A Framework for Probability Density Estimation

John Shawe-Taylor[1] and Alex Dolia[2]
Shai Ben-David[3]

[1] Centre for Computational Statistics and Machine Learning
University College London

[2] Southampton Statistical Sciences Research Institute
University of Southampton

[2]Department of Computer Science,
University of Waterloo

December, 2007

NIPS Workshops '07

Motivation

1. Learning a PDF equips us to solve all tasks that might arise: BUT PDF learning impossible in the L_1 sense:
   - Batu et al. showed that for example you need a number of samples that is $\Omega(1.587^d)$ for learning a distribution over d-dimensional boolean vectors in order to distinguish distributions in the L_1 sense.

2. Success with learning for one task such as a classification problem

3. Compromise between 1. and 2.: aim to learn PDF that is good for the subset of tasks that might arise in an application

One class vs PDF learning

- If we use a positive kernel κ such that

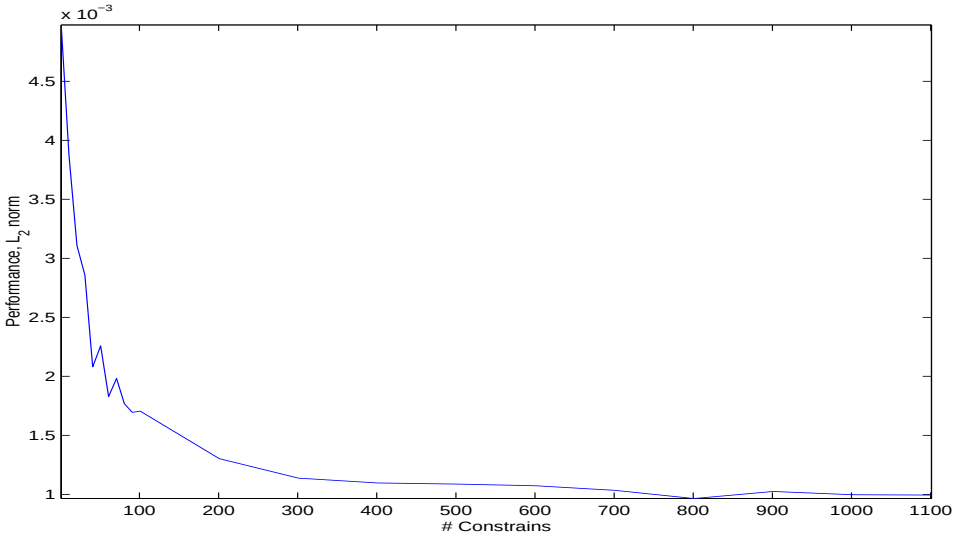
$$\|\kappa(\mathbf{x}, \cdot)\|_{L_1} = 1$$

  (such as a normalised Gaussian) then a one class SVM produces as output a pdf since it includes the constraint $\sum_{i=1}^{m} \alpha_i = 1$

- 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



- 0 constraints is 1-class SVM while 1100 (all) constraints corresponds to Mukherjee & Vapnik
- 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 $\mathcal{F}$.

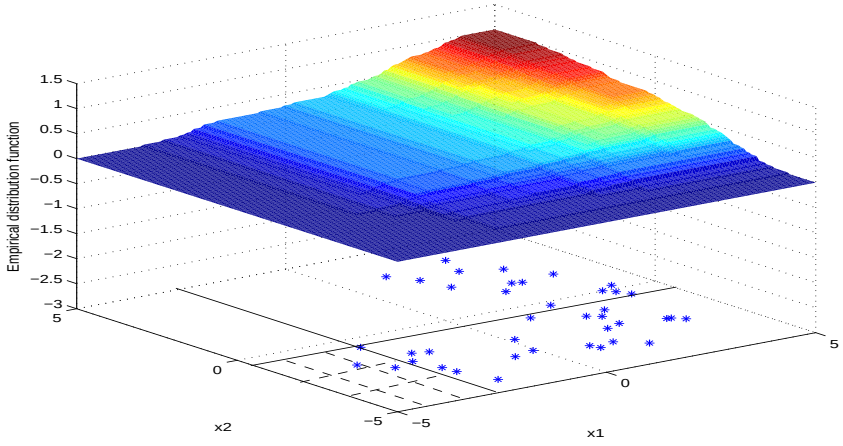
Given an unknown pdf function p, the *error* $\mathsf{err}(\hat{p})$ of an approximate pdf function $\hat{p}$ is defined as

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

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



For this example if f is the indicator function of the downward closed set defined by the two unbroken lines then

$$\hat{\mathbb{E}}[f] = \frac{1}{m_x} \sum_{i=1}^{m_x} f(x_i) = \frac{3}{m_x}$$

Examples cont.

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

$$d_{\mathcal{A}}(P, P') = \sup_{A \in \mathcal{A}} |P(A) - P'(A)|$$

measuring the distance between distributions P and P', since $P(A) = \int_X I_A(x) dP(x)$, where I_A is the indicator function of the set A.

Note that Glivenko-Cantelli theorems are concerned with convergence of empirical averages to true expectations over a class of functions. Hence, they can be seen as defining a Touchstone class that can be estimated by taking the empirical distribution.

Example 3

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 $p(\mathbf{x})$ over $\mathbf{x} \in \{0,1\}^n$. 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 $\mathbf{v} \in \{0,1\}^{|\mathbf{i}|}$

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

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

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 $\mathbf{v}$, i.e.

$$\mathbb{E}_p[\pi_{\mathbf{i},\mathbf{v}}] = p\left(\wedge_{j=1}^{|\mathbf{i}|} \mathbf{x}_{i_j} = \mathbf{v}_j \right)$$

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

- The distribution $\mathcal{P}_{\mathcal{F}}$ governs the likelihood of a 'test' function from the Touchstone class being chosen.
- In Example 1 (Mukherjee and Vapnik) the distribution $\mathcal{P}_{\mathcal{F}}$ mirrors the distribution of the data but in general they need not be related.
- 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 class $\mathcal{P}_{\mathcal{F}}$ of regression functions mapping from $\mathcal{F}$ to the reals, indexed by the elements of $\mathcal{P}$:

$$\mathcal{P}_{\mathcal{F}} = \left\{ f \in \mathcal{F} \mapsto \mathbb{E}_p[f] = \int_X f(x)p(x)dx \in \mathbb{R} : p \in \mathcal{P} \right\}.$$

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 $\mathcal{F}$.

Proof sketch

- The basic strategy is to consider the problem as one of learning a regressor from the set $\mathcal{P}_{\mathcal{F}}$.
- Hence, sample a suitably large set of m_f functions.
- Problem is that we aren't given the 'correct' output values $\mathbb{E}_p[f]$, so we need to use a suitable sample of inputs to get good enough estimates of these values, $\mathbb{E}_p[f] \approx \frac{1}{m_x} \sum_{i=1}^{m_x} f(x_i)$.
- Putting the two error bounds together ensures that the estimated distribution is a good approximator.

Support vector density estimation

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

- 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

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

- Note that without the 3rd and 4th lines (and the second term in the objective) this would be the optimisation for a 1-class SVM
- The extra constraints are to ensure a good fit between the empirical and estimated expectations for the sample of Touchstone functions $f_j, j = 1, \ldots, m_f$.

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

- This is for the kernel defined density estimation class and the bound ensures that learning will be effective.
- Note that there are two ways of measuring the complexity: $C_\kappa^2 \|f_i\|_{L_1}^2$ and $\|f_i\|_{L_1}\|f_i\|_{L_\infty}$.

Proof sketch

We can view the function class as a set of linear functions

$$\begin{aligned}
\mathbb{E}_{q_{\mathbf{w}}}[f] &= \int_{\mathcal{X}} q_{\mathbf{w}}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \\
&= \int_{\mathcal{X}} \langle \mathbf{w}, \phi(\mathbf{x}) \rangle f(\mathbf{x}) d\mathbf{x} \\
&= \int_{\mathcal{X}} \langle \mathbf{w}, f(\mathbf{x}) \phi(\mathbf{x}) \rangle d\mathbf{x} \\
&= \left\langle \mathbf{w}, \int_{\mathcal{X}} f(\mathbf{x}) \phi(\mathbf{x}) d\mathbf{x} \right\rangle ,
\end{aligned}$$

hence we can apply a known result bounding the Rademacher complexity of linear function classes.

Bounding SVDE

Theorem 3. *Suppose that we learn a pdf function based on a sample of m_x observations and a sample of m_f functions from the Touchstone Class $\mathcal{F}$. 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*

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

where L is the Lipschitz constant of the loss function.

Comparison with Glivenko-Cantelli

- It is worth noting that this result could be applied to all measurable functions whose indicator functions have L_1 norm bounded by a constant C .

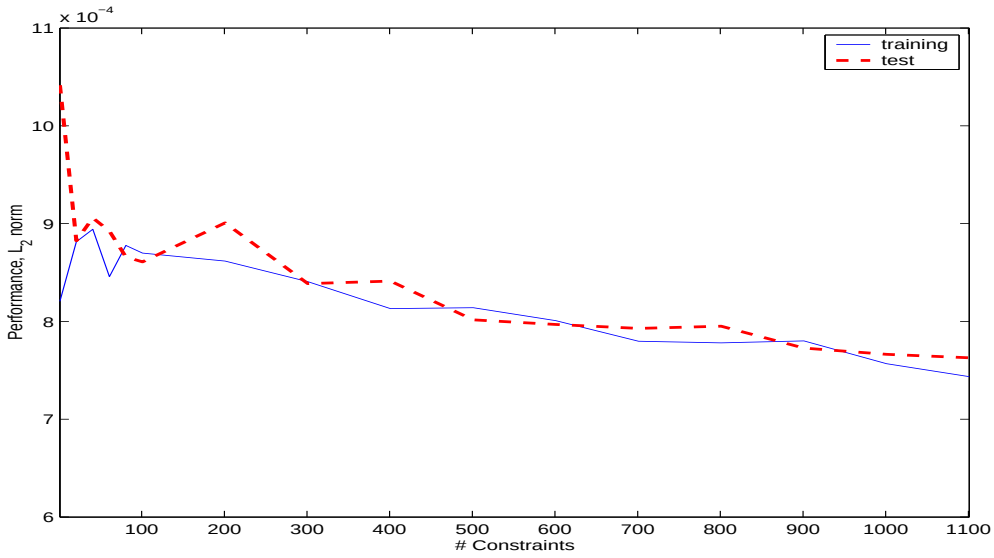
- The Glivenko-Cantelli result would not apply in this case since it requires bounded fat-shattering dimension of the class of functions.

- The quality of the approximation $\hat{\mathbb{E}}_f[\ell(\mathbb{E}_{\hat{p}}[f], \hat{\mathbb{E}}_x[f])]$ would however depend on the particular distribution and the class $\mathcal{P}_{\mathcal{F}}(B)$ used to approximate it.

- Hence, the result is not comparable with G-C, potentially allowing more complex sets of functions to be approximated.

Experiments with Half spaces

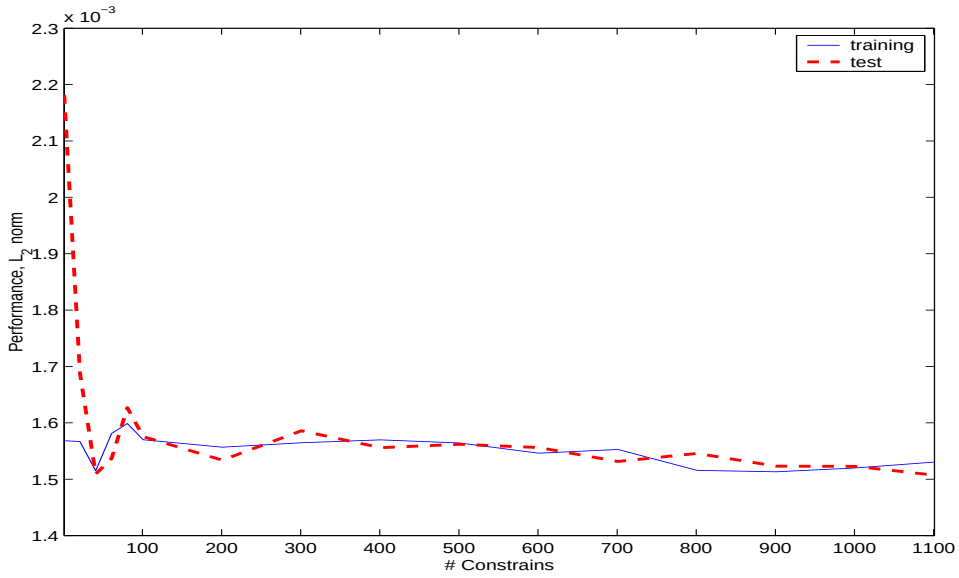
10 dimensional, 100 inputs generated by a mixture of Gaussians. Half spaces sampled using a Gaussian distribution.



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)

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

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