ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Least 1-Norm SVMs: a New SVM Variant between Standard and LS-SVMs

Jorge López and José R. Dorronsoro *

Universidad Autónoma de Madrid
Departamento de Ingeniería Informática and
Instituto de Ingeniería del Conocimiento
C/ Francisco Tomás y Valiente 11, 28049 Madrid, Spain

Abstract. Least Squares Support Vector Machines (LS-SVMs) were proposed by replacing the inequality constraints inherent to L1-SVMs with equality constraints. So far this idea has only been suggested for a least squares (L2) loss. We describe how this can also be done for the sum-of-slacks (L1) loss, yielding a new classifier (Least 1-Norm SVMs) which gives similar models in terms of complexity and accuracy and that may also be more robust than LS-SVMs with respect to outliers.

1 Introduction

Assuming a binary classification context, we have a sample of N preclassified patterns $\{X_i, y_i\}, i = 1, \ldots, N$, where the outputs $y_i \in \{+1, -1\}$. If we further assume linear inseparability and consider slack variables to allow for misclassifications, the primal of an LS-SVM [1] is

$$\min_{W,b,\xi} \ \frac{1}{2}\|W\|^2 + \frac{C}{2}\sum_i \xi_i^2 \quad s.t. \quad y_i\left(W \cdot \Phi\left(X_i\right) + b\right) = 1 - \xi_i \ \ \forall i, \tag{1}$$

where $\cdot$ denotes inner product, and $\Phi\left(X_i\right)$ is the image of X_i in the feature space with feature map $\Phi\left(\cdot\right)$. The corresponding dual is

$$\min_{\alpha} \ \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \tilde{K}_{ij} - \sum_i \alpha_i \quad s.t. \quad \sum_i \alpha_i y_i = 0, \tag{2}$$

with the modified kernel $\tilde{K}_{ij} = k\left(X_i, X_j\right) + \delta_{ij}/C$, δ_{ij} standing for Kronecker's delta symbol and $k\left(X_i, X_j\right) = \Phi\left(X_i\right) \cdot \Phi\left(X_j\right)$ the original kernel.

LS-SVMs were originally derived in [1] from the so-called L1-SVMs [2], whose primal changes (1) in three aspects: 1) the objective function uses the L1 loss $C\sum_i \xi_i$ instead of the L2 loss, 2) the equality constraints become inequality ones, and 3) there is the additional requirement that $\xi_i \geq 0$. In turn, L2-SVMs, also described in [2], lie somewhere in between, since their primal is identical to (1), but with the equality constraints still transformed into inequality ones.

To our knowledge, there is no current classifier that combines equality constraints with the L1 loss. It is desirable to fill this gap mainly because of two

*With partial support of Spain's TIN 2007–66862 project and Cátedra IIC en Modelado y Predicción. The first author is kindly supported by FPU-MICINN grant reference AP2007–00142.

	Squared Slacks	Slacks
Inequality constraints	L2-SVMs	L1-SVMs
Equality constraints	LS-SVMs	?

Table 1: Types of SVMs according to how slacks and constraints are treated.

facts: 1) in practice L1-SVMs and the L1 loss have become the standard, 2) the influence of a given pattern (i.e. the value of its coefficient α_i) in the model is not bounded when using the L2 loss, so L2 and LS-SVMs are more sensitive to outliers than L1-SVMs.

The central idea of this work is to simplify L1-SVMs similarly to LS-SVMs, but keeping the L1 loss, giving rise to the so-called Least 1-Norm SVMs, which fill the gap above and are expected to preserve the robustness to outliers. The rest of the paper is organized as follows: in Section 2 we give the primal and dual of Least 1-Norm SVMs and discuss briefly their KKT optimality conditions. Section 3 explains how the popular SMO algorithm can be adapted to solve the Least 1-Norm dual. Section 4 reports some experiments that illustrate how they can be more robust to outliers than LS-SVMs while being as accurate as them, and discusses the varied convergence speeds observed. Finally, Section 5 gives pointers to future possible extensions.

2 Least 1-Norm SVMs

In order to simplify the L1-SVM primal, one may think that it suffices to force equality constraints $y_i\left(W\cdot\Phi\left(X_i\right)+b\right)=1-\xi_i$, while keeping the inherent requirement $\xi_i\geq 0$. However, this is not correct because it implies that slacks are only allowed in one direction, something which is obviously not convenient. Therefore, we propose to remove the constraints $\xi_i\geq 0$ and minimize the 1-Norm of the slack vector, which gives the Least 1-Norm SVM primal

$$\min_{W,b,\xi}\ \frac{1}{2}\|W\|^2+C\sum_i|\xi_i| \quad s.t.\quad y_i\left(W\cdot\Phi\left(X_i\right)+b\right)=1-\xi_i\ \ \forall i. \tag{3}$$

Now we use the cast of 1-Norm problems as Linear Programming problems [3, p. 294]: minimizing (3) can be reformulated as

$$\min_{W,b,t}\ \frac{1}{2}\|W\|^2+C\sum_i t_i \quad s.t.\quad -t_i\leq 1-y_i\left(W\cdot\Phi\left(X_i\right)+b\right)\leq t_i\ \ \forall i, \tag{4}$$

Note that (4) transforms the desired equalities of (3) into inequalities, but otherwise the objective function is not differentiable. Using standard Lagrangian theory with (4) and denoting β_i (γ_i) as the multipliers associated with $-t_i$ ($+t_i$) we obtain the following dual, where $\alpha_i=\gamma_i-\beta_i$:

$$\min_{\alpha}\frac{1}{2}\sum_i\sum_j\alpha_i\alpha_j y_i y_j K_{ij}-\sum_i\alpha_i \quad s.t.\quad \sum_i\alpha_i y_i=0,\ -C\leq\alpha_i\leq C\ \ \forall i, \tag{5}$$

which happens to be identical to the L1-SVM dual but with the lower bound $-C$ instead of 0, so that negative values are allowed, as in LS-SVMs.

Since all the formulations above are convex with affine constraints, the KKT optimality conditions are necessary and sufficient for optimality [3]. The KKT conditions for (4) are analogous to the well-known ones for L1-SVMs, substituting just the lower bound $-C$ for 0, which yields:

$$\begin{aligned} y_i \left(W \cdot \Phi\left(X_i\right) + b\right) &= 1 \quad \forall i \mid -C < \alpha_i < C, \\ y_i \left(W \cdot \Phi\left(X_i\right) + b\right) &\leq 1 \quad \forall i \mid \alpha_i = C, \\ y_i \left(W \cdot \Phi\left(X_i\right) + b\right) &\geq 1 \quad \forall i \mid \alpha_i = -C, \end{aligned} \tag{6}$$

together with the dual constraints $W = \sum_i \alpha_i y_i \Phi\left(X_i\right)$ and $\sum_i \alpha_i y_i = 0$. These are common to LS-SVMs, whose only primal KKT condition [1] is

$$y_i \left(W \cdot \Phi\left(X_i\right) + b\right) = 1 - \alpha_i / C \ \ \forall i, \tag{7}$$

which shows why LS-SVMs are very sensible to outliers: outliers are characterized by a large $|\xi_i|$, which in view of (7) and (1) implies a large $|\alpha_i|$. On the other hand, in Least 1-Norm SVMs this influence is limited because $|\alpha_i| \leq C$.

It also shows another drawback of LS-SVMs: they are not sparse because $\alpha_i = C\xi_i \ \forall i$, so a pattern takes part in the model whenever $\xi_i \neq 0$, which is almost certain to happen. Observe that this is also the case for Least 1-Norm SVMs, since $\alpha_i = 0$ implies $y_i \left(W \cdot \Phi\left(X_i\right) + b\right)$ is exactly 1, so they are not likely to be sparse either. L1-SVMs are indeed sparse because, instead of $-C$, patterns with $y_i \left(W \cdot \Phi\left(X_i\right) + b\right) > 1$ are assigned $\alpha_i = 0$.

3 Least 1-Norm SMO

We will adapt SMO for Least 1-Norm SVMs basing on a maximum gain viewpoint (for more details see [4]). In general, SMO performs updates of the form $W' = W + \delta_L y_L X_L + \delta_U y_U X_U$. The constraint $\sum \alpha_i y_i = 0$ implies $\delta_U y_U = -\delta_L y_L$ and the updates become $W' = W + \delta y_L \left(X_L - X_U\right)$, where we write $\delta = \delta_L$ and, hence, $\delta_U = -y_U y_L \delta$. As a consequence, the multiplier updates are $\alpha'_L = \alpha_L + \delta$, $\alpha'_U = \alpha_U - y_U y_L \delta$ and $\alpha'_j = \alpha_j$ for other j. Therefore, denoting the dual function in (5) as $D\left(\alpha\right)$, $D\left(\alpha'\right)$ can be written as

$$D\left(\alpha'\right) = D\left(\alpha\right) - \frac{\left(\Delta_{U,L}\right)^2}{\left\|Z_{L,U}\right\|^2},$$

where we write $\Delta_{U,L} = W \cdot \left(X_U - X_L\right) - \left(y_U - y_L\right)$ and $Z_{L,U} = X_L - X_U$. Ignoring the denominator, we can approximately maximize the gain in $D\left(\alpha'\right)$ by choosing $L = \arg\min_j \left\{W \cdot X_j - y_j\right\}$ and $U = \arg\max_j \left\{W \cdot X_j - y_j\right\}$, so that the violation extent $\Delta_{U,L}$ is largest. Writing $\Delta = \Delta_{U,L}$ and $\lambda' = \Delta / \left\|Z_{U,L}\right\|^2$, we then have $\Delta > 0$, $\lambda' > 0$, $\delta = y_L \lambda'$ and the α updates become $\alpha'_L = \alpha_L + y_L \lambda'$, $\alpha'_U = \alpha_U - y_U \lambda'$. Thus, α'_L or α'_U will decrease if $y_L = -1$ or $y_U = 1$, which requires the corresponding α_L and α_U to be greater than $-C$. In turn, they will

increase if $y_L = 1$ or $y_U = -1$, which requires the corresponding α_L and α_U to be less than C. Hence, we must replace the previous L, U choices with

$$L = \arg\min_j \{W \cdot X_j - y_j : j \in \mathcal{I}_L\}, \; U = \arg\max_j \{W \cdot X_j - y_j : j \in \mathcal{I}_U\}, \quad (8)$$

where we use the notations $\mathcal{I}_U = \{i : (y_i = 1, \alpha_i > -C) \vee (y_i = -1, \alpha_i < C)\}$ and $\mathcal{I}_L = \{i : (y_i = 1, \alpha_i < C) \vee (y_i = -1, \alpha_i > -C)\}$. Moreover, to make sure that α'_L and α'_U remain then in the interval $[-C, C]$, we may have to clip λ' with

$$\lambda' = \min\{\lambda', C - y_L\alpha_L, C + y_U\alpha_U\}. \quad (9)$$

4 Numerical Experiments

In this section we will illustrate empirically how the Least 1-Norm SVM may be more robust to outliers than its LS-SVM counterpart, as well as its good generalization properties. The training algorithm is SMO; the Least 1-Norm variant explained above and the LS-SVM version in [5]. The stopping criterion is final KKT violation, specifically when it is less than $\epsilon = 10^{-3}$. For LS-SVMs this means

$$\max_i \left\{\tilde{W} \cdot \tilde{\Phi}(X_i) - y_i\right\} - \min_i \left\{\tilde{W} \cdot \tilde{\Phi}(X_i) - y_i\right\} \leq \epsilon, \quad (10)$$

where the tilde indicates that we use the modified kernel $\tilde{k}$ as in (2). For Least 1-Norm SVMs, it means

$$\max_{\mathcal{I}_U} \{W \cdot \Phi(X_i) - y_i\} - \min_{\mathcal{I}_L} \{W \cdot \Phi(X_i) - y_i\} \leq \epsilon. \quad (11)$$

The derivation of these KKT based criteria is given in [6] for LS-SVMs and [7] for L1-SVMs. Firstly, to show generalization we take 4 datasets from [8] with 100 training–test splits each. We compare the performance of Least 1-Norm and LS-SVMs. We use the RBF kernel $k(X_i, X_j) = \exp\left(-\|X_i - X_j\|^2/\sigma\right)$. The values for the hyperparameters C and σ are sought with a grid in the logarithmic range $[0, 2]$ for C and $[0, 4]$ for σ. Each point of the grid is evaluated with a 10-times-10-fold cross-validation over the whole dataset. We report in Table 2 the accuracy and number of support vectors obtained in the final models, as well as the number of iterations needed by the corresponding SMO version to stop.

	LS			Least 1-Norm		
	% err.	#SV	#It.	% err.	#SV	#It.
Titanic	22.4±1.0	150.0±0.0	390.1±14.7	22.4±1.0	71.4±9.8	53.7±6.1
Heart	15.6±3.2	169.9±0.3	261.3±7.8	15.6±3.5	169.9±0.3	120.1±14.7
Cancer	25.7±4.5	199.8±0.5	424.6±7.9	25.9±4.5	195.9±2.3	3510.7±701.9
German	23.3±2.1	699.4±0.8	3329.8±38.0	23.3±2.2	699.9±0.3	16167.4±1419.3

Table 2: Average accuracies, number of support vectors and number of iterations obtained by a Least 1-Norm SVM and an LS-SVM.

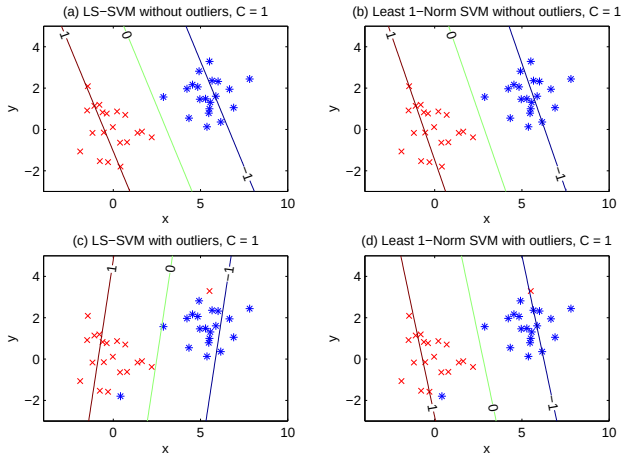

Fig. 1: Contours of function $W \cdot \Phi(X) + b$ for a toy problem trained with an LS-SVM (left) and a Least 1-Norm SVM (right). Top: original problem. Bottom: modified problem with one outlier for each class.

It can be seen how the accuracies obtained are similar for both kinds of SVM and also similar to the ones reported in [8] for an L1-SVM. Regarding the number of support vectors, as expected, none is sparse, except the Least 1-Norm SVM for dataset *titanic*, which we think is due to the existence of identical points with different tags. Finally, concerning the number of iterations, it is somewhat puzzling, sometimes the LS-SVM is remarkably faster and sometimes the Least 1-Norm SVM is. This of course depends on the hyperparameters chosen, but it is not clear what is the exact influence of them. Care must also be taken since (10) and (11), though formally similar, may require quite different number of iterations since the W vectors are different. Further study is clearly needed to better characterize what the convergence speed will be for each case.

Secondly, to show robustness we use the toy bidimensional problem depicted in 1, where 20 patterns belong to each class. The positive class' patterns are drawn from a normal distribution with mean $(0, 0)$, whereas the negative class has a mean of $(5, 2)$. In both cases the covariance matrix is the unit one. In the top part of the figure we train an LS-SVM (a) and a Least 1-Norm SVM (b) with this training set, which is linearly separable, with $C = 1$ and no specific kernel (just the inner product). Note that the final hyperplanes are very similar and the "support"hyperplanes traverse their corresponding cloud of points.

In the bottom part of the figure, we introduce two outliers by switching the class labels of two points, so that the classes are no longer linearly separable, training again the LS-SVM (c) and the Least 1-Norm SVM (d). Observe that the final LS-SVM hyperplane has remarkably changed its orientation because of the outliers' influence, whereas the Least 1-Norm one changes quite less because

their influence is limited.

5 Conclusions and further work

In this work we have presented Least 1-Norm SVMs, a new SVM classifier. As LS-SVMs did with L1-SVMs, they are derived by substituting inequality for equality constraints in the primal. The arising dual is almost identical to the L1 one, with box constraints $[-C, C]$ in lieu of $[0, C]$. This implies that the outliers' influence is also limited, but sparsity is lost because now the points for which $y_i W \cdot \Phi(X_i) > 1$ are assigned an $\alpha_i = -C$ instead of being zero. We have also seen how it can be trained with an adaptation of the well-known SMO algorithm, giving models with similar test accuracies. Which particular SVM variant converges faster seems to be problem and parameter dependent.

As a possible future extension, the training phase for Least 1-Norm SVMs can be accelerated by making use of the 2nd order variant of the SMO algorithm as was done for L1-SVMs in [7]. This method has been shown to not always accelerate LS-SVM training [6]. As mentioned above, the convergence properties of SMO for Least 1-Norm SVMs will be further studied.

References

[1] J. A. K. Suykens and J. Vandewalle. Least Squares Support Vector Machine Classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

[2] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[3] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[4] J. López, Á. Barbero, and J. R. Dorronsoro. On the Equivalence of the SMO and MDM Algorithms for SVM Training. In *Lecture Notes in Computer Science: Machine Learning and Knowledge Discovery in Databases*, volume 5211, pages 288–300. Springer, 2008.

[5] S. S. Keerthi and S. K. Shevade. SMO Algorithm for Least-Squares SVM Formulations. *Neural Computation*, 15(2):487–507, 2003.

[6] J. López and J. A. K. Suykens. First and Second Order SMO Algorithms for Large Scale LS-SVM training. Technical Report 09-179, Katholieke Universiteit Leuven, 2009.

[7] R. E. Fan, P. H. Chen, and C. J. Lin. Working Set Selection using Second Order Information for Training Support Vector Machines. *Journal of Machine Learning Research*, 6:1889–1918, 2005.

[8] G. Rätsch. *Benchmark Repository*, 2000. Datasets available at `http://ida.first.fhg.de/projects/bench/benchmarks.htm`.