One-Class Support Vector Machines and Density Estimation: The Precise Relation

Alberto Muñoz[1] and Javier M. Moguerza[2]

[1] University Carlos III, c/ Madrid 126, 28903 Getafe, Spain
alberto.munoz@uc3m.es
[2] University Rey Juan Carlos, c/ Tulipán s/n, 28933 Móstoles, Spain
j.moguerza@escet.urjc.es

Abstract. One-Class Support Vector Machines (SVM) afford the problem of estimating high density regions from univariate or multivariate data samples. To be more precise, sets whose probability is specified in advance are estimated. In this paper the exact relation between One-Class SVM and density estimation is demonstrated. This relation provides theoretical background for the behaviour of One-Class SVM when the Gaussian kernel is used, the only case for which successful results are shown in the literature.

1 Introduction

Density estimation [14] arises explicitly in a number of pattern recognition tasks involving interesting problems such as outlier (novelty) detection [8, 11, 15] or cluster analysis (see for instance [10, 5]). The density estimation task can be regarded as a particular type of inverse problem. In this setting, we consider a mapping $H_1 \xrightarrow{A} H_2$, where H_1 represents a metric function space and H_2 represents a metric space in which the observed data (which could be functions) live. In the density estimation problem, H_1 and H_2 are both function spaces and A is a linear integral operator given by: $(Af)(x) = \int K(x,y)f(x)dx$, where K is a predetermined kernel function and f is the density function we are seeking. The problem to solve is $Af = F$, where F is the distribution function. As far as F is unknown, the empirical distribution function F_n is used instead, where n is the number of data points, and the inverse problem to solve is $Af = y$, with $y = F_n$. Within the framework of regularization theory [16], if H_1 is chosen as a reproducing kernel Hilbert space (RKHS) [2], by the representer theorem [9], the estimator $\hat{f}$ of f takes the form $\hat{f}(x) = \sum_{i=1}^{n} c_i K(x, x_i)$. Taking $K(x, x_i) = K_h(x, x_i) = e^{-\|x-x_i\|^2/h}$, we obtain the well-known kernel density estimator with Gaussian kernel (see [14]), where each $c_i = 1/nh^d$, $h > 0$ and d is the data dimension.

One-Class Support Vector Machines [13, 15] are designed to solve density estimation related problems with tractable computational complexity.

The concrete problem to solve is the estimation of minimum volume sets of the form $S_\alpha(f) = \{x | f(x) \geq \alpha\}$, such that $P(S_\alpha(f)) = 1 - \nu$, where f is the

A. Sanfeliu et al. (Eds.): CIARP 2004, LNCS 3287, pp. 216–223, 2004.

density function and $0 < \nu < 1$. These sets are known in the literature as density contour clusters at level α [4, 6]. One-Class SVMs deal with a problem related to that of estimating $S_\alpha(f)$. The method computes a binary function that takes the value $+1$ in 'small' regions containing most data points and -1 elsewhere.

The rest of the paper is organized as follows: in Section 2 One-Class SVMs are briefly described. Section 3 makes explicit the relation between One-Class SVMs and classic density estimation. In Section 4 experiments that corroborate the theoretical findings are shown. Section 5 concludes.

2 One-Class SVMs in a Nutshell

The strategy of One-Class support vector methods is to map the data points into the feature space determined by the kernel function, and to separate them from the origin with maximum margin. Thus, it follows the general scheme of SVMs.

In order to build a separating hyperplane between the origin and the mapped points $\{\Phi(x_i)\}$, the One-Class SVM method solves the following quadratic optimization problem:

$$\begin{aligned} \min_{w,\rho,\xi} \; & \frac{1}{2}\|w\|^2 - \nu n\rho + \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \; & \langle w, \Phi(x_i)\rangle \geq \rho - \xi_i \,, \\ & \xi_i \geq 0, \qquad\qquad i = 1, \ldots, n \,, \end{aligned} \tag{1}$$

where Φ is the mapping defining the kernel function, ξ_i are slack variables, $\nu \in [0, 1]$ is an a priori fixed constant which represents the fraction of outlying points, and ρ is the decision value which determines if a given point belongs to the estimated high density region. The decision function will take the form $h(x) = sign(w^{*T}\Phi(x) - \rho^*)$, where w^* and ρ^* are the values of w and ρ at the solution of problem (1).

In [13] the mapping induced by the exponential (Gaussian) kernel $K_c(x, y) = e^{-\|x-y\|^2/c}$ is used. This kernel maps the data onto the unit hypersphere within the positive orthant. Figure 1 illustrates the situation.

In the following we will refer to 'quadratic One-Class SVM' simply as 'One-Class SVM'. Notice the difference with linear One-Class SVM, which were stated in [12].

The dual problem of (1) (see [3] for details on the derivation of the dual formulation) is:

$$\begin{aligned} \max_{\alpha} \; & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \\ \text{s.t.} \; & \sum_{i=1}^{n} \alpha_i = \nu n \,, \\ & 0 \leq \alpha_i \leq 1 \,, \qquad\qquad i = 1, \ldots, n \,. \end{aligned} \tag{2}$$

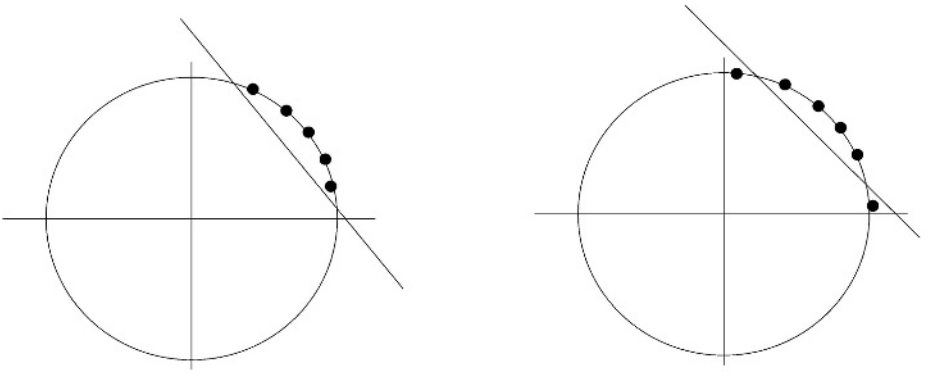

Fig. 1. Left: separating hyperplane in feature space. Right: the same for a percentage of points.

Points x_i such that at the solution of problem (2) satisfy $\alpha_i > 0$ are called *support vectors*. It can be shown that $h(x_i) > 0$ for the non-support vector points, that is, those points such that $\alpha_i = 0$ at the solution of problem (2).

The numerical results in [13] show that, for the exponential kernel, the performance of the method is similar to that of a kernel density estimator (Parzen windows).

3 Density Estimation and One-Class SVMs

In this section we show the strong relation that exists between One-Class SVM and kernel density estimation. This relation provides theoretical background for the behaviour of One-Class SVM with the exponential kernel (also known as Gaussian or RBF kernel), the only case illustrated with examples in [13]. In that work a relation in terms of (loose) probability bounds is given but, as stated by its authors, the relation is not conclusive.

The exponential kernel is a Mercer's kernel; therefore there exists a map $\phi : \mathbb{R}^d \longrightarrow \mathbb{R}^\infty$ such that $K(x, y) = \phi(x)^T \phi(y)$. Denote by d^* the distance induced by the kernel K in the feature space: $d^*_{ij} = d(\phi(x_i), \phi(x_j)) = \|\phi(x_i) - \phi(x_j)\|_K$. Considering that $d^{*2}_{ij} = K_{ii} + K_{jj} - 2K_{ij}$ and $K_{ii} = 1$ for the exponential kernel, a direct calculation shows that $K_{ij} = 1 - d^{*2}_{ij}/2$.

Hence, the One-Class SVM in its dual formulation (2) for the exponential kernel can be stated as the following equivalent optimization problem:

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}(\nu n)^2 - \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j d^{*2}_{ij} \\ &\text{s.t. } \sum_{i=1}^{n}\alpha_i = \nu n\,, \\ &\quad 0 \le \alpha_i \le 1\,, \quad i = 1, \ldots, n\,. \end{aligned} \tag{3}$$

Note that in the One-Class SVM problem with exponential kernel ν represents the fraction of outlying points (see [13]).

In order to minimize the objective function of problem (3), the term $\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j d_{ij}^{*2}$ has to be maximized. As a consequence of Proposition 4 in [13], as $n\to\infty$, the α_i's become 1 or 0. Thus, we have to choose a subset of νn points from the sample such that the sum of the distances among their images in the feature space is maximized. Equivalently, we can find the $(1-\nu)n$ points such that the sum of distances among their images in the feature space is minimized. For notational simplicity, assume these points are the first $(1-\nu)n$ in the sample.

Lemma 1. *Given n points x_i in a metric space, the following equality holds: $\sum_i\sum_j d^2(x_i,x_j) = 2n\sum_i d^2(x_i,\bar{x})$, where $\bar{x}$ stands for the sample mean.*

Proof. $\sum_i\sum_j d^2(x_i,x_j) = \sum_i\sum_j \|x_i-x_j\|^2 = \sum_i\sum_j \|x_i-\bar{x}+\bar{x}-x_j\|^2 = \sum_i\sum_j \left(\|x_i-\bar{x}\|^2+\|x_j-\bar{x}\|^2\right) - 2\sum_i\sum_j (x_i-\bar{x})^T(x_j-\bar{x})$ and the last term becomes zero by definition of sample mean: $\sum_j(x_j-\bar{x})=0$. □

By the preceding lemma and Proposition 4 in [13], as $n\to\infty$,

$$\sum_{i=1}^{(1-\nu)n}\sum_{j=1}^{(1-\nu)n} d_{ij}^{*2} = 2(1-\nu n)\sum_{i=1}^{(1-\nu)n} d^{*2}(\phi(x_i),\overline{\phi(x)}),$$

where $\overline{\phi(x)}$ stands for the average of the mapping $\phi(x_i)$ of the $(1-\nu)n$ points. This quantity will be minimized choosing the $(1-\nu)n$ points closest to their average.

Thus, we have proved the following theorem.

Theorem 1. *Consider the One-Class SVM with the exponential kernel. Asymptotically, the points obtained as non-support vectors correspond to those whose sum of distances to their mean in the feature space is minimum.*

The next theorem relates kernel density estimation with One-Class SVM.

Theorem 2. *Consider the One-Class SVM with the exponential kernel. Asymptotically, the points obtained as non-support vectors are those closest to the mode estimator calculated on this set of $(1-\nu)n$ non-support vectors, using a kernel density estimator with Gaussian kernel (Parzen windows) .*

Proof. Again for notational simplicity, assume the non-support vector points are the first $(1-\nu)n$ points in the sample. Consider the kernel density estimator with exponential kernel $\hat{f}(x) = \frac{1}{(1-\nu)nh^d}\sum_{i=1}^{(1-\nu)n} K_h(x,x_i)$ where $K_{xx_i} = K_h(x,x_i) = e^{-\|x-x_i\|^2/h}$. Since $K_{xx_i} = 1 - 1/2d^{*2}(\phi(x),\phi(x_i))$, a simple calculation shows that $\hat{f}(x) = \frac{1}{h^d} - \frac{1}{2nh^d}\sum_{i=1}^{(1-\nu)n} d^{*2}(\phi(x),\phi(x_i))$.

Now consider the mean of the $(1-\nu)n$ non-support vector points, $\overline{\phi(x)}$.

Since $\frac{1}{h^d} - \frac{1}{2nh^d}\sum_{i=1}^{(1-\nu)n} d^{*2}(\overline{\phi(x)},\phi(x_i))$ is maximum (see the proof of Theorem 1), Theorem 1 and standard continuity arguments guarantee that the points

obtained as non-support vectors will be the points nearest to the maximum of $\hat{f}(x)$. □

Statistical properties of the mode estimator using kernel density estimators have been studied in [7]. In particular, the estimator is consistent.

Remark 1. In case of existence of ϕ^{-1} (which is not guaranteed), the anti-image through ϕ, $x^* = \phi^{-1}(\overline{\phi(x)})$ would be the mode estimator. In fact, since $\hat{f}(x^*) = \frac{1}{h^d} - \frac{1}{2nh^d}\sum_{i=1}^{(1-\nu)n} d^{*2}(\overline{\phi(x)}, \phi(x_i))$ and by Theorem 2, the second term of $\hat{f}(x^*)$ is minimum; being the first term a constant, $\hat{f}(x^*)$ will be maximum.

Remark 2. The kernel density estimator $\hat{f}(x)$ relies critically on the value of the smoothing parameter h. Therefore the performance of the One-Class SVM will critically depend on a good choice of such parameter, and on the solution of the optimization problem itself.

4 Experiments

The main aim of the paper is to provide a deeper understanding of One-Class SVMs, by demonstrating its relation to already existing density estimation techniques. Anyhow, for the sake of completeness, next we show a couple of applications derived from the previous theoretical results.

4.1 An Example of Biased Behaviour

The asymptotical result in Theorem 1 suggests a suboptimal performance for One-Class SVM with asymmetrical data for non-huge data sets: for spherically symmetric distributions, asymptotically, the average of the whole set of points will converge to the true mean (which coincides with the mode) and so will happen with the $(1-\nu)n$ points closest to their average. This can not be guaranteed for asymmetric distributions. To check the behaviour of One-Class SVM in this case, we have generated 2000 points from a gamma $\Gamma(\alpha, \beta)$ distribution, with $\alpha = 1.5$ and $\beta = 3$. Figure 2 shows the histogram, the gamma density curve, the true mode $(\alpha - 1)/\beta$ as a bold vertical line, and the One-Class SVM (five lines) estimations of the 50% highest density region. The parameters have been chosen applying the widely used rule $c = hd$ in $K_c(x, y)$, where $h \in \{0.1, 0.2, 0.5, 0.8, 1.0\}$ and d is the data dimension (see for instance [13]). The bias is apparent, since none of the five estimated support sets contains the true mode (and they should).

4.2 Improving One-Class SVM Performance

To illustrate a practical consecuence of Remark 2 next we show an example from the pattern recognition field, where the choice of the parameter c of kernel $K_c(x, y)$ is crucial. The database used contains nearly 4000 instances of handwritten digits from Alpaydin and Kaynak [1]. Each digit is represented by a

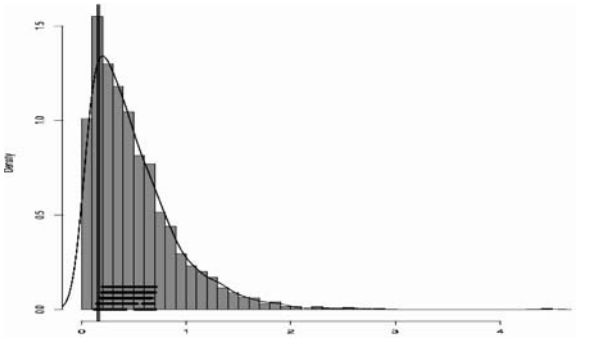

Fig. 2. Gamma sample with 2000 points. The figure shows the histogram, the density curve, a vertical line at the true mode, and One-Class SVM (five lines) estimations of the 50% highest density region.

vector in $\mathbb{R}^{64}$ constructed from a 32×32 bitmap image. Figure 3 shows a sample from the data base. The calligraphy of the digits in the database seems to be easily perceivable, which is supported by the high success rate of various classifiers. In particular, for each digit, nearest neighbour classifiers accuracy is always over 97% [1]. From this database we have selected a set of 409 data points made up by the 389 instances of digit '3' and the first 20 instances of digit '4' (approximattely 5% of the selected sample). The underlying hypothesis is that the support of the data is constituted by instances corresponding to digit '3', while the outlying points should correspond to instances of digit '4'.

We have run a set of experiments applying the rule for the choice of c described in the previous example. In the five experiments using this rule none of the outlying digits was detected by the One-Class SVM.

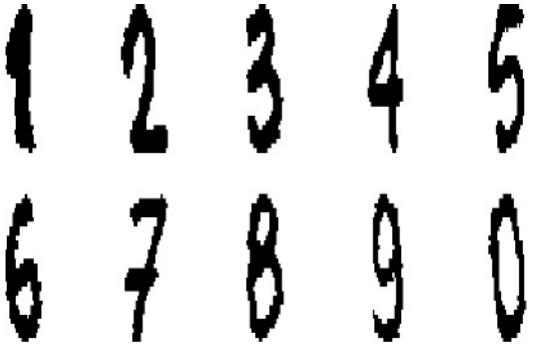

Fig. 3. A sample of the Alpaydin and Kaynak digit data base.

In order to improve this behaviour, and taking into account the result in Theorem 1, next we give a simple rule to choose the parameter c. This rule tries to minimize the numerical errors arising from the use of an exponential function. Thus we choose $c = \max\{d_{ij}^2\}$, where d_{ij} stands for the Euclidean distance between data points x_i and x_j. This value implies that the argument inside the exponent of $K_c(x_i, x_j)$ will be upper bounded for the data set, avoiding as far as possible numerical errors. Using this rule, 50% of the outlying instances were detected by the One-Class SVM, which certainly represents a remarkable improvement. The results were similar when different pair of digits were used for the experiments.

5 Conclusions

One-Class Support Vector Machines (SVM) afford the important task of estimating high density regions from data samples, a problem strongly related to the classical problem of density estimation. In this paper we have clearly stated the relation that exists between One-Class SVM and kernel density estimation. This relation provides theoretical background for the suboptimal behaviour of One-Class SVM when the Gaussian kernel is used, which is corroborated in the paper with some data examples. Finally, a simple rule to fix the parameter of the Gaussian kernel is given.

Acknowledgments

This work was partially supported by Spanish grants BEC2000-0167 (DGICYT), TIC2003-05982-C05-05 (MCyT) and PPR-2003-42 (URJC).

References

1. E. Alpaydin and C. Kaynak. *Cascading Classifiers.* Kybernetika, 34(4):369-374, 1998.
2. N. Aroszajn. *Theory of Reproducing Kernels.* Transactions of the American Mathematical Society, vol. 68, Issue 3, 1950, pp. 337-404.
3. M.S. Bazaraa, H.D. Sherali and C.M. Shetty. *Nonlinear Programming: Theory and Algorithms, 2nd Ed.* Wiley, New York, 1993.
4. S. Ben-David and M. Lindenbaum. *Learning distributions by their density levels: a paradigm for learning without a teacher.* Journal of Computer and System Sciences, 55:171-182, 1997.
5. A. Ben-Hur, D. Horn, H. Siegelmann and V. Vapnik. *Support Vector Clustering.* Journal of Machine Learning Research, 2:125-137, 2001.
6. A. Cuevas and R. Fraiman. *A plug-in approach to support estimation.* The Annals of Statistics, 25(6):2300-2312, 1997.
7. L. Devroye. *Recursive estimation of the mode of a multivariate density.* The Canadian Journal of Statistics, 7(2):159-167, 1979.
8. L. Devroye and Wise, G. *Detection of abnormal behavior via nonparametric estimation of the support.* SIAM J. Appl. Math., 38:480-488, 1980.

9. G.S. Kimeldorf and G. Wahba. *A Correspondence between Bayesian Estimation on Stochastic Processes and Smoothing by Splines.* Annals of Mathematical Statistics, 2:495-502, 1971.
10. J.M. Moguerza, A. Muñoz and M. Martin-Merino. *Detecting the Number of Clusters Using a Support Vector Machine Approach.* Proc. ICANN 2002, LNCS 2415:763-768, Springer, 2002.
11. A. Muñoz and J. Muruzabal. *Self-Organizing Maps for Outlier Detection.* Neurocomputing, 18:33-60, 1998.
12. G. Rätsch, S. Mika, B. Schölkopf and K.R. Müller. *Constructing Boosting Algorithms from SVMs: an Application to One-Class Classification.* IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(9):1184-1199, 2002.
13. B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola and R.C. Williamson. *Estimating the Support of a High Dimensional Distribution.* Neural Computation, 13(7):1443-1471 , 2001.
14. B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 1990.
15. D.M.J. Tax and R.P.W. Duin. *Support Vector Domain Description.* Pattern Recognition Letters, 20:1991-1999, 1999.
16. A.N. Tikhonov and V.Y. Arsenin. *Solutions of ill-posed problems.* John Wiley & Sons, New York, 1977.