Spectral methods for learning a set

L. Rosasco[1,4] A. Toigo[2] E. De Vito[3,4]

[1]CBCL, M.I.T., Boston, USA [2]Dip. Matematica, Politecnico di Milano, Italy

[3]DIMA, Università di Genova, Italy [4]Slipguru, DISI, Genova, Italy

Workshop in honour of Mario Bertero

DISI, Genova 2th February 2011

Plan of the talk

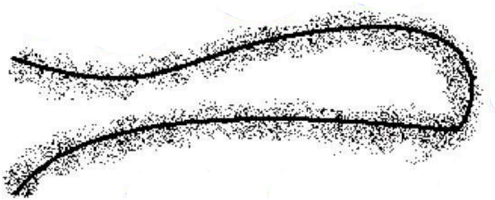
1. The problem: learning a set from random data
2. The ingredients: a **completely regular** reproducing kernel Hilbert space and a low-pass filter
3. The results: a kernel estimator and its consistency
4. (Preliminary) experiments

The problem

- we have a sample of n-examples $x_1, \ldots, x_n$
- we fix a (possibly high dimensional) representation

$$x_i = (x_i^1, \ldots, x_i^d) \in \mathbb{R}^d \qquad \text{with } d \gg n$$

- we believe that the points similar to the examples do not live in a fat region of $\mathbb{R}^d$, but they belong to a thin subset
- we aim to learn some properties of this thin subset from the examples



The mathematical setting

- we assume that the examples are sampled independently according to an **unknown probability measure** ρ defined on a compact subset X of $\mathbb{R}^d$
- we model the thin subset as the **smallest closed** subset C_ρ such that $\rho(C_\rho) = 1$, *i.e* C_ρ is the support of the measure ρ
- the goal is to define a set C_n, depending on the examples, such that C_n is *close* to C_ρ with respect some distance among sets for example the Hausdorff distance

$$\mathrm{d}_H(C_n, C_\rho) = \max\{\sup_{x \in C_n} d(x, C_\rho), \sup_{x \in C_\rho} d(x, C_n)\}$$

Note that $\mathrm{d}_H(C_n, C_\rho)$ is a random variable

State of the art

Many different frameworks

1. support density estimation
2. level set density estimation
3. novelty/anomaly detection
4. one-class classifier
5. spectral manifold learning
6. dimensionality reduction

Our approach is based on the idea of "spectral regularization" and

i) ρ is not assumed to have a density with respect to the Lebesgue measure
ii) C_ρ is not assumed to be a Riemannian submanifold
iii) Our algorithm is easy to implement (at the cost of an SVD)

Our results

Three steps

1. we define a continuous function $F : X \to [0,1]$ such that

$$C_\rho = \{x \in X \mid F(x) = 1\}$$

   which explicitly depends on ρ

2. we define a continuous estimator $F_n : X \to [0,1]$ of F such that
   a) F_n only depends on the examples through a matrix $\mathbf{K}_n$
   b) F_n converges uniformly to F

3. The plug-in estimator is given by

$$C_n = \{x \in X \mid F_n(x) \geq 1 - \tau_n\}$$

   where τ_n is a tuning parameter.

Ingredients

We need

- A completely regular reproducing kernel Hilbert space
  - Example: the Abel kernel

$$K(x,\tilde{x}) = e^{-\gamma\|x-\tilde{x}\|} \propto \text{Fourier transform of the Poisson kernel}$$

    where $\gamma > 0$ is a fixed parameter

- A low-pass filter r_λ in the frequency domain, where λ is a regularization parameter
  - Example: the Tikhonov filter

$$r_\lambda(\sigma) = \frac{\sigma}{\sigma + \lambda}$$

Reproducing Kernel Hilbert space (RKHS)

A Hilbert space $\mathcal{H}$ is a RKHS if

- the elements of $\mathcal{H}$ are functions $f : X \to \mathbb{R}$ with the pointwise operations
- for any $x \in X$ there is a unique $K_x \in \mathcal{H}$ such that

$$\text{reproducing formula} \qquad f(x) = \langle f, K_x \rangle \qquad f \in \mathcal{H}$$

- the reproducing kernel $K : X \times X \to \mathbb{R}$

$$K(x, \tilde{x}) = K_x(\tilde{x}) = \langle K_{\tilde{x}}, K_x \rangle$$

  is continuous (so that the elements of $\mathcal{H}$ are continuous functions)
- $K_x \neq K_{\tilde{x}}$ for all $x \neq \tilde{x}$ and $K(x, x) = 1$ for all x

The feature map Φ

$$X \ni x \mapsto K_x \in \mathcal{H}$$

is a continuous embedding of X into the linear space $\mathcal{H}$ ($\dim \mathcal{H} \gg d$)

Mercer theorem (revisited)

- The integral operator on $L^2(X, \rho)$

$$(Lf)(x) = \int_X K(x, \tilde{x})\, f(\tilde{x})\; d\rho(\tilde{x})$$

  is a positive Hilbert-Schmidt operator with range into $\mathcal{H}$
- There is a base $(\varphi_k)_{k\in\mathbb{N}}$ of eigenfunctions of L with the corresponding sequence of eigenvalues $(\sigma_k)_{k\in\mathbb{N}}$: $L\varphi_k = \sigma_k\varphi_k$

Mercer theorem

$$\sum_k \sigma_k |\varphi_k(x)|^2 = K(x, x) = 1 \qquad x \in C_\rho$$

$$\sum_k \sigma_k |\varphi_k(x)|^2 \neq K(x, x) \qquad x \notin C_\rho \qquad ?$$

YES, provided that $\mathcal{H}$ separes C_ρ:

for any $x \notin C_\rho$ there exists $f \in \mathcal{H}$

$$f(x) \neq 0 \qquad \text{and} \qquad f(\tilde{x}) = 0 \quad \forall \tilde{x} \in C_\rho$$

Separating property and universal kernels

- C_ρ is separated by $\mathcal{H}$ if there exists a closed subspace $\mathcal{K}$ such that

$$\Phi(C_\rho) = \mathcal{K} \cap \Phi(X)$$

- a completely regular RKHS is able to separate any closed subset

Examples

- Sobolev spaces with smoothness $s > \frac{d}{2}$

$$\mathcal{H}^s = \{f \in L^2 \mid \int_{\mathbb{R}^d} |\hat{f}(p)|^2 |p|^{2s} \, dp < +\infty\}$$

  are completely regular
- The Abel kernel $K(x, \tilde{x}) = e^{-\gamma \|x - \tilde{x}\|}$ $(\mathcal{H} \simeq \mathcal{H}^{\frac{d+1}{2}})$ is completely regular
- the linear kernel is able to separate only linear subspaces!

The function F

Let $\mathcal{H}$ be a completely regular RKHS $\mathcal{H}$ with normalized kernel K.

The continuous function

$$F : X \to \mathbb{R} \qquad F(x) = \sum_k \sigma_k |\varphi_k(x)|^2$$

is such that

$$C_\rho = \{x \in X \mid F(x) = 1\}$$

A little bit of algebra

$$F(x) = \sum_{\sigma_k > 0} |\sqrt{\sigma_k}\varphi_k(x)|^2 = \sum_{\sigma_k > 0} |\langle \sqrt{\sigma_k}\varphi_k, K_x \rangle|^2 = \left\langle T^\dagger T K_x, K_x \right\rangle$$

where $T = L_{|\mathcal{H}} \in \mathcal{L}(\mathcal{H})$ and $T^\dagger$ is the generalized inverse
Note that $T^\dagger T$ is the spectral projection associated with the strictly positive eigenvalues of T

A good empirical estimator of T

- define the finite rank positive operator on $T_n : \mathcal{H} \to \mathcal{H}$

$$(T_n\ f)(x) = \frac{1}{n} \sum_{i=1}^{n} K(x, x_i) f(x_i),$$

  depending on the examples $x_1, \ldots, x_n$

- Hoeffeding inequality for Hilbert space valued random variables gives

$$\lim_{n\to\infty} \frac{\sqrt{n}}{\log n} \|T_n - T\|_{\mathrm{HS}} = 0 \qquad \text{with probability 1}$$

- Naive idea: $F_n(x) = \left\langle T_n^\dagger T_n\, K_x, K_x \right\rangle$

- Since T is compact, then 0 is an accumulation point for the spectrum and

$$\left\langle T_n^\dagger T_n\, K_x, K_x \right\rangle \text{ does not converge to } \langle T^\dagger T K_x, K_x \rangle$$

  The instability is due to the fact that $T^\dagger$ is unbounded

A filter function: (Groetsch, C.W. Boll.Un.Mat.Ital. B **17** (1980) 1411–1419)

Take a filter function $r_\lambda : [0,1] \to [0,1]$ depending on a regularization parameter $\lambda > 0$ satisfying

1. $r_\lambda(0) = 0$ so that $r_\lambda(\sigma) = \sigma g_\lambda(\sigma)$
2. $\lim_{\lambda \to 0} r_\lambda(\sigma) = 1$ for all $\sigma > 0$
3. $|r_\lambda(\sigma) - r_\lambda(\tilde{\sigma})| \leq C_\lambda |\sigma - \tilde{\sigma}|$ for all $\lambda > 0$

then

i) $\lim_{\lambda \to 0} \sup_{x \in X} |\langle r_\lambda(T) K_x, K_x \rangle - \left\langle T^\dagger T K_x, K_x \right\rangle| = 0$

ii) $\|r_\lambda(T) - r_\lambda(T_n)\|_{\mathrm{HS}} \leq C_\lambda \|T - T_n\|_{\mathrm{HS}}$ (simple proof due to A. Maurer)

where $\|T - T_n\|_{HS}$ is the Hilbert-Schmidt (Frobenius) norm.

Item ii) is also consequence of the theory of double operator integrals due to Birman and Solomyak

Examples

1. Tikhonov

$$r_\lambda(\sigma) = \frac{\sigma}{\sigma + \lambda} \qquad C_\lambda = \frac{1}{\lambda}$$

2. Spectral Cut-Off

$$r_\lambda(\sigma) = \begin{cases} 1 = \frac{\sigma}{\sigma} & \sigma \geq \lambda \\ \frac{\sigma}{\lambda} & \sigma \leq \lambda \end{cases} \qquad C_\lambda = \frac{1}{\lambda}$$

3. Landweber

$$r_m(\sigma) = \sigma \sum_{k=0}^{m} (1 - \sigma)^m \qquad C_m = m + 1$$

4. Truncated SVD (kernel PCA)

$$r_\lambda(\sigma) = \begin{cases} 1 & \sigma \geq \lambda \\ 0 & \sigma < \lambda \end{cases}$$

it is not a Lipschitz function

A regularized empirical estimator as kernel method

Define

$$F_{n,\lambda}(x) = \langle r_\lambda(T_n) K_x, K_x \rangle = \underbrace{\left\langle (T_n + \lambda I)^{-1} T_n K_x, K_x \right\rangle}_{\text{Tikhonov}}$$

- $\mathbf{k}_x$ is the n-dimensional column vector

$$\mathbf{k}_x^t = (K(x, x_1), \ldots, K(x, x_n))$$

- $\mathbf{K}_n$ the $n \times n$-matrix $(\mathbf{K}_n)_{ij} = K(x_i, x_j)$ $\qquad \mathbf{K}_n \hat{v}_k = \hat{\sigma}_k \hat{v}_k$

$$\begin{aligned} F_n^\lambda(x) &= \frac{1}{n} \mathbf{k}_x^t \, g_\lambda\left(\frac{\mathbf{K}_n}{n}\right) \mathbf{k}_x \\ &= \frac{1}{n} \sum_{k=1}^n g_\lambda(\hat{\sigma}_k) |\mathbf{k}_x^t \hat{v}_k|^2 = \underbrace{\sum_{i=1}^n y_i(x) e^{-\gamma \|x - x_i\|}}_{\text{Abel kernel}} \end{aligned}$$

$$= \underbrace{\mathbf{k}_x^t (\mathbf{K}_n + n\lambda I)^{-1} \mathbf{k}_x}_{\text{Tikhonov}}$$

A kernel method point of view

1. given n-examples $x_1, \ldots, x_n \in C_\rho$ and a new point $x \in X$
2. label the examples according to the similarity function K

$$y_i = K(x_i, x) = e^{-\gamma \|x - x_i\|} \qquad \begin{cases} y_i \sim 1 & x_i \sim x \\ y_i \sim 0 & x_i \not\sim x \end{cases}$$

3. consider the linear inverse problem

$$\text{find} \quad f \in \mathcal{H} \quad \text{such that} \quad f(x_i) = y_i \quad \Longleftrightarrow \quad \underbrace{S_n}_{\text{sampling operator}} f = y$$

4. find the regularized solution according to the filter function g_λ

$$f_n^\lambda = g_\lambda(S_n^* S_n) S^* y \implies f_n^\lambda(x) = F_n^\lambda(x)$$

5. x is estimated to be in C_ρ if and only if $y = f_n^\lambda(x) \geq 1 - \tau_n$

Consistency

If we choose the regularization parameter λ_n so that

- $\lim_{n\to\infty} \lambda_n = 0$
- $\limsup_{n\to\infty} C_{\lambda_n} \frac{\log n}{\sqrt{n}} < +\infty$ Tikhonov filter: $\lambda_n = \frac{\log n}{\sqrt{n}}$

$$\lim_{n\to\infty} \sup_{x\in X} |F_n^{\lambda_n}(x) - F(x)| = 0 \qquad \text{with probability 1}$$

Define $C_n = \{x \in X \mid F_n^{\lambda_n}(x) \geq 1 - \tau_n\}$

- $\lim_{n\to\infty} \tau_n = 0$ $\qquad \limsup_{n\to\infty} \frac{\|F_n - F\|_\infty}{\tau_n} \leq 1$

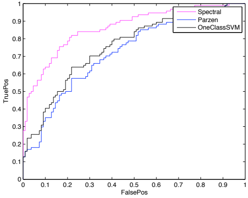
$$\lim_{n\to\infty} \underbrace{\mathrm{d}_H(C_n, C_\rho)}_{\text{Hausdorff distance}} = 0 \qquad \text{with probability 1}$$

With the Abel kernel the above results also hold for non-compact X

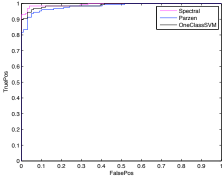
Some numerical experiments

- The final algorithm has 3 tuning parameters
  - kernel width ($K(x, \tilde{x}) = e^{-\gamma \|x - \tilde{x}\|}$) $\rightarrow$ the median 10-NN distance
  - regularization parameter ($r_\lambda(\sigma) = \frac{\sigma}{\sigma + \lambda}$) $\rightarrow$ eigenvalues decay of $\mathbf{K}_n$
  - threshold parameter ($C_\tau = \{x \in X \mid F_n(x) \geq 1 - \tau\}$)$\rightarrow$ ROC curve
- The database is MNIST (hand-written digits)
  - training set with 500 images of the same digit
  - test set of 200 images of two different digits
  - Each experiment consists of training on one class and testing on two different classes and was repeated for 20 trials over different training set choices.

MNIST 9 vs 4



MNIST 1 vs 7



EIGENVALUES DECAY

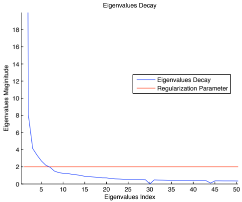


Figure: ROC curves for the estimator in two different tasks. Left: digit 9 vs 4, Center: digit 1 vs 7, Right : Eigenvalues decay

	3 vs 8	8 vs 3	1 vs 7	9 vs 4
Spectral	0.8371 ± 0.0056	0.7830 ± 0.0026	$0.9921 \pm 4.7283e-04$	0.8651 ± 0.0024
Parzen	0.7841 ± 0.0069	0.7656 ± 0.0029	$0.9811 \pm 3.4158e-04$	$0.0.7244 \pm 0.0030$
1CSVM	0.7896 ± 0.0061	0.7642 ± 0.0032	$0.9889 \pm 1.8479e-04$	0.7535 ± 0.0041

Table: Average and standard deviation of the AUC for the different estimators on the considered tasks.

Thank you and we are ready for the cake