Dipartimento di Informatica e
Scienze dell'Informazione

Non standard Support Vector Machines and Regularization Networks

Andrea Caponnetto, Lorenzo Rosasco

Technical Report DISI-TR-04-03

DISI, Università di Genova
v. Dodecaneso 35, 16146 Genova, Italy http://www.disi.unige.it/

Abstract

Many recently proposed learning algorithms are clearly inspired by Suppport Vector Machines. Some of them were developed while trying to simplify the quadratic programming problem that has to be solved in training SVMs. Some others have been proposed to solve problems other than binary classification (for example one-class SVM for novelty detection).

Though indeed attractive, for most of the learning machine community the above algorithms lack of a clear theoretical motivation. In this context it seems that the connection to regularization networks is most promising both from a theoretical and a practical point of view and might be of great use to understand the mathematical properties of various SV algorithms. In this paper we contribute to fill the existing gap reviewing several SV algorithms from a regularization point of view.

Keywords: SVM, Reproducing kernel Hilbert spaces, Quadratic programming, Consistency, Representer theorem, Regularization networks.

1. Introduction

A binary patten classification problem amounts to determine a classificator able to distinguish between elements of two classes. In learning from examples we are given a training set of instances of the two classes and we have to determine a classificator capable of generalization, that is able to perfom well on new instances.

Support Vector Machines (SVM) algorithms proved to be extremely effective in a variety of different application fields and, as shown in Evgeniou et al. (2000), they 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). The latter proved to be a convenient formulation in order to study the mathematical properties of SVMs.

Recently several learning algorithms were inspired by SVMs. Some of them resulted from the effort to obtain algorithms easier to implement. In fact it is well known that training standard SVMs involves a QP dual problem which can be hard to solve. A possible way to overcome these difficulties is slightly modifying the dual problem formulation. The above approach led to a number of learning algorithms to which we refer as non standard Support Vector Machines (NS-SVM). Some other algorithms, as for instance one-class SVM for novelty detection, shared the same geometrical intuition of the original SVM.

Though indeed promising from a practical viewpoint the above algorithms lack a clear theoretical foundation. In this paper we show that all the above algorithms can be described in a general regularization framework and this automatically makes available many existing theoretical results. In particular through the formulation in terms of RN we can easily apply known existence and uniqueness results, describe the explicit form of the solution and dicuss generalization properties.

2. Learning from examples

background

We now briefly review the main concepts and notations of the binary patterns classification problem in a statistical learning framework (for details see Evgeniou et al. (2000), Vapnik (1988), Hastie et al. (2001), Devroye et al. (1996)).

2.1 Input and Output: the Sample Space

We consider an input space $X \subset \mathbb{R}^n$ and an output space $Y = \{-1, 1\}$. The standard assumption is requiring X and Y to be compact. The space $X \times Y$ is endowed with a probability measure $\rho(\mathbf{x}, y) = \rho(\mathbf{x})\rho(y|\mathbf{x})$, where $\rho(\mathbf{x})$ denotes the marginal probability measure on X and $\rho(y|\mathbf{x})$ the conditional probability measure of y given $\mathbf{x}$. We can think of the data space as some vectors $\mathbf{x}$ belonging to two classes labelled respectively with $y = 1$ and $y = -1$. In a learning problem the probability measure ρ is fixed but unknown and the data we are given are ℓ pairs of examples $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$, called training set, that we assume to be drawn i.i.d. with respect to ρ.

sec:bin_class

2.2 The Binary Classification

Roughly speaking, given the training set S the classification problem amounts to find a deterministic classification rule in such a way that the misclassification error on future examples is as small as possible. That is, we look for a function $f_S : X \to \mathbb{R}$ such that the expected misclassification error

$$R(f_S) = \int_{X \times Y} \theta(-y f_S(\mathbf{x})) d\rho(\mathbf{x}, y) = \Pr(f(\mathbf{x}) \neq y),$$

(here $\theta(\tau) = 1$ if $\tau > 0$ and $\theta(\tau) = 0$ otherwise) is as close as possible to the Bayes risk

$$R^* = \inf_{f \in \mathcal{F}} R(f),$$

with $\mathcal{F}$ the set of all measurable functions.

In other words the goal of the learning procedure is to ensure that the error rate of the solution converges in probability to the Bayes risk possibly with a fast rate of convergence. The following definition formalizes these ideas.

Definition 1 (Bayes Consistency) *We say that a learning algorithm achieves Bayes consistency if the following convergence in probability holds for every $\epsilon > 0$*

$$\lim_{\ell \to \infty} \operatorname*{Prob}_S \left(R(f_S) - R^* > \epsilon \right) = 0$$

where f_S is the solution returned by the learning algorithm given a training set S.

We do not give more details on this topic, we just conclude this Section noting that the rate of convergence in the above definition is crucial to obtain good generalization: the faster the convergence rate the fewer examples we need to obtain meaningful solutions.

3. Learning Algorithms and Regularization Networks

sec:RN

A learning algorithm is a map that, given the training set S, provides us with a classification rule f_S. While trying to actually design such an algorithm we soon realize that the framework presented in the previous Section is too general to deal with, the main problems being the followings:

1. trying to directly minimize the number of misclassification errors on the data leads to algorithms that are not computationally feasible

2. the space $\mathcal{F}$ is too large and we have to consider a smaller hypothesis space $\mathcal{H}$ (usually a RKHS with kernel K).

The first problem is usually solved considering a convex approximation of the misclassification loss $\theta(-yf(\mathbf{x}))$ and the problem arises to quantify how much we are actually changing the original problem. In Section 6.2 we discuss how to deal with this problem.

The class of algorithms that we consider throughout are the so called Regularization Networks (RN), which amount to solve the following minimization problem

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

where the first term measure how well the function f approximates the given data and the second term is the squared norm of f in the RKHS $\mathcal{H}$ which controls the complexity (smoothness) of the solution. The parameter λ is the regularization parameter that balances the tradeoff between the two terms.

The term $V(y, f(\mathbf{x}))$ can be interpreted as a measure of the penalty (loss) when classifying by $f(\mathbf{x})$ a new input $\mathbf{x}$ whose label is indeed y. Usually in the classification setting the loss function V depends on its arguments through the product $yf(\mathbf{x})$, for this reason in the following we will use indifferently the expressions $V(y, f(\mathbf{x}))$ and $V(yf(\mathbf{x}))$. However we want to stress that this particular form of the loss function implicitly models situations in which false negative ($y = +1$ and $f(\mathbf{x}) < 0$) and false positive ($y = -1$ and $f(\mathbf{x}) > 0$) errors are equally penalized. More general situations have been considered in the literature (see for example Lin et al. (2002)), in the general case an extra factor depending on y has to be added to the loss function

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

We will return on this last fact while considering one-class SVM. We can obtain different learning algorithms choosing different loss functions V (Evgeniou et al., 2000). Some choices we will 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 practice an unpenalized offset term is often added to the classifier in Prob. (1), yielding

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

In Section 3.2 we discuss in details how the introduction of the unpenalized offset can be interpreted in terms of RKHS.

The problem above can be generalized to include arbitrary penalization of the offset term, in particular let us consider the additive penalization b^c, with c a fixed parameter

$$\min_{f\in\mathcal{H}, b\in\mathbb{R}}\{\frac{1}{\ell}\sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i) + b) + \lambda(\|f\|^2_{\mathcal{H}} + b^c)\}. \tag{4}$$ genpenoffset

Porb. (3) is subsumed for $c = 0$. Moreover we will show in the following that both the cases $c = 1$ and $c = 2$ realize learning algorithms known in the literature.

3.1 Standard SVM & Regularized Least Squares Classification

We now briefly recall the minimization problem arising in standard SVM and Regularized Least Squares Classification (RLSC). The following minimization problem describes the SVM algorithm in its primal formulation

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

An alternative formulation equivalent to the previous one can be stated in terms of slack variables and margin

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

s.t. $\forall\, i$

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

Finally we recall that the RLSC amounts to consider the following minimization problem

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

sec:offset

3.2 Offset Term and RKHS

Before reviewing a number of different training approaches for SVMs from a regularization theoretical point of view, in this section we report two results which clarify the meaning of the offset b. When considering RN algorithms one of the following cases usually occurs

1. no offset term is considered,
2. an unpenalized offset term is considered,
3. a quadratic penalization of the offset is added.

In this section we discuss the connections between these situations. First, we cast the case in which an unpenalized bias term is considered in a RKHS framework. Then we show that case 3 is subsumed by case 1.

In order to state the first result of this Section we previously note the following facts. First, we recall that the space of constants functions $\mathcal{B}=\mathbb{R}$ is a RKHS with kernel $K_{\mathcal{B}}(\mathbf{s},\mathbf{x})=1$. Second, we note that when we consider the minimization problem with an offset term, we separately minimize on $f\in\mathcal{H}$ and $b\in\mathcal{B}=\mathbb{R}$, that is we are considering couples $(f,b)\in\mathcal{H}\times\mathbb{R}$. Third, since we are looking for a solution of the form $\bar{f}(\mathbf{x})+\bar{b}$ it is useful to consider the sum space

$$\mathcal{S}=\mathcal{H}+\mathcal{B}=\mathcal{H}+\mathbb{R}. \tag{5}$$

The hypothesis space $\mathcal{S}$ is again a RKHS and the associated kernel is simply the sum of the kernels of $\mathcal{H}$ and $\mathcal{B}$ that is $K+1$. We are now ready to state the following theorem which is a simple reformulation of Theorem 3 in De Vito et al. (2003b). See De Vito et al. (2003b) for details and proof.

teo2 **Theorem 2** *Let Q be the orthogonal projection on the subset of functions orthogonal to $\mathcal{B}$ w.r.t. the scalar product in $\mathcal{S}$, that is the following closed subspace of $\mathcal{S}$*

$$\mathcal{S}_0 = \{s \in \mathcal{S} \mid \langle s, g \rangle_{\mathcal{S}} = 0 \quad \forall g \in \mathcal{B}\}. \tag{6}$$

We have that the couple $(\bar{f}, \bar{b})$ is a solution of the problem

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

if and only if $\bar{f} = Q\bar{s}$ and $\bar{b} = \bar{s} - Q\bar{s}$, with $\bar{s}$ (that is $\bar{f} + \bar{b}$) a solution of the problem

$$\min_{s \in \mathcal{S}} \{\frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, s(\mathbf{x}_i)) + \lambda \|Qs\|_{\mathcal{S}}^2\}. \tag{8}$$

The above result shows that the case in which an unpenalized offset term appears in the minimization problem corresponds to the following situation: we are simply considering the minimization of a functional in the space $\mathcal{S}$ but the penalty term appearing in the functional is a seminorm. Moreover it is clear that if we are considering the unpenalized offset term the two kernels $K(\mathbf{s}, \mathbf{x})$ and $K(\mathbf{s}, \mathbf{x}) + 1$ are completely equivalent.

But what does it happen when we consider an offset term and we do penalize it? The next theorem formally answers to this question. Intuitively it shows that we are considering the minimization of a functional in the sum space $\mathcal{S}$ but this time the penalty term is the standard squared norm in the space. Before proving the theorem which formalizes the above statements, let us introduce some notation. Let $s \in \mathcal{S}$, then it is well known (see Aronszajn (1950) or Schwartz (1964)) that

$$\|s\|_{\mathcal{S}}^2 = \min_{\substack{(f,b) \in \mathcal{H} \times \mathbb{R} \\ \text{s.t. } f+b=s}} \{\|f\|_{\mathcal{H}}^2 + b^2\}. \tag{9}$$ norm

Moreover the minimum is attained by the couple (f_s, b_s) given by $f_s = Qs$ and $b_s = s - Qs$ (here Q is the projector defined in the text of Theorem 2). This last fact is a consequence of Lemma 6 in De Vito et al. (2003b) applied to the case of finite dimensional $\mathcal{B}$.

We are now ready to state the theorem

theo_b_squared **Theorem 3** *The couple $(\bar{f}, \bar{b})$ is a solution of the problem*

$$\min_{(f,b) \in \mathcal{H} \times \mathbb{R}} \{\frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, f(\mathbf{x}_i) + b) + \lambda(\|f\|_{\mathcal{H}}^2 + b^2)\} \tag{10}$$ proffset

if and only if $\bar{f} = Q\bar{s}$ and $\bar{b} = \bar{s} - Q\bar{s}$, with $\bar{s}$ (that is $\bar{f} + \bar{b}$) a solution of the problem

$$\min_{s \in \mathcal{S}} \{\frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, s(\mathbf{x}_i)) + \lambda \|s\|_{\mathcal{S}}^2\}. \tag{11}$$ prnoffset

Proof

Let us preliminarily notice that the minima achieved by the two functionals in the text are equal, in fact

$$\inf_{(f,b)\in\mathcal{H}\times\mathbb{R}}\{\frac{1}{\ell}\sum_{i=1}^{\ell}V(y_i,f(\mathbf{x}_i)+b)+\lambda(\|f\|_{\mathcal{H}}^2+b^2)\} \tag{12}$$

prnoffset

$$=\inf_{s\in\mathcal{S}}\inf_{\substack{(f,b)\in\mathcal{H}\times\mathbb{R}\\ \text{s.t. } f+b=s}}\{\frac{1}{\ell}\sum_{i=1}^{\ell}V(y_i,s(\mathbf{x}_i))+\lambda(\|f\|_{\mathcal{H}}^2+b^2)\} \tag{13}$$

$$=\inf_{s\in\mathcal{S}}\{\frac{1}{\ell}\sum_{i=1}^{\ell}V(y_i,s(\mathbf{x}_i))+\lambda\|s\|_{\mathcal{S}}^2\}, \tag{14}$$

where the last equality descends from Eq.(9).

Now assume that $\bar{s}$ is a solution of Prob.(11), then

$$\frac{1}{\ell}\sum_{i=1}^{\ell}V(y_i,\bar{s}(\mathbf{x}_i))+\lambda(\|f_{\bar{s}}\|_{\mathcal{H}}^2+(b_{\bar{s}})^2)=\frac{1}{\ell}\sum_{i=1}^{\ell}V(y_i,\bar{s}(\mathbf{x}_i))+\lambda\|\bar{s}\|_{\mathcal{S}}^2. \tag{15}$$

Recalling that we set $\bar{f}=Q\bar{s}$ and $\bar{b}=\bar{s}-Q\bar{s}$, it follows that the couple $(f_{\bar{s}},b_{\bar{s}})$ is a solution of Prob.(10) since it attains the common minimum. This proves one half of the theorem.

On the other hand assume that $(\bar{f},\bar{b})$ is a solution of Prob.(10), then setting $\bar{s}=\bar{f}+\bar{b}$ and recalling again Eq.(9), we get

$$\frac{1}{\ell}\sum_{i=1}^{\ell}V(y_i,\bar{s}(\mathbf{x}_i))+\lambda\|\bar{s}\|_{\mathcal{S}}^2\leq\frac{1}{\ell}\sum_{i=1}^{\ell}V(y_i,\bar{f}(\mathbf{x}_i)+\bar{b})+\lambda(\|\bar{f}\|_{\mathcal{H}}^2+\bar{b}^2). \tag{16}$$

But the r.h.s. of the inequality above is the common minimum of problems (11) and (10), then equality must hold. It follows both that $\bar{s}$ is a solution of Prob.(11) and that $(\bar{f},\bar{b})=(f_{\bar{s}},b_{\bar{s}})$. This concludes the proof.

■

4. Non Standard SVM revisited

In this section we review a number of SVM-like algorithms and show that each one is indeed equivalent to a particular regularization network algorithm. We stress that our analysis will not make use of the dual formulation of the algorithms.

4.1 Preliminary remarks

preliminary

We add a few remarks before starting our discussion.

- the linear case $K(\mathbf{x},\mathbf{s})=\mathbf{x}\cdot\mathbf{s}$ has received much attention in machine learning literature (it realizes for example the original linear SVM). However we will not treat separately linear and non-linear cases, in fact we can always assume the existence of a mapping $\mathbf{\Phi}$ from X to a (possibly infinite dimensional) *feature space*, such that $K(\mathbf{x},\mathbf{s})=\mathbf{\Phi}(\mathbf{x})\cdot\mathbf{\Phi}(\mathbf{s})$,

- the general primal formulation in terms of slack variables and feature map becomes

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

s.t. $\forall\ i$

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

- note that the relation between the parameter C in the standard slack variables formulation and the parameter λ in the regularization network formulation is simply $C = \frac{1}{2\lambda\ell}$,
- finally, the list of algorithms that we are going to review can be divided in two groups: the first composed of regularization networks not penalizing the offset term, the second group penalizing it.

4.2 Regularization Networks with unpenalized offset term

4.2.1 L2-SVM

We consider (see Cristianini and Shawe Taylor (2000) for reference) the problem

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

s.t. $\forall\ i$

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

that is we consider the squares of the slack variables and do not penalize the bias term b.

The above problem can be seen as the regularization network obtained considering the truncated square loss (see Section 3) and penalizing the estimators by the seminorm $\|Q\cdot\|_{\mathcal{S}}$ in the sum space $\mathcal{S}$ defined in Theorem 2. In fact we are simply considering the minimization problem

$$\min_{s \in \mathcal{H}} \{\frac{1}{\ell} \sum_{i=1}^{\ell} |1 - y_i s(\mathbf{x}_i)|_+^2 + \lambda \|Qs\|_{\mathcal{S}}^2\}.$$

We have seen in Section 3.2 that the above problem can be also written in the (more direct) form

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

4.2.2 Least Square SVM, LS-SVM

We consider (see Cristianini and Shawe Taylor (2000) and Suykens et al. (2002) for reference) the problem

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

s.t. $\forall\, i$

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

the inequality constraints are now replaced by equality constraints and the square of the slack variables is considered, moreover an unpenalized offset term is added. It is easy to see that the above problem corresponds to a regularization network with square loss, the sum space $\mathcal{S}$ as hypothesis space and again penalty term $\lambda\left\|Qs\right\|_{\mathcal{S}}^2$. It follows that in this case the minimization problem can be written as

$$\min_{s\in\mathcal{S}}\{\frac{1}{\ell}\sum_{i=1}^{\ell}(1-y_i s(\mathbf{x}_i))^2+\lambda\left\|Qs\right\|_{\mathcal{S}}^2\}.$$

As we have seen in Subsection 3.2 another equivalent formulation is the following

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

We now consider algorithms where the offset term is also penalized.

4.3 Regularization Networks penalizing the offset term

For all the following algorithms the bias term appears explicitly in the form of the solution but it is now penalized by the complexity term.

4.3.1 Modified SVM

We consider (see Mangasarian and Musicant (2001) for reference) the problem

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

s.t. $\forall\, i$

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

that is, we now penalize the bias term b. The above problem can be easily interpreted in the sum space $\mathcal{S}$ with kernel $K+1$. In fact by Theorem 3in Subsection 3.2, it is equivalent to the problem

$$\min_{s\in\mathcal{S}}\{\frac{1}{\ell}\sum_{i=1}^{\ell}|1-y_i s(\mathbf{x}_i)|_+ +\lambda\left\|s\right\|_{\mathcal{S}}^2\}.$$

As we mentioned before the above situation is formally equivalent to the case in which no offset term is considered.

4.3.2 Smooth SVM, SSVM

We consider (see Lee and Mangasarian (2001) for reference) the problem

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

s.t. $\forall\, i$

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

that is we consider the square of the slack variables and do penalize the bias term b.

The above problem can be seen as the regularization network obtained considering the truncated square loss and searching for a solution in the sum space $\mathcal{S}$. In fact we are simply considering the minimization problem

$$\min_{s\in\mathcal{S}}\{\frac{1}{\ell}\sum_{i=1}^{\ell}|1-y_i s(\mathbf{x}_i)|_+^2+\lambda\left\|s\right\|_{\mathcal{S}}^2\}.$$

4.3.3 Proximal SVM, PSVM

We consider (see Fung and Mangasarian (2001) for reference) the problem

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

s.t. $\forall\, i$

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

the inequality constraints are now replaced by equality constraints and the squares of the slack variables are considered, moreover the bias term b is penalized.

The above problem can be easily interpreted in the sum space $\mathcal{S}$. In fact we are simply considering the minimization problem

$$\min_{s\in\mathcal{S}}\{\frac{1}{\ell}\sum_{i=1}^{\ell}(1-y_i s(\mathbf{x}_i))^2+\lambda\left\|s\right\|_{\mathcal{S}}^2\}.$$

It should be apparent that the above problem is completely equivalent to the RLSC algorithm in the RKHS $\mathcal{S}$.

5. One-class Support Vector machines as a regularization network

oneclass

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 (Tax and Duin (1999), Schlkopf et al. (2001)). In this setting the component of the training set $(\mathbf{x}_i,y_i)_{i=1}^{\ell}$ labelled 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, in its simplest formulation, is looking for the smallest sphere enclosing the examples in the data space. Hence the training procedure amounts to the solution of the following constrained minimization problem with respect to the balls of center $\mathbf{a}$ and radius R in the input space

$$\min_{R,\xi_i}\{C\sum_{i=1}^{\ell_+}\xi_i + R^2\}, \tag{17}$$ oneclassprob1

conditioned to the existence of a vector $\mathbf{a} \in \mathbb{R}^n$ s.t. $\forall\ i \leq \ell_+$

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

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, that is the squared distance appearing in (17) is now replaced by the expression $(\mathbf{\Phi}(\mathbf{x}_i) - \mathbf{a}) \cdot (\mathbf{\Phi}(\mathbf{x}_i) - \mathbf{a})^T$.

It can be shown (see for example Cucker and Smale (2002)) 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 problem can be written as follows

$$\min_{R^2,\xi_i}\{C\sum_{i=1}^{\ell_+}\xi_i + R^2\}, \tag{18}$$ oneclassprob2

requiring that there exists a vector $f \in \mathcal{H}$ s.t. $\forall\ i \leq \ell_+$

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

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 complexity measure run over unconstrained reals. This is obtained simply by introducing the real variable ρ and formulating the modified problem as follows

$$\min_{\rho,\xi_i}\{C\sum_{i=1}^{\ell_+}\xi_i + \rho\}, \tag{19}$$ oneclassprob3

conditioned to the existence of a function $f \in \mathcal{H}$ s.t. $\forall\ i \leq \ell_+$

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

The two problems are indeed virtually equivalent, in fact it is straightforward to verify that whenever $C > \frac{1}{\ell_+}$ problems (18) and (20) have the same solutions. On the other hand if

$C < \frac{1}{\ell_+}$ both problems are trivial: any solution of (18) attains $R = 0$ while no solution of (20) exists.

Introducing the offset b and the fixed bias function $\mathcal{D}(\mathbf{x})$, and suitably rescaling the parameter C and the kernel K, the previous problem becomes

$$\min_{f,b,\xi_i} \{\widetilde{C} \sum_{i=1}^{\ell_+} \xi_i + \frac{1}{2}(\|f\|^2_{\widetilde{\mathcal{H}}} + b)\}, \tag{20}$$ oneclassprob3

s.t. $\forall\, i \leq \ell_+$

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

where we set

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

The couple (f, b) in (20) runs over $\widetilde{\mathcal{H}} \times \mathbb{R}$, with $\widetilde{\mathcal{H}}$ the RKHS of kernel $\widetilde{K}$.

Finally, by standard reasoning the problem can be rewritten in terms of loss function, penalty term and original two-class training set, in fact considering the loss function

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

which matches the general form in Eq.(2), we easily obtain

$$\min_{f \in \widetilde{\mathcal{H}}, b \in \mathbb{R}} \{\frac{1}{\ell} \sum_{i=1}^{\ell} V(y_i, \mathcal{D}(\mathbf{x}_i) + f(\mathbf{x}_i) + b) + \lambda(\|f\|^2_{\widetilde{\mathcal{H}}} + b)\}. \tag{21}$$ oneclassprob4

By definition the fixed bias function $\mathcal{D}(\mathbf{x})$ is null for translation invariant kernel functions (e.g. the gaussian kernel). In this case problem (21) fits the general regularization network form (4), with $c = 1$.

6. Properties of Regularization Networks

In this Section we summarize some recent results that allow to completely characterize the mathematical properties of most of the presented algorithms. We deal in particular with the following issues: existence, uniqueness and explicit form of the solution of problem (1) and (3), discussion of a general scheme to study the Bayes consistency of the RN algorithms.

6.1 On the Solution of Regularization Networks with Convex Loss

An exhaustive analysis of the mathematical properties of the solution of problems (1) and (3) has been recently proposed considering a very general context (De Vito et al., 2003b). In the following we report without proofs the main results in the aforementioned paper, specialized to the discrete classification problem and constant offset term.

6.1.1 Quantified Representer Theorem

We start by reviewing the result about the explicit form of the solution. The following is a slightly modified version of Corollary 12 in De Vito et al. (2003b) holding for convex loss function satisfying some extra very weak conditions (which are satisfied for all the loss functions considered in these notes).

repre **Theorem 4** *Let the loss function V be convex and $\mathcal{H}$ be the RKHS with kernel K. The following two conditions are equivalent.*

1. *The pair $(f^\lambda, b^\lambda) \in \mathcal{H} \times \mathbb{R}$ is a solution of*

$$\min_{f \in \mathcal{H}, b \in \mathbb{R}} \left(\frac{1}{\ell} \sum_i V(y_i(f(\mathbf{x}_i) + b)) + \lambda \|f\|_{\mathcal{H}}^2 \right)$$

2. *There are $\alpha_1, \ldots, \alpha_\ell \in \mathbb{R}$ such that*

$$f^\lambda = \sum_{i=1}^{\ell} y_i \alpha_i K_{\mathbf{x}_i} = \sum_{i=1}^{\ell} y_i \alpha_i (K_{\mathbf{x}_i} + 1).$$

where

$$\frac{-1}{2\lambda \ell y_i} V'_+(y_i, f^\lambda(\mathbf{x}_i) + b^\lambda) \leq \quad \alpha_i \quad \leq \frac{-1}{2\lambda \ell y_i} V'_-(y_i, f^\lambda(\mathbf{x}_i) + b^\lambda) \tag{22}$$ alfa

$$\sum_{i=1}^{\ell} y_i \alpha_i \quad = \quad 0 \tag{23}$$

The message of the above result can be summarized as follows: exploiting a convexity assumption on the loss function it is possible to deduce a quantified version of the well known representer theorem holding both in the the standard (with no offset) case and in the semiparametric (considering the offset term) case. The main difference with classic results on the representation of the solution is that this time the expression for the coefficients α_i appearing in the solution is given in closed form through the subgradient of the loss function (22).

6.1.2 Existence an Uniqueness

We again report without proof the main results in De Vito et al. (2003b). For similar results see also Steinwart (2002b) and Burges and Crisp (2002).

The first result shows that when the offset term is not considered (or when it is penalized) strict convexity and coerciveness of the penalty term trivially imply both existence and uniqueness of the solution.

Proposition 5 *Given $\lambda > 0$, there exists a unique solution of the problem*

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

The following proposition discuss existence when an unpenalized constant offset term is considered.

existence_c

Proposition 6 *Assume that the following conditions hold*

1. $\lim_{w\to-\infty} V(1,w) = +\infty$ *and* $\lim_{w\to+\infty} V(-1,w) = +\infty$
2. *there is* $C > 0$ *such that* $\sqrt{K(\mathbf{x},\mathbf{x})} \leq C$ *for all* $\mathbf{x} \in \operatorname{supp}\nu$
3. $\exists\ \mathbf{x}_i, \mathbf{x}_j$ *s.t* $y_i = 1$ *and* $y_j = -1$.

Then there is at least one solution of the problem

$$\min_{f\in\mathcal{H}, b\in\mathbb{R}} \left(\frac{1}{\ell} \sum_i V(y_i, f(\mathbf{x}_i) + b) + \lambda \|f\|_{\mathcal{H}}^2 \right)$$

We note that condition *1* holds for all the loss functions we considered, while condition *2* is always satisfied when the input space X is compact. Finally, condition *3* tells us that existence is ensured whenever we have at least one example for each class.

We now deal with uniqueness of the solution. It is easy to see that if the loss function is strictly convex uniqueness is always ensured (see Proposition 10 in De Vito et al. (2003b)). A careful analysis is required if the considered loss is just convex. Next proposition discuss the case in which the hinge loss is considered.

Proposition 7 *Let* f^λ, b^λ *be a solution of*

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

and define

$$I_+ = \{i \,|\, y_i = 1,\ (f^\lambda + b^\lambda) < 1\} \qquad I_- = \{i \,|\, y_i = -1,\ (f^\lambda + b^\lambda) > -1\}$$
$$B_+ = \{i \,|\, y_i = 1,\ (f^\lambda + b^\lambda) = 1\} \qquad B_- = \{i \,|\, y_i = -1,\ (f^\lambda + b^\lambda) = -1\}.$$

The solution is unique if and only if

$$\#I_+ \neq \#I_- + \#B_- \tag{24}$$ pa1

and

$$\#I_- \neq \#I_+ + \#B_+, \tag{25}$$ pa2

where # denotes set cardinality.

If the solution is not unique, the solution family is parameterized as $s^\lambda + b$, where b runs on a closed, not necessarily bounded interval. However, if there is at least one example for each class, b lies in the bounded interval $[b_-, b_+]$ and one can easily show that

1. for the solution with $b = b_-$, only inequality (24) holds;
2. for the solution with $b = b_+$, only inequality (25) holds;
3. for the solution with $b_- < b < b_+$, neither (24) nor (25) hold, from which it follows that $\#I_+ = \#I_-$ and $\#B_+ = \#B_- = 0$.

sec:bayes_cons

6.2 Bayes Consistency of Regularization Networks

We finally consider the problem of Bayes consistency of regularization networks. For most of the material in this section we refer to Bartlett et al. (2003a) and Bartlett et al. (2003b). We have seen that one of the keys to design feasible algorithms is considering convex approximations $V(yf(\mathbf{x}))$ to the (computationally intractable) misclassification loss. In doing so we typically no longer control the deviation of the solution's misclassification error[1] from the Bayes risk

$$R(f_S) - R^*, \tag{26}$$ error_risk

we instead deal with the expected risk

$$I[f] = \int_{X \times Y} V(yf(\mathbf{x}))d\rho(\mathbf{x}, y),$$

and we control the deviation of the expected risk of the solution f_S from the minimum attainable expected risk, that is

$$I[f_S] - \inf_{f \in \mathcal{F}} I[f], \tag{27}$$ expected_risk

where $\mathcal{F}$ is again the space of all measurable functions. Then the task of proving Bayes consistency (see Section 2.2) is twofold. First, we need a method to relate (26) and (27). Second we have to ensure that (27) converges to zero in probability typically solving an estimation/approximation problem. The following result from Bartlett et al. (2003a) and Bartlett et al. (2003b) provide us with the comparison results we are looking for.

The text of the following Theorem refers to the function ψ, it results from a functional transform of the loss function V. However the details regarding the precise definition of ψ and the proofs of the results below are outside of the scope of these notes. It suffices to note that as a consequence of the properties of the loss function (in particular convexity), it follows that for any sequence $\theta_i \in [0, 1]$

$$\psi(\theta_i) \to 0 \text{ if and only if } \theta_i \to 0. \tag{28}$$ psiprop

In particular ψ can be explicitly computed for all the loss functions considered in Section 3

- hinge: $\psi(\theta) = \frac{\theta}{2}$,
- square: $\psi(\theta) = \theta^2$,
- truncated square: $\psi(\theta) = \theta^2$.

We are now ready to state the promised Theorem.

Theorem 8 *For any nonnegative loss function $V(yf(\mathbf{x}))$, measurable function $f \in \mathbb{R}^X$ and probability measure ρ on $X \times Y$*

$$\psi(R(f_S) - R^*) \leq I[f_S] - \inf_{f \in \mathcal{F}} I[f].$$

1. Recall that the misclassification error is simply the expected risk with respect to the misclassification loss.

The previous Theorem together with property (28) allow us to conclude that for every sequence of measurable functions $f_i : X \to \mathbb{R}$ and probability measure ρ on $X \times Y$

$$I[f_i] \to \inf_{f \in \mathcal{F}} I[f] \quad \text{implies} \quad R(f_i) \to R^*.$$

Given the above comparison result, the next step amounts to controlling the deviation in (27). This is usually accomplished studying an approximation and a sampling problem. We give some details specializing the discussion to regularization networks. For sake of simplicity we do not take into account the bias term. We define f_S^λ as the minimizer of (1) and f^λ as the solution of the problem

$$\min_{f \in \mathcal{H}} \{ \int_{X \times Y} V(yf(\mathbf{x})) d\rho(\mathbf{x}, y) + \lambda \|f\|_{\mathcal{H}}^2 \}.$$

Roughly speaking we can think of f^λ as the solution that the RN algorithm would return assuming to have enough data to completely characterize the problem. We can now split the deviation in (27) as follows

$$I[f_S^\lambda] - \inf_{f \in \mathcal{F}} I[f] \leq \underbrace{I[f_S^\lambda] - I[f^\lambda]}_{\text{sample error}} + \underbrace{I[f^\lambda] - \inf_{f \in \mathcal{H}} I[f]}_{\text{approximation error}} + \underbrace{\inf_{f \in \mathcal{H}} I[f] - \inf_{f \in \mathcal{F}} I[f]}_{\text{irreducible error}}. \qquad (29)$$

gen_error

The above terms can be interpreted in the following way:

1. sample error: the error component due to finite sampling,
2. approximation error: the error component due to the chosen regularization level, coded by the value for the parameter λ,
3. irreducible error: the error component due to the limited approximation capability of the RKHS $\mathcal{H}$ in $\mathcal{F}$.

Results regarding how to measure the above errors abound in the literature and cover all the algorithms that we are going to consider. Some recent results can be found in Cucker and Smale (2002), Steinwart (2002a), De Vito et al. (2003a).

conc

7. Remarks and Conclusions

In this paper we reviewed several SV algorithms as particular instances of functional minimization on a RKHS. We showed that each algorithm can be recovered in the general RN framework by choosing appropriate loss function and penalty term. From this formulation many theoretical results easily follow.

Consistency results are available for all the NS-SVM algorithms considered (see for example Steinwart (2002a) and reference therein). A reasonable question would be to ask which algorithm provides best performances in the classification task, but unfortunately the theory shows that there cannot be a general answer to this question (see Devroye et al. (1996)). Moreover most of the theoretical results concerning consistency do not shed light on the role -if any at all- played by the underling loss function (some results can be found in Rosasco et al. (2003), Bartlett

et al. (2003a), Zhang (2001)) and most important the probabilistic bounds at the basis of the consistency results are often too loose to be useful in practice.
Representation theorems for the solution of the various methods can be applied. In particular recent results (De Vito et al. (2003b), Zhang (2001)) express the solution in closed form exploting a convexity assumption on the loss function.
Results about existence and uniqueness can be found in Burges and Crisp (2002) and De Vito et al. (2003b) for all the NS-SVM algorithms except for L2-SVM though it should be easily treated using results on the standard SVM.

It would be interesting to extend our review to algorithms using more general penalty terms (as for instance L-1 norm), in such cases a careful mathematical analysis is required since we are no longer working in an hilbertian setting (see von Luxburg and Bousquet (2004) and Micchelli and M. Pontil (2003)).
Other interesting topics could be the extension of known results about consistency and generalization to one-class SVM and to the case of asymmetric loss function.

References

aron N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 686:337–404, 1950.

bjm-ccrb-03 Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, U.C. Berkeley, 2003a.

bjm-d-03 Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Discussion of boosting papers. *The Annals of Statistics*, 2003b. To appear.

burges1 C. Burges and D. Crisp. Uniqueness theorems for kernel methods. Technical Report MSR-TR-2002-11, Microsoft Research, 2002.

cri-sh N. Cristianini and J. Shawe Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

smale F. Cucker and S. Smale. On the mathematical foundation of learning. *Bull. A.M.S.*, 39:1–49, 2002.

model E. De Vito, A. Caponnetto, L. Rosasco, M. Piana, and A. Verri. Model selection for regularized least squares. Technical report, DISI, Dipartimento di Informatica e Scienze dell'Informazione, Genova, 2003a.

representer E. De Vito, L. Rosasco, A. Caponnetto, M. Piana, and A. Verri. Some properties of regularized kernel methods. *submitted*, 2003b.

DevGyoLug96 L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

epp00 T. Evgeniou, Pontil M., and T. Poggio. Regularization networks and support vector machines. *Adv. Comp. Math.*, 13:1–50, 2000.

funman01 G. Fung and O. L. Mangasarian. Proximal support vector machine classifiers. Technical Report 01-02, Data Mining Institute - University of Wisconsin - Madison, February 2001.

htf T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

leeman01 Y. J. Lee and O. L. Mangasarian. Ssvm: A smooth support vector machine for classification. *Computational Optimization and Applications*, 20(1):5–22, october 2001.

lilewa02 Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. *Machine Learning*, 46:191–202, 2002.

manmus01b O. L. Mangasarian and D. R. Musicant. Data discrimination via nonlinear generalized support vector machines. In M. C. Ferris, O. L. Mangasarian, and J.S. Pang, editors, *Complementarity: Applications, Algorithms and Extensions*, pages 233–251. Kluwer Academic Publishers, 2001.

micpon04 C. A. Micchelli and M M. Pontil. A function representation for learning in banach spaces. Technical Report Research Note RN/04/03, Dept of Computer Science, UCL, 2003.

loss L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *to appear in Neural Computation*, 2003.

schpla01 B. Schlkopf, 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.

schwartz L. Schwartz. Sous-espaces hilbertiens d'espaces vectorials topologiques and noyaux associès. *Journal d'Analyse Mathematique*, 13:115–256, 1964.

ingocons02 I. Steinwart. Consistency of support vector machines and other regularized kernel machines. *submitted to IEEE Transactions on Information Theory*, 2002a.

ingo03 I. Steinwart. Sparseness of support vector machines. *submitted to IEEE Transactions on Information Theory*, 2002b.

suyges02 J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, 2002.

taxdui99 D.M.J. Tax and R.P.W. Duin. Data domain description using support vectors. In *Proceedings of European Symposium on Artificial Neural Networks '99, Brugge*, 1999.

vapnik V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1988.

luxbou04 U. von Luxburg and O. Bousquet. Distance-based classification with lipschitz functions. 2004. Submitted.

zhang T. Zhang. Convergence of large margin separable linear classification. In T.G. Len, T.K. Dietterich and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 357–363. MIT Press, 2001.