Support Vectors Algorithms as Regularization Networks

Andrea Caponnetto[1,2] Lorenzo Rosasco[2] Francesca Odone[2]
Alessandro Verri[2]

1- INFM - DISI, Università di Genova,
v. Dodecaneso 35,
16146 Genova, Italy

2- CBCL - Massachusetts Institute of Technology,
Bldg.E25-201, 45 Carleton St.,
Cambridge, MA 02142

Abstract. In this paper we show that several Support Vector methods, including one-class SVM and a number of non-standard SVM classification techniques, can be viewed as special implementations of a general regularization network. Formally, the connection is obtained by choosing the appropriate loss function and parametrized by the exponent of the offset in the penalty term. The mathematical properties of the underlying algorithms can then be more conveniently studied within the theoretical framework of regularization networks.

1 Introduction

As shown in [7] Support Vector Machines (SVM) can be seen as a particular instance of a wider class of learning algorithms usually called Regularization Networks (RN). In the RN formulation the original SVM problem is stated as the minimization problem of a regularized functional on a Reproducing Kernel Hilbert Space (RKHS).

Recently, several learning algorithms were inspired by SVMs. In the attempt to obtain algorithms easier to implement, for example, a number of non standard support vector methods have been proposed [**?**]. In addition, support vector methods for novelty detection (or one-class SVM [**?**]) have been introduced through an analogy with the geometrical intuition of the original SVM for binary classification.

In this paper we show that the algorithms of above can be regarded as regularization networks. As a result, we can formally justify the various methods within a unifying framework and analyze the mathematical properties of existence and uniqueness, explicit form of the solution, and generalization properties in a coherent setting.

The paper is organized as follows. In Section 2 we introduce the general formulation of RN and discuss the meaning of the offset term. In Section 3 we deal with the case of one-class SVM, while in Section 4 with several non standard SVMs. Finally, we draw the conclusions of our analysis in Section 5.

2 Binary Classification and Regularization Networks

Throughout the paper we assume some familiarity with the learning from examples problem. The notation is consistent with [12], while for comprehensive introduction see, for example [7, 17, 6].

A learning algorithm is a map that, given a training set D of $\{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ examples, with $\mathbf{x}_i \in X \subset \mathbf{R}^n$ and $y_i = \pm 1$ for all i drawn i.i.d. with respect some fixed but unknown probability distribution, provides us with a classification rule. The class of algorithms we consider are the so called Regularization Networks solution of the following minimization problem

$$\min_{f \in \mathcal{H}} \{ \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i)) + \lambda \|f\|_{\mathcal{H}}^2 \} \tag{1}$$

where the first term measures how well the function f fits the data and the second, the squared norm of f in the RKHS $\mathcal{H}$, is a penalty term controlling the complexity (or smoothness) of the solution. The parameter λ is the regularization parameter tradeoff between the two terms.

The function $V(y, f(\mathbf{x}))$ can be interpreted as a measure of the loss we incur when classifying by $f(\mathbf{x})$ a new input $\mathbf{x}$ the label of which is y. In the classification setting the loss function V usually depends on its arguments through the product $yf(\mathbf{x})$. This particular dependency rests on the implicit assumption that false negatives ($y = +1$ and $f(\mathbf{x}) < 0$) and false positives ($y = -1$ and $f(\mathbf{x}) > 0$) are equivalent. More general situations have been considered in the literature (see for example [11]) leading to describe a loss function as

$$V(y, f(\mathbf{x})) = L(y) V(yf(\mathbf{x})). \tag{2}$$

Different loss functions lead to different learning algorithms [7]. The choices we consider in the following are

- the square loss $V(y, w) = (w - y)^2 = (1 - wy)^2$,
- the hinge loss $V(y, w) = \max\{1 - wy, 0\} =: |1 - wy|_+$,
- the truncated square loss $V(y, w) = \max\{1 - wy, 0\}^2 =: |1 - wy|_+^2$.

In the most general case a regularization network can be obtained as the minimization of a functional written as

$$\min_{f \in \mathcal{H}, b \in \mathbf{R}} \{ \frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i) + b) + \lambda(\|f\|_{\mathcal{H}}^2 \; + b^c) \} \tag{3}$$

where b is an offset term and c a parameter.

For $c = 0$ the constant offset is not penalized. As shown in [5] this is equivalent to a functional minimization in which the penalty term is a semi-norm in a RKHS. For $c = 2$ the offset term can effectively be absorbed by redefining the kernel function as $K+1$ and leaving the rest unchanged. The case $c=1$ singles out the case of one-class SVM in which it is also necessary to make full use of the more general definition of loss function given in (2) due to the presence of examples from one class only in the training stage.

3 One-class Support Vector Machine

One-class classification techniques were originally developed to cope with binary classification problems in which statistics for one of the two classes was virtually absent ([16, 13]). In this setting the component of the training set $(\mathbf{x}_i, y_i)_{i=1}^{\ell}$ labeled according to the minority class ($y=-1$ in the following) is intentionally removed, generating the reduced one-class training set $(\mathbf{x}_i)_{i=1}^{\ell_+}$.

Intuitively the idea behind one-class SVM algorithm is to look for the smallest sphere enclosing the examples in data space. Hence, the training procedure amounts to the solution of the following constrained minimization problem with respect to balls of center $\mathbf{a}$ and radius R in input space

$$\min_{R,\xi_i}\{C\sum_{i=1}^{\ell_+}\xi_i + R^2\}, \tag{4}$$

s.t. the existence of a vector $\mathbf{a}$ for which $\forall\ i \leq \ell_+$

$$(\mathbf{x}_i - \mathbf{a})\cdot(\mathbf{x}_i - \mathbf{a})^T \leq R^2 + \xi_i, \quad \xi_i \geq 0.$$

A non-linear extension of the previous algorithm can be directly achieved by substituting scalar products with kernel functions. From a more geometrical point of view we consider balls in a suitable feature space. It can be shown (see for example [4]) that the centers $\mathbf{a}$ can be mapped one to one with the functions f of the RKHS $\mathcal{H}$ of kernel $K(\mathbf{x},\mathbf{s}) = \mathbf{\Phi}(\mathbf{x})\cdot\mathbf{\Phi}(\mathbf{s})$. This correspondence is such that $\mathbf{\Phi}(\mathbf{x})\cdot\mathbf{a} = f(\mathbf{x})$, so that the constraints of the problem (4) can be rewritten as the existence of a function $f \in \mathcal{H}$ for which $\forall\ i \leq \ell_+$

$$K(\mathbf{x}_i,\mathbf{x}_i) - 2f(\mathbf{x}_i) + \|f\|^2_{\mathcal{H}} \leq R^2 + \xi_i, \quad \xi_i \geq 0.$$

In the minimization problems above the relevant complexity measure R^2 runs over non negative real values. In order to interpret the above problems from a regularization point of view it results convenient slightly modify the problem allowing the penalty to run over unconstrained reals. This is obtained simply by replacing R^2 with the real number ρ. The two problems are essentially equivalent since it is straightforward to verify that whenever $C > \frac{1}{\ell_+}$ they have the same solution. On the other hand if $C < \frac{1}{\ell_+}$ both problems are trivial: for any solution of (4) we have $R=0$ while no solution exists if R^2 is replaced by ρ.

Introducing the offset b and the fixed bias function $\mathcal{D}(\mathbf{x})$, and suitably rescaling the parameter C and the kernel K, problem (5) in which R^2 is replaced by ρ becomes

$$\min_{f,b,\xi_i}\{\widetilde{C}\sum_{i=1}^{\ell_+}\xi_i + \frac{1}{2}(\|f\|^2_{\widetilde{\mathcal{H}}} + b)\}, \tag{5}$$

s.t. $\forall\ i \le \ell_+$

$$\mathcal{D}(\mathbf{x}_i) + f(\mathbf{x}_i) + b \ge -\xi_i, \quad \xi_i \ge 0,$$

with

$$\begin{aligned} b &= \frac{1}{2}(\rho - \|f\|^2_{\mathcal{H}} - K(\mathbf{0},\mathbf{0})), \\ \mathcal{D}(\mathbf{x}) &= \frac{1}{2}(K(\mathbf{x},\mathbf{x}) - K(\mathbf{0},\mathbf{0})), \\ \widetilde{C} &= \frac{1}{4}C, \\ \widetilde{K} &= 2K. \end{aligned}$$

Finally, problem (5) considering the loss function

$$V(y,w) = \theta(y) \cdot |-wy|_+, \tag{6}$$

which matches the general form of Eq.(2), can be rewritten as

$$\min_{f\in\widetilde{\mathcal{H}},b\in\mathbf{R}} \{\frac{1}{\ell}\sum_{i=1}^{\ell} V(y_i, \mathcal{D}(\mathbf{x}_i) + f(\mathbf{x}_i) + b) + \lambda(\|f\|^2_{\widetilde{\mathcal{H}}} + b)\}. \tag{7}$$

For translation invariant kernel functions (e.g. the Gaussian kernel), the fixed bias function $\mathcal{D}(\mathbf{x})$ vanishes and problem (7) fits the general regularization network form (3) with $c = 1$.

4 Non Standard SVMs Revisited

In the following each method is viewed in the primal formulation considering the general non-linear setting through the use of the feature map $\mathbf{\Phi}(\mathbf{x})$ (see [17]).

4.1 Constant Offset not in the Penalty Term

We first consider L2-SVM and least square SVM, algorithms in which the offset term is not penalized. We start off with *L2*-SVM (see [3] for reference) which looks for the solution of the problem

$$\min_{\mathbf{w},b,\xi_i} \{C\sum_{i=1}^{\ell} \xi_i^2 + \frac{1}{2}\langle\mathbf{w},\mathbf{w}\rangle\} \tag{8}$$

s.t. $\forall\ i$

$$y_i(\mathbf{w}\cdot\mathbf{\Phi}(\mathbf{x}_i) + b) \ge 1 - \xi_i,\ \xi_i \ge 0.$$

In the primal formulation the constrained minimization problem (8) can be rewritten using the *truncated square loss* as

$$\min_{f\in\mathcal{H},b\in\mathcal{B}} \{\frac{1}{\ell}\sum_{i=1}^{\ell} |1 - y_i(f(\mathbf{x}_i) + b)|^2_+ + \lambda\,\|f\|^2_{\mathcal{H}}\} \tag{9}$$

which is a specialization of the regularization network (3) with $c = 0$.

Similarly, *least-square* SVM (see [3] and [15] for reference) aims at solving the same problem bu replacing the inequality constraints with equality constraints. Therefore, a least square SVM looks for the solution of problem (8) s.t. $\forall\ i$

$$y_i(\mathbf{w} \cdot \mathbf{\Phi}(\mathbf{x}_i) + b) = 1 - \xi_i,\ \ \xi_i \geq 0.$$

It is easy to see that, in the primal formulation, this leads to a constrained minimization problem which is a specialization of the regularization network (3) with $c = 0$ in which the loss function is the *square loss*.

4.2 Constant Offset in the Penalty Term

Finally, we consider *(i)* modified SVM, *(ii)* smooth SVM, and *(iii)* proximal SVM, algorithms for which the constant offset appears explicitly in the penalty term.

In the *modified* SVM (see [9] for reference) the problem is

$$\min_{\mathbf{w},b,\xi_i} \{C \sum_{i=1}^{\ell} \xi_i + \frac{1}{2}(\langle \mathbf{w}, \mathbf{w} \rangle + b^2)\} \tag{10}$$

s.t. $\forall\ i$

$$y_i(\mathbf{w} \cdot \mathbf{\Phi}(\mathbf{x}_i) + b) \geq 1 - \xi_i,\ \ \xi_i \geq 0,$$

The equivalence with the regularization network of (3) with $c = 2$ with the *hinge loss* as loss function is readily established.

The only difference of *smooth* SVM (see [10] for reference) is in the squares of the slack variables leading to the same regularization network of above in which the hinge loss is replaced by the *truncated square loss*.

Finally, *proximal* SVM, see [8] for reference, solving the same problem of smooth SVM in which the inequality constraints are replaced by equality constraints, lead to a regularization network of (3) with $c = 2$ with the *square loss* as loss function, that is a standard least squares regularization network.

5 Conclusions

In this paper we showed that several Support Vector methods are particular instances of functional minimization on a RKHS. Each algorithm can be derived from the general RN framework by choosing the appropriate loss function and exponent for the constant offset in the penalty term. In the case of one-class SVM it turns out that the underlying problem is truly a binary classification task for which no example from one of the two classes is available in training. From our analysis many theoretical results easily follow.

Consistency results are available for all the non standard SVM algorithms considered (see for example [14] and reference therein). Representation theorems for the form of the solution in the various cases can be applied. In particular, recent results ([5], [18]) express the solution in closed form exploiting a convexity assumption on the loss function. Results about existence and uniqueness can be found in [2] and [5] for all the non standard SV algorithms with the exception of L2-SVM (which however can be easily dealt with using results on the standard SVM).

References

[1] Bartlett, P. L. and Jordan, M. I. and McAuliffe, J. D.: Convexity, classification, and risk bound. Department of Statistics, U.C. Berkeley , Technical report, 638, 2003.

[2] Burges, C. and Crisp, D.: Uniqueness Theorems for Kernel Methods. to appear in Neurocomputing.

[3] Cristianini, N. and Shawe Taylor, J.: An Introduction to Support Vector Machines. Cambridge University Press, 2000, Cambridge, UK.

[4] Cucker, F. and Smale, S.: On the mathematical foundation of learning. Bull. A.M.S., 39, 1-49, 2002.

[5] De Vito, E. and Rosasco, L. and Caponnetto, A. and Piana, M. and Verri, A.: Some Properties of Regularized Kernel Methods. submitted, 2003.

[6] Devroye, L. and Györfi, L. and Lugosi G.: A Probabilistic Theory of Pattern Recognition. Springer, Applications of mathematics, 31, 1996, New York.

[7] Evgeniou, T. and Pontil M. and Poggio, T.: Regularization networks and support vector machines. Adv. Comp. Math., 13, 1-50, 2000.

[8] Fung, G. and Mangasarian, O. L.: Proximal Support Vector Machine Classifiers. 01-02, February, 2001, Data Mining Institute - University of Wisconsin - Madison.

[9] Mangasarian, O. L. and Musicant, D. R.: Data Discrimination via Nonlinear Generalized Support Vector Machines. Complementarity: Applications, Algorithms and Extensions, 233-251, Kluwer Academic Publishers, 2001.

[10] Lee, Y. J. and Mangasarian, O. L.: SSVM: A Smooth Support Vector Machine for Classification. Computational Optimization and Applications, 1, 20, 5-22, 2001.

[11] Lin, Y. and Lee, Y. and Wahba, G.: Support Vector Machines for Classification in Nonstandard Situations. Machine Learning, 46, 191-202, 2002.

[12] Rosasco,L. and De Vito, E. and Caponnetto, A. and Piana, M. and Verri, A.: Are Loss Functions all the Same?. Neural Computation, Vol. 16, Issue 6, 1063-1076 2004.

[13] Schölkopf, B. and J.C. Platt and J. Shawe-Taylor and A.J. Smola and R.C. Williamson: Estimating the support of a high-dimensional distribution. Neural Computation, Vol. 13, Issue 7, 1443-1471, 2001.

[14] Steinwart, I.: Consistency of support vector machines and other regularized kernel machines. submitted to IEEE Transactions on Information Theory, 2002.

[15] Suykens, J. A. K. and Van Gestel, T. and De Brabanter, J. and De Moor, B.: Least Squares Support Vector Machines. World Scientific, 2002.

[16] Tax, D.M.J. and Duin, R.P.W.: Data Domain Description using Support Vectors. Proceedings of European Symposium on Artificial Neural Networks '99, Brugge, 1999.

[17] Vapnik, V.: Statistical Learning Theory. Wiley, 1998, New York

[18] Zhang, T.: Convergence of Large Margin Separable Linear Classification. Advances in Neural Information Processing Systems 13, 357-363, MIT Press, 2001.