}^{(n+1)m}$: solvers (L-BFGS)

Solving in Vector-Valued RKHSs

Finite expansion of the solution

There exist $(\alpha_{ij})_{i,j=1}^{n,m} \in \mathbb{R}^{n \times m}$ and $(\beta_j)_{j=1}^{m} \in \mathbb{R}^m$ such that for $\forall (x, \nu) \in \mathcal{X} \times (0, 1)$,

$$h^*(x)(\nu) = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} k_{\mathcal{X}}(x, x_i) k_{\nu}(\nu, \nu_j)$$

$$t^*(\nu) = \sum_{j=1}^{m} \beta_j k_b(\nu, \nu_j)$$

- Weak regularizer but still representer
- Classical convex problem in $\mathbb{R}^{(n+1)m}$: solvers (L-BFGS)

Numerical experiments: Infinite One-Class SVM

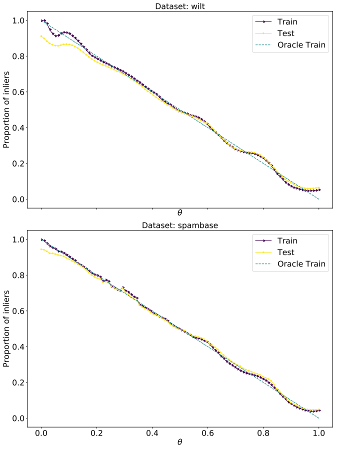

Figure 6: Level set estimation: the ν-property is approximately satisfied. Top: Wilt benchmark; bottom: Spambase dataset.

Perspectives

Perspectives

Investigate:

- Algorithmic guarantees

Perspectives

Investigate:

- Algorithmic guarantees
- New regularization term : $\sum_j \left\| h(\cdot)(\theta_j) \right\|_{\mathcal{H}_{k_{\mathcal{X}}}}$

Perspectives

Investigate:

- Algorithmic guarantees
- New regularization term : $\sum_j \| h(\cdot)(\theta_j) \|_{\mathcal{H}_{k_{\mathcal{X}}}}$
- Hard monotony constraints

Perspectives

Investigate:

- Algorithmic guarantees
- New regularization term : $\sum_j \left\| h(\cdot)(\theta_j) \right\|_{\mathcal{H}_{k_{\mathcal{X}}}}$
- Hard monotony constraints
- Scaling up : ORFF (Brault et al., 2016)

References

Schölkopf, Bernhard et al. (2000). "New support vector algorithms." In: *Neural computation* 12.5, pp. 1207–1245.

Carmeli, Claudio, Ernesto De Vito, and Alessandro Toigo (2006). "Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem." In: *Analysis and Applications* 4 (4), pp. 377–408.

Kadri, Hachem et al. (2015). "Operator-valued kernels for learning from functional response data." In: *Journal of Machine Learning Research* 16, pp. 1–54.

Brault, Romain, Markus Heinonen, and Florence d'Alché-Buc (2016). "Random Fourier Features For Operator-Valued Kernels." In: *Asian Conference on Machine Learning (ACML)*, pp. 110–125.

Sangnier, Maxime, Olivier Fercoq, and Florence d'Alché-Buc (2016). "Joint quantile regression in vector-valued RKHSs." In:

Advances in Neural Information Processing Systems (NIPS),
pp. 3693–3701.