MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1653 April 1999
C.B.C.L Paper No. 170

Multivariate Density Estimation: a Support Vector Machine Approach

Sayan Mukherjee and Vladimir Vapnik

This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.
The pathname for this publication is: ai-publications/1500-1999/AIM-1653.ps

Abstract

A Support Vector Machine (SVM) algorithm for multivariate density estimation is developed based on regularization principles and bounds on the convergence of empirical distribution functions. The algorithm is compared to Gaussian Mixture Models (GMMs). Our algorithm outperforms GMMs for data drawn from mixtures of gaussians in $\mathbb{R}^2$ and $\mathbb{R}^6$. Our algorithm is also automated with respect to parameters.

This report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences and at the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology. This research is sponsored by a grant from the Office of Naval Research under contract No. N00014-93-13085. Additional support is provided by Eastman Kodak Company, Daimler-Benz AG, Siemens Corporate Research Inc., AT&T, Digital Equipment and Central Research Institute of Electric Power Industry.

1 Introduction

Estimating probability densities from a set of observed data points is a basic problem in statistics. In this paper we formulate a novel algorithm for multivariate density estimation that is mathematically well-founded, shows promise in preliminary experiments, and is automated with respect to parameter setting.
Traditional approaches to estimating densities have been parametric. One assumes that the data are drawn from a parametric family of distributions and one then estimates the parameters of the family based upon the maximum likelihood principle. Gaussian Mixture Models (GMMs) are a particular case of the parametric approach. It is well known, even by peeople using this approach, that it is mathematically problematic [1] [11].
The problem of density estimation is an inverse operator problem which is ill-posed and stochastic. These problems are amenable to regularization approaches. The classical approach of the Parzen's window technique [6] can be derived from Tikhonov regularization [9], unfortunately this technique has not yielded good results for high-dimensional problems. We apply the residual method [7] of regularization to the density problem where the residual is set based on bounds on the convergence of empirical distribution functions. Setting the error functional in the residual method to a norm in a Reproducing Kernel Hilbert Space (RKHS) results in a Support Vector Machine (SVM) formulation. Previously, an SVM approach to density estimation was attempted [13], however it was only applied to univariate density estimation and performed very badly for the multivariate problem. Our algorithm works for both multivariate and univariate problems. We also present experimental results that show better performance than a GMM approach for 2 and 6 dimensional data.

2 A Support Vector Machine (SVM) approach

2.1 Density estimation

A probability density, $p(t)$, is defined as the solution of the following equation

$$\int_{-\infty}^{x} p(t)\, dt = F(x), \tag{1}$$

where $F(x)$ is the distribution function $F(x) = \mathbb{P}\{\xi < x\}$, ξ is a random variable. Estimating a probability density from data means solving this integral equation on a given set of densities $p(t, \alpha)$, $\alpha \in \Lambda$, when the distribution function $F(x)$ is unknown and given a random independent sample $\{x_i\}_{i=1}^{\ell}$ obtained from this distribution.
The empirical distribution function

$$F_\ell(x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \Theta(x - x_i), \tag{2}$$

is a good approximation of the actual distribution function $F(x)$. The rate of convergence of this approximation is known asymptotically. For the univariate case, the random variable k

$$k = \sqrt{\ell} \sup_x |F(x) - F_\ell(x)| = \sqrt{\ell} ||F(x) - F_\ell(x)||_\infty,$$

is independent of the distribution function.
Hence one can consider the problem of density estimation as a problem of solving equation (1) using $F_\ell(x)$ instead of $F(x)$. $F_\ell(x)$ converges to $F(x)$ at a fast rate, $O(1/\sqrt{\ell})$.

2.2 The density estimation problem is stochastic and ill-posed

It is known that the problem of solving linear integral equation

$$Af = F$$

is ill-posed when the sequence $F_\ell(x)$ of approximations are used on the right hand side instead of $F(x)$.

Two forms of regularization can be used to solve ill-posed problems using approximations $||F(x) - F_\ell(x)|| = \delta_\ell$. The idea of these methods is to introduce the so called regularizing functional $\Omega(f)$ (semi-continuous, positive functional for which $\Omega(f) \leq c$ is a compactum for all positive c) and define the solution f which is a trade-off between the value of the functional $\Omega(f)$ and accuracy $||Af - F_\ell||$.

The two forms of this trade-off or regularization turn out asymptotically equivalent [12], they are the methods of Tikhonov [9] and Phillips [7] respectively,

$$\min_f \left(||Af - F_\ell||^2 + \gamma_\ell \Omega(f) \right), \quad \gamma_\ell > 0,$$

$$\min_f \Omega(f) \quad s.t. \quad ||Af - F_\ell|| < \varepsilon_\ell, \quad \varepsilon_\ell > 0,$$

where for certain conditions on constants δ_ℓ, γ_ℓ, and ε_ℓ the sequence of approximations converges to the desired solution.

Both principles can be generalized for the stochastic case [11], in particular for the Tikhonov method it was shown that if function $F_\ell(x)$ converges in probability to $F(x)$ and $\gamma_\ell \to 0$ then for any positive ν and μ there exists a positive number $n(\nu, \mu)$ such that for $\ell > n(\nu, \mu)$ the following inequality holds

$$P(\rho_{E_1}(f, f_\ell) > \nu) \leq P(\rho_{E_2}(F, F_\ell) > \sqrt{\gamma_\ell \mu}) \tag{3}$$

where $\rho_{E_1}(f, f_\ell)$, $\rho_{E_1}(f, f_\ell)$ are metrics in the space of functions f and F.

Since the empirical distribution function $F_\ell(x)$ converges in probability to $F(x)$ from equation (3) one concludes that estimating the density by solving integral equation (1) using $F_\ell(x)$ is always consistent.

The main problem in density estimation using a finite number of observations is specifying the regularization parameters.

2.3 Regularization method for density estimation

We solve the density estimation problem using the Philips' regularization form. We minimize the regularization functional $\Omega(f)$ subject to constraint

$$||Af - F_\ell|| \leq \sigma_\ell,$$

where σ_ℓ is the known discrepancy between $F(x)$ and $F_\ell(x)$.

Usually it is not easy to evaluate σ_ℓ. However for the problem of density estimation one can obtain an exact estimate of σ_ℓ.

As we discussed in section 2.1 for the one dimensional case the random variable $k = \sqrt{\ell}||F(x) - F_\ell(x)||_\infty$ has an universal distribution. Therefore one can choose $\sigma_\ell = \frac{c}{\sqrt{\ell}}$ where c is a constant, for example the median of this distribution 0.6. For the multivariate case with probability $1 - \eta$ the inequality

$$\sigma_\ell \leq c(\eta) \left(\ell^{-k(d)}\right)$$

holds true where $k(d)$ is defined by the dimensionality d of the space [4] [8]. These types of bounds are used to set the regularization parameter.

2.4 SVM for density estimation

To use the support vector method for the density estimation problem we look for a solution f in the set of functions that belong to some Reproducing Kernel Hilbert Space (RKHS) where we define the regularization functional $\Omega(f)$ as a norm in the RKHS

$$\Omega(f) = ||f||_H^2. \tag{4}$$

We minimize the functional (4) subject to constraints in a set of functions $f \in H$. It is known [3] that the solution of the optimization problem for $f \in H$ has a form

$$f(x) = \sum_{i=1}^{\ell} \alpha_i K(x_i, x)$$

where the coefficients minimize the functional

$$\Omega(\alpha) = \sum_{i=1}^{\ell} \alpha_i^2 \tag{5}$$

subject to constraints

$$||\hat{F} - F_\ell||_\infty \leq \sigma,$$

where $\hat{F}(x) = \int_\infty^x \hat{p}(t)dt$. Minimizing (5) can be thought of as finding the smoothest function $\hat{F}(x)$ which is within an σ-tube of $F_\ell(x)$.
A diagnostic is needed to check whether an approximation $\hat{F}(x)$ exists that stays within the σ-tube. This is done by adding slack variables, $\xi_i \geq 0$. The addition of the slack variables to functional (5) and properties of the RKHS give us the following functional to minimize (which is identical to the functional minimized for SVMs for inverse operator problems [10])

$$\min_{\alpha,\xi_i} ||\alpha||^2 + C \sum_{i=1}^{\ell} \xi_i \tag{6}$$

subject to

$$|\hat{F}(x_i) - F_l(x_i)| \leq \sigma + \xi_i, \ \forall\ x_i, \tag{7}$$

where ξ_i are slack variables, C is set to infinity (in implementations C is set to a large number), and since we want to estimate a density the following constraints are added:

$$\sum_{i=1}^{\ell} \alpha_i = 1 \ \ , \alpha_i \geq 0. \tag{8}$$

Minimizing functional (6) with respect to the constraints in equations (7) and (8) is a quadratic programming problem with linear constraints and can be solved in either a primal or dual formulation [2]. The solution has the form

$$\hat{p}(t, \alpha) = \sum_{i=1}^{\ell} \alpha_i K(t, x_i, \lambda),$$

where $K(t, x_i, \lambda)$ are kernel functions and λ is another regularization parameter and most of the α_i's are typically zero.
The extension to the the multivariate case, $\mathbf{x}_i \in \mathbb{R}^d$, only involves constructing a multivariate empirical distribution function

$$F_\ell(\mathbf{x}) = \sum_{i=1}^{\ell} \prod_{j=1}^{d} \Theta(\mathbf{x}^j - \mathbf{x}_i^j), \tag{9}$$

where $\mathbf{x}_i^j$ is the j^{th} component of the point $\mathbf{x}_i$. Again the solution will have the form

$$\hat{p}(\mathbf{t}, \alpha) = \sum_{i=1}^{N} \alpha_i K(\mathbf{t}, \mathbf{x}_i, \lambda), \tag{10}$$

where N are the points $\mathbf{x}_i$ for which α_i is nonzero, these are the Support Vectors, in general $N \ll \ell$. The parameter λ, an example of which is the variance of a gaussian, is set by a line search: we select the largest λ for which there exists an estimate within a σ-tube of the empirical distribution function.
Note that this algorithm has only one free parameter σ which is set based on bounds on the convergence of $F_\ell(\mathbf{x})$ to $F(\mathbf{x})$.

2.5 Comparison to Gaussian Mixture Models (GMMs)

When the kernel functions are gaussians the SVM solution is a linear combination of gaussians just like the GMM case. The difference between the two cases is how the inverse operator problem is solved. In the GMM case one uses the expectation maximization (EM) algorithm to estimate the parameters that maximize the log-likelihood. However, this maximum may not exist, from Basu and Micchelli "*Generally a maximum likelihood does not exist.*" [1]. Regularization heuristics are imposed to deal with this problem. The two basic heuristics are to set a lower bound on the variance of the gaussians and also to specify an upper bound on the number of gaussians in the model. The problem is that there is no rigorous way to set these bounds which serve as the regularization parameters.
Both the SVM and GMM methods require regularization to effectively solve the density estimation problem. The problem for the GMM approach is that it is difficult to set the regularization parameters in a mathematically rigorous way. In the SVM method the only free regularization parameter σ_ℓ is set based upon the convergence properties of empirical distribution functions.

3 Simulations

In this section we apply our algorithm to a 2-dimensional density estimation problem and a 12-dimensional density estimation problem. We compare the results to those achieved by a GMM.

3.1 2-dimensional case

Training data for 100 trials was generated by randomly sampling 200 data points for each trial from the following distribution

$$p(x, y) = \frac{1}{8\pi} e^{-((x-1)^2/2+(y-1)^2/8)} + \frac{1}{8\pi} e^{-((x+1)^2/8+(y+1)^2/2)}.$$

For the test data 1000 points were sampled from the above distribution. We applied our algorithm using a gaussian kernel

$$K(\{t_x, t_y\}, \{x_i, y_i\}, \lambda) = \frac{1}{2\pi\lambda^2} e^{-((t_x-x_i)^2+(t_y-y_i)^2)/2\lambda^2},$$

and $\sigma = \frac{.6}{\sqrt{200}}$. We compared our algorithm to that of a GMM. The GMM algorithm consisited of a vector quantization and k-means clustering step followed by the expectation maximization (EM) algorithm to compute the maximum likelihood. The GMM required as parameters the maximum number of gaussians allowed and the minimum variance allowed. We set the maximum number of gaussians to 4 and 2 and the minimum variance allowed to .01. For each trial we computed the ℓ^1 error or average absolute error between the estimate and true distribution. Figure (1a) shows that our algorithm does slightly better than the GMM with two gaussians and much better than the GMM with four gaussians. Figure (2) plots the true density and the three estimated densities for the first trial.

3.2 6-dimensional case

Training data for 50 trials was generated by randomly sampling 600 data points for each trial from the following distribution

$$\begin{aligned} p(\mathbf{x}) &= \frac{1}{24\pi^3}\left(\frac{1}{\det|\Sigma_1|} e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma_1^{-1}(\mathbf{x}-\mu_1)} + \frac{1}{\det|\Sigma_2|} e^{-\frac{1}{2}(\mathbf{x}-\mu_2)^T\Sigma_2^{-1}(\mathbf{x}-\mu_2)}\right. \\ &\left. + \frac{1}{\det|\Sigma_3|} e^{-\frac{1}{2}(\mathbf{x}-\mu_3)^T\Sigma_3^{-1}(\mathbf{x}-\mu_3)}\right) \end{aligned} \tag{11}$$

where $\mathbf{x} \in \mathbb{R}^6$, the covariance matrices are diagonal with the following elements
$\mathbf{\Sigma}_1 = \{1.0, 2.0, 1.0, 2.0, 1.0, 2.0\}$,
$\mathbf{\Sigma}_2 = \{2.0, 1.0, 2.0, 1.0, 2.0, 1.0\}$,
$\mathbf{\Sigma}_3 = \{2.0, 1.0, 2.0, 1.0, 2.0, 1.0\}$,
$\mu_1 = \{1.0, 1.0, 1.0, 1.0, 1.0, 1.0\}$,
$\mu_2 = \{-1.0, -1.0, -1.0, -1.0, -1.0, -1.0\}$,
$\mu_3 = \{0.0, 0.0, 0.0, 0.0, 0.0, 0.0\}$.
For the test data 6000 points were sampled from the above distribution. We applied our algorithm using a gaussian kernel

$$K(\mathbf{t}, \mathbf{x}_i, \lambda) = \frac{1}{8\pi^3\lambda^6} e^{-||\mathbf{t}-\mathbf{x}_i||^2/2\lambda^2},$$

and $\sigma = \frac{.5}{600^{1/3}}$. We again compared our algorithm to that of a GMM. We set the maximum number of gaussians to 8 and 4 and the minimum variance allowed to .01. For each trial we computed the ℓ^1 error or average absolute error between the estimate and true distribution. Figure (1b) shows that our algorithm does slightly better than the GMM with four gaussians and much better than the GMM with eight gaussians.

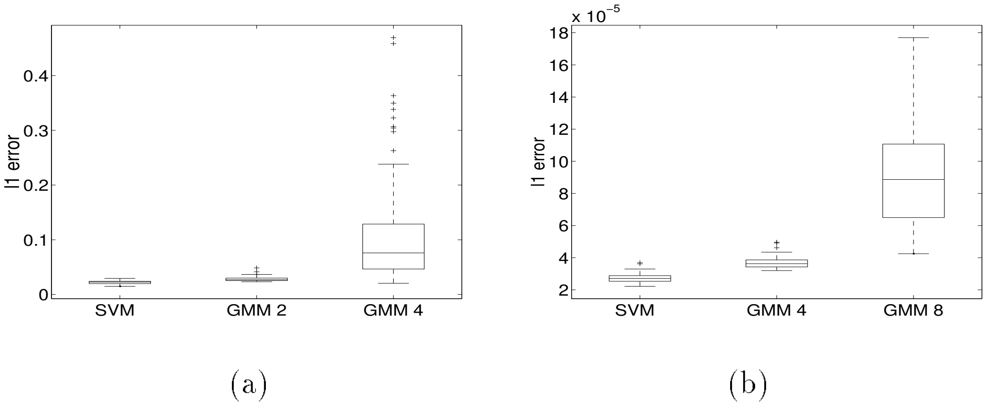

Figure 1: A boxplot of (a) the ℓ^1 error over 100 trials for SVM, GMM with 2 gaussians, and GMM with 4 gaussians in the two dimensional case and (b) the ℓ^1 error over 50 trials for SVM, GMM with 4 gaussians, and GMM with 8 gaussians in the six dimensional case.

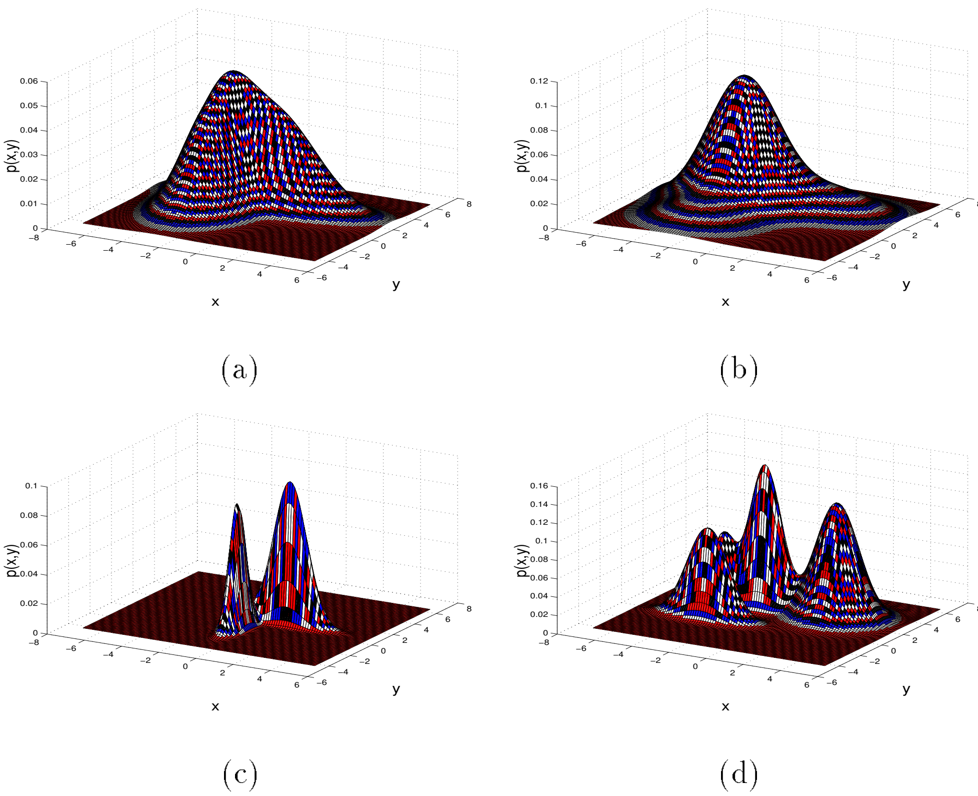

Figure 2: For the first trial in the two dimensional case (a) the true density (b) the SVM estimate (c) the GMM estimate with 2 or fewer gaussians (d) the GMM estimate with 4 or fewer gaussians.

4 Future work

In order to apply this algorithm to high dimensional problems with many data points in the training set we will implement the dual formulation using the decomposition algorithm of Osuna et. al. [5]. We also need a better understanding of which bounds to use to obtain the σ parameter. We will look at speech data as well as simulated data to examine general convergence behavior for practical problems so as to determine this parameter.

References

[1] S. Basu and C.A. Micchelli. Parametric density estimation for the classification of acoustic feature vectors in speech recognition. In *Nonlinear Modeling, Advanced Black-Box Techniques*. Kluwer Publishers, 1998.

[2] M.S. Bazarra, H.D. Sherali, and C.M. Shetty. *Nonlinear Programming Theory and Algorithms*. John Wiley, New York, NY, 1993.

[3] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10:1455–1480, 1998.

[4] P. Massart. Strong approximation for multivariate empirical and related processes, via k.m.t. constructions. *Ann. Probab.*, 21:266–291, 1989.

[5] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *IEEE Neural Network for Signal Processing*, 1997.

[6] E. Parzen. On estimation of a probability density function and mode. *Ann. Math. Statis.*, 33:1065–1076, 1962.

[7] D.L. Phillips. A technique for the numerical solution of integral equations of the first kind. *J.Assoc. Comput. Machinery*, 9:84–97, 1962.

[8] E. Rio. Local invariance principles and its application to density estimation. *Probabilty Theory and Related Fields*, 98:21–45, 1994.

[9] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038, 1963.

[10] V. Vapnik, S.E. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In *Advances in Neural Information Processings Systems 9*, pages 281–287, San Mateo, CA, 1997. Morgan Kaufmann Publishers.

[11] V. N. Vapnik. *Statistical learning theory*. J. Wiley, 1998.

[12] V.V. Vasin. Relationship of several variational methods for the approximate solution of ill-posed problems. *Math Notes*, 7:161–166, 1970.

[13] J. Weston, A. Gammerman, M. Stitson, V. Vapnik, V. Vovk, and C. Watkins. *Support Vector Density Estimation*. M.I.T. Press, Cambridge, MA, 1999.