Targeted PDF Learning

John Shawe-Taylor
Centre for Computational Statistics
and Machine Learning
University College London
jst@cs.ucl.ac.uk

Joint work with Alex Dolia and Tijl De Bie

July, 2006

Open House '06

Motivation

1. Challenge of PDF Learning: impossible in the L_1 sense – Batu et al.

2. Success with learning for one task

3. Compromise: aim to learn for set of tasks that might arise

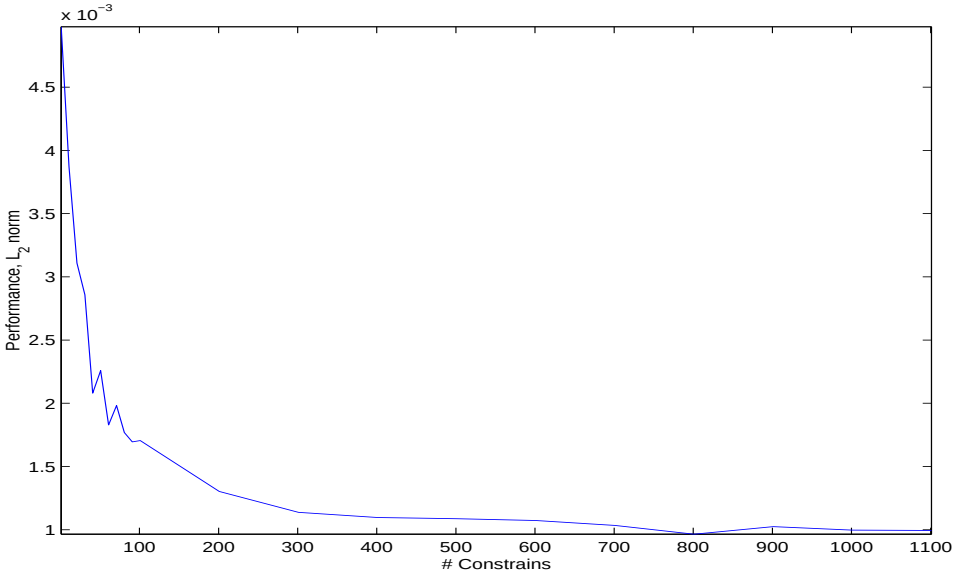
4. Needs new framework but first motivational applications

One class vs PDF learning

- Mukherjee and Vapnik added constraints to the one class SVM to fit the cumulative distribution up to data points to the estimated distribution.

- What happens if we only add some of the constraints but see how well we do on all of them? Maybe we don't need to include all the constraints?

One class vs PDF learning



- Note how only a small number of constraints is sufficient to significantly reduce the loss
- Curve then levels off as more constraints are added (note shifted axes)

Touchstone Class

A *Touchstone class* for learning a probability density function (pdf) on a measurable space $\mathcal{X}$ is

- a class of measurable real-valued functions $\mathcal{F}$ on $\mathcal{X}$ with a distribution $P_{\mathcal{F}}$ defined over $\mathcal{F}$.

- Given an unknown pdf function p, the *error* $\mathsf{err}(\hat{p})$ of an approximate pdf function $\hat{p}$ is defined as

$$\mathsf{err}(\hat{p}) = \mathbb{E}_{f \sim P_{\mathcal{F}}} \left[\ell(\mathbb{E}_p[f], \mathbb{E}_{\hat{p}}[f]) \right],$$

- where ℓ is a loss function such as the absolute value, its square or an ϵ-insensitive version of either – could also be an ϵ-insensitive classification

Examples

1. Mukherjee and Vapnik: $\mathcal{F}$ are indicator functions of downward closed sets - could also be Kolmogorov and Smirnov in 1 dimension

2. Generalise to indicator functions of a class of sets: $\mathcal{A}$-distance of He, Ben-David and Tong

3. Marginals of sets of variables. Typically two processes: estimating probabilities of the model and performing inference. Approach can be used to combine the two – see next slide

Example 3

Consider a distribution over $\{0,1\}^n$. The touchstone class $\mathcal{F}_\mathcal{I}$ is taken as a set of 'projection' functions $\pi_{\mathbf{i},\mathbf{v}}$ onto subsets $\mathbf{i} = \{i_1, \ldots, i_{|\mathbf{i}|}\} \in \mathcal{I}$ of variables drawn from a set $\mathcal{I} \subseteq 2^{\{1,\ldots,n\}}$ with prescribed values $\mathbf{v} \in \{0,1\}^{|\mathbf{i}|}$

$$\mathcal{F}_\mathcal{I} = \left\{ \pi_{\mathbf{i},\mathbf{v}} \colon \mathbf{i} \in \mathcal{I}, \mathbf{v} \in \{0,1\}^{|\mathbf{i}|} \right\}, \text{ where}$$

$$\pi_{\mathbf{i},\mathbf{v}}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x}_{i_j} = \mathbf{v}_j, \text{ for } j = 1, \ldots, |\mathbf{i}|, \\ 0; & \text{otherwise.} \end{cases}$$

For this case the expectation $\mathbb{E}_p[\pi_{\mathbf{i},\mathbf{v}}]$ is the marginal for the variables indexed by $\mathbf{i}$ set to the values $\mathbf{v}$.

Distribution of $\mathcal{P}_{\mathcal{F}}$

- In example 1 derived from input distribution but in general would be unrelated.
- It should encode our prior belief about which functions are most likely to arise in practice.
- If we simply wish to be good at all the functions we should use a uniform distribution
- Using an epsilon insensitive classification loss makes it possible to interpret the error as a probability that a randomly drawn function will be estimated with accuracy less than ϵ

Theory of learning

- $\hat{p} \in \mathcal{P}$ is an ϵ-approximation of the true density p with respect to the Touchstone Class $\mathcal{F}$, if $\mathsf{err}(\hat{p}) \leq \epsilon$
- $\mathcal{P}$ is learnable if there is an algorithm $\mathcal{A}$ such that given any $p \in \mathcal{P}$, $\epsilon > 0$ and $\delta > 0$, $\mathcal{A}$ given a sample of m i.i.d. points where m is polynomial in $\frac{1}{\epsilon}$ and $\frac{1}{\delta}$, returns an estimate $\hat{p} \in \mathcal{P}$ that with probability $1 - \delta$ is an ϵ-approximation of p
- For a class $\mathcal{P}$ of distributions and a Touchstone Class $\mathcal{F}$ of functions we define the $\mathcal{F}$-derived class of functions to be

$$\mathcal{P}_{\mathcal{F}} = \{f \in \mathcal{F} \mapsto \mathbb{E}_p[f] : p \in \mathcal{P}\}.$$

First result

Theorem 1. *Let $\mathcal{F}$ and $\mathcal{P}$ be such that there exists a polynomial Q with the property that for $m \geq Q(1/\epsilon)$,*

$$R_m(\mathcal{P}_\mathcal{F}) \leq \epsilon,$$

where the associated symmetric loss function ℓ has range $[0,1]$, satisfies the triangle inequality and is Lipschitz continuous with constant L. Then an algorithm that can select a function from $\mathcal{P}_\mathcal{F}$ that minimises the empirical ℓ loss can learn $\mathcal{P}$ with respect to the function class $\mathcal{F}$.

Support vector density estimation

A kernel κ normalised: $\int_{\mathcal{X}} \kappa(\mathbf{x}, \mathbf{z}) d\mathbf{x} = 1$.

The standard choice for κ is a normalised Gaussian

$$\kappa(\mathbf{x}, \mathbf{z}) = \frac{1}{\left(\sqrt{2\pi\sigma}\right)^d} \exp\left(-\frac{\|\mathbf{x} - \mathbf{z}\|^2}{2\sigma^2}\right)$$

If we now consider learning a density function in a dual representation $q(\mathbf{x}) = \sum_{i=1}^m \alpha_i \kappa(\mathbf{x}_i, \mathbf{x})$, the constraint $\sum_{i=1}^m \alpha_i = 1$ ensures that the density is correctly normalised,

The corresponding space $\mathcal{P}_{\mathcal{F}}(B)$ is given by

$$\mathcal{P}_{\mathcal{F}}(B) = \left\{ q_{\mathbf{w}} : f \mapsto \mathbb{E}_{q_{\mathbf{w}}}[f] \,\middle|\, \|\mathbf{w}\| \leq B, q_{\mathbf{w}}(\mathcal{X}) = 1 \right\}.$$

Optimisation problem

$$\begin{array}{ll} \min_{\alpha,\xi} & \sum_{i,j=1}^{m_x} \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j) + D \sum_{j=1}^{m_f} \xi_j \\ \text{subj to} & \sum_{i=1}^{m_x} \alpha_i = 1 \\ & \ell\left(\sum_{i=1}^{m_x} \alpha_i \int_{\mathcal{X}} \kappa(\mathbf{x}_i, \mathbf{x}) f_j(\mathbf{x}) d\mathbf{x}, \frac{1}{m_x} \sum_{i=1}^{m_x} f_j(x_i)\right) \leq \xi_j \\ & \quad \text{and } \xi_j \geq 0 \text{ for } j = 1, \ldots, m_f, \\ & \alpha_i \geq 0 \text{ for } i = 1, \ldots, m_x. \end{array}$$

Bounding SVDE

Theorem 2. *The empirical Rademacher complexity of* $\mathcal{P}_{\mathcal{F}}(B)$ *on the sample* $\{f_1, \ldots, f_{m_f}\}$ *is bounded by*

$$\hat{R}_{m_f}(\mathcal{P}_{\mathcal{F}}(B)) \leq \frac{2B}{m_f} \sqrt{\sum_{i=1}^{m_f} \min\left(C_\kappa^2 \|f_i\|_{L_1}^2, \|f_i\|_{L_1} \|f_i\|_{L_\infty}\right)}.$$

where

$$C_\kappa := \sup_{\mathbf{z},\mathbf{z}'} \sqrt{\kappa(\mathbf{z},\mathbf{z}')} = \sqrt{\kappa(\mathbf{x},\mathbf{x})} \text{ for all } \mathbf{x}.$$

Bounding SVDE

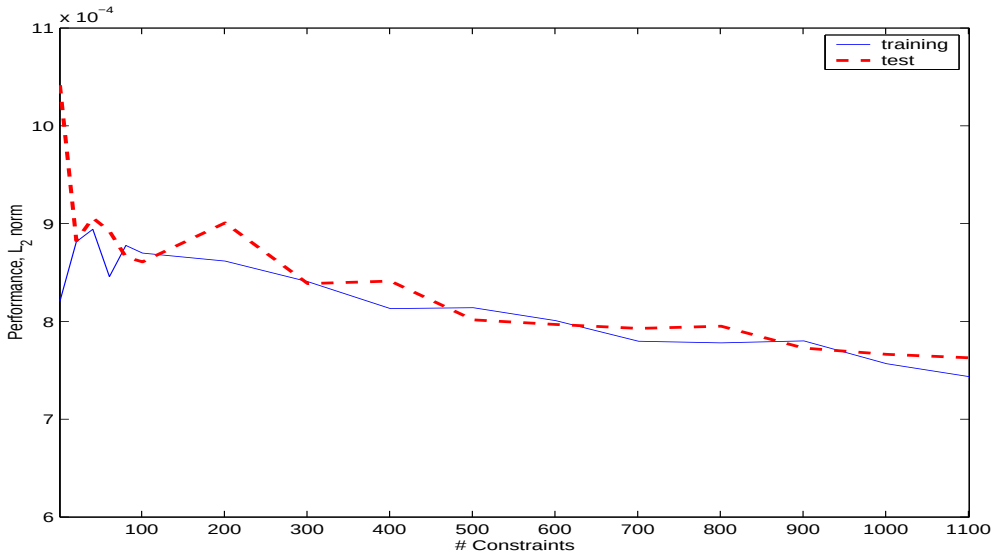
Theorem 3. *Suppose that we learn a pdf function based on a sample of m_x inputs and m_f sample functions from the space $\mathcal{F}$. Then with probability at least $1-\delta$ over the generation of the two samples we can bound the error of $\hat{p} \in \mathcal{P}_{\mathcal{F}}(B)$ by*

$$\begin{aligned} \mathit{err}(\hat{p}) \quad \leq \quad & L\sqrt{\frac{2}{m_x}\ln\frac{4m_f}{\delta}} + \hat{\mathbb{E}}_f[\ell(\mathbb{E}_{\hat{p}}[f], \hat{\mathbb{E}}_x[f])] + \\ & \frac{2BC_\kappa}{m_f}\sqrt{\sum_{i=1}^{m_f}\|f_i\|_{L_1}^2} + \sqrt{\frac{9}{2m_f}\ln\frac{4}{\delta}} \end{aligned}$$

where L is the Lipschitz constant of the loss function.

Experiments with Half spaces

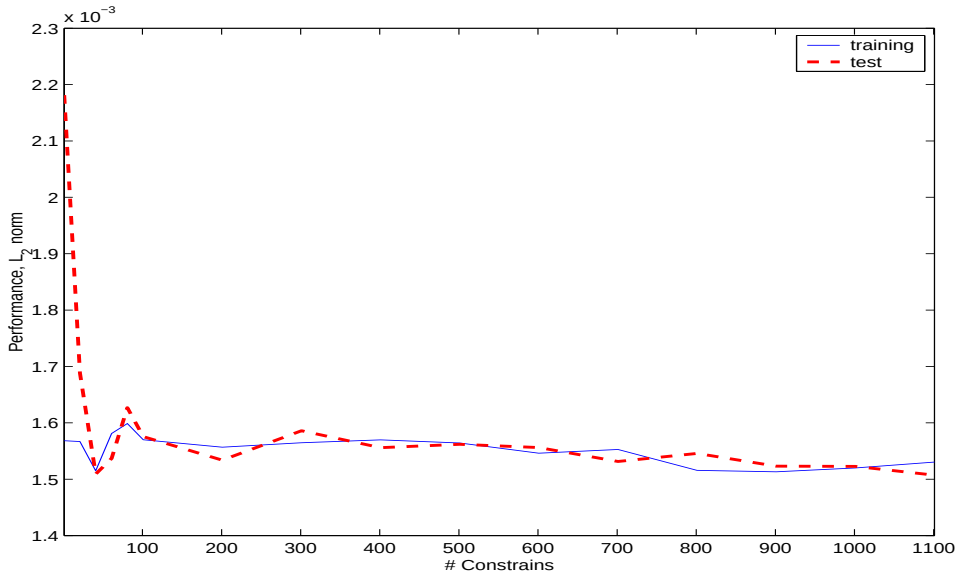
10 dimensional, 100 inputs generated by a mixture of Gaussians. Half spaces sampled using a Gaussian distribution.



The average training (blue unbroken) and test (red dashed) L_2 error as a function of the number of constraints (size of the sample m_f)

Experiments with Half spaces

10 dimensional, 500 inputs generated by a mixture of Gaussians. Half spaces sampled using a Gaussian distribution.



The average training (blue unbroken) and test (red dashed) L_2 error as a function of the number of constraints (size of the sample m_f)

Semi-supervised application

Consider using learning to get localised estimates of the density using the Touchstone class:

$$\mathcal{F} = \{f_{\mathbf{x}}(\cdot) = \kappa(\mathbf{x}, \cdot) : \mathbf{x} \sim p_L\}$$

We can use the unlabelled data to train a pdf targeted for $\mathcal{F}$.

Now use the information to guide the margin measurement when learning to classify the labelled data:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{2}\|\mathbf{w}\|^2, \\ \text{s.t.} \quad & y_i^l \langle \mathbf{x}_i^l, \mathbf{w} \rangle \geq \mathbb{E}_p[f_{\mathbf{x}_i^l}], \end{aligned}$$

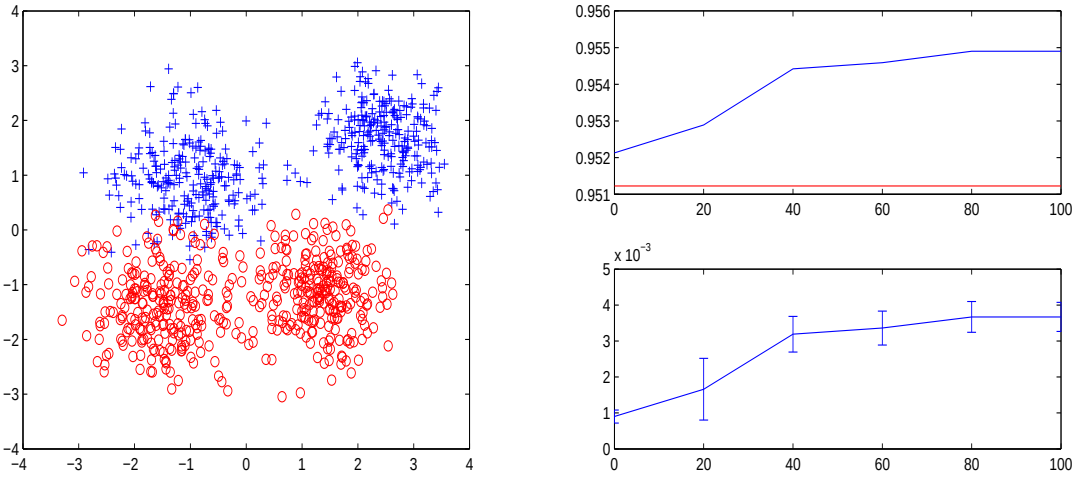
Semi-supervised application

The dual of which is given by (expressed in terms of the kernel functions $\kappa(\mathbf{x}_i, \mathbf{x}_j)$):

$$\begin{aligned} \min_\beta \quad & \frac{1}{2} \sum_{i,j=1}^{n_l} \beta_i \beta_j \kappa(\mathbf{x}_i, \mathbf{x}_j) - \sum_i \mathbb{E}_p[f_{\mathbf{x}_i^l}] \beta_i, \\ \text{s.t.} \quad & \beta_i \geq 0. \end{aligned}$$

Here we denote the dual variables with β_i to distinguish those with the variables α_i from the density estimate.

Experimental results



Left: A typical Ripley data set.

Right: The performance of the semi-supervised learning method as a function of the number of constraints used to learn the distribution of the test data. Below difference with sd's.

Conclusions

- Introduced a framework for learning a pdf targeted for a set of tasks
- Theoretical justification that approach will work under reasonable conditions
- Experiments demonstrating that fast learning can kick in quite quickly
- Application to semi-supervised learning

Future work

- Using the approach for probabilistic inference
- Theoretical analysis for ϵ-insensitive classification loss
- Applications to sensor networks – retain a range of information that might be required later
- Other applications?