Robust cross-validation score function for non-linear function estimation

J. De Brabanter, K. Pelckmans, J.A.K. Suykens, J. Vandewalle

K.U.Leuven ESAT-SCD/SISTA, Kasteelpark Arenberg 10, B-3001 Leuven Belgium
{jos.debrabanter,johan.suykens}@esat.kuleuven.ac.be

Abstract. In this paper a new method for tuning regularisation parameters or other hyperparameters of a learning process (non-linear function estimation) is proposed, called robust cross-validation score function ($\mathrm{CV}^{Robust}_{S-fold}$). $\mathrm{CV}^{Robust}_{S-fold}$ is effective for dealing with outliers and non-Gaussian noise distributions on the data. Illustrative simulation results are given to demonstrate that the $\mathrm{CV}^{Robust}_{S-fold}$ method outperforms other cross-validation methods.
Keywords. Weighted LS-SVM, Robust Cross-Validation Score function, Influence functions, Breakdown point, M-estimators and L-estimators

1 Introduction

Most efficient learning algorithms in neural networks, support vector machines and kernel based methods require the tuning of some extra learning parameters, or *hyperparameters*, denoted here by θ. For practical use, it is often preferable to have a data-driven method to select θ. For this selection process, many data-driven procedures have been discussed in the literature. Commonly used are those based on the cross-validation criterion of Stone [5] and the generalized cross-validation criterion of Craven and Wahba [1]. In recent years, results on L_2, L_1 cross-validation statistical properties have become available [9]. However, the condition $E\left[e_k^2\right] < \infty$ (respectively, $E\left[|e_k|\right] < \infty$) is necessary for establishing weak and strong consistency for L_2 (respectively, L_1) cross-validated estimators. On the other hand, when there are outliers in the y observations (or if the distribution of the random errors has a heavy tail so that $E\left[|e_k|\right] = \infty$), then it becomes very difficult to obtain good asymptotic results for the $L_2\,(L_1)$ cross-validation criterion. In order to overcome such problems, a robust cross-validation score function is proposed in this paper.

This paper is organized as follows. In section 2 the classical S-fold crossvalidation score function is analysed. In section 3 we construct a robust crossvalidation score function based on the trimmed mean. In section 4 the repeated robust S-fold crossvalidation score function is described. In section 5 we give an illustrative example.

2 Analysis of the *S*-fold cross-validation score function

The cross-validation procedure can basically be split up into two main parts: (a) constructing and computing the cross-validation score function, and (b) finding

the hyperparameters as: $\theta^* = \text{argmin}_\theta \left[CV_{S-fold}(\theta)\right]$. In this paper we focus on (a).

Let $\{z_k = (x_k, y_k)\}_{k=1}^N$ be an independent identically distributed (i.i.d.) random sample from some population with distribution function $F(z)$. Let $F_N(z)$ be the empirical estimate of $F(z)$. Our goal is to estimate a quantity of the form

$$T_N = \int L\left(z, F_N(z)\right) dF(z), \tag{1}$$

with $L(\cdot)$ the loss function (e.g. the L_2 or L_1 norm) and where $E\left[T_N\right]$ could be estimated by cross-validation. We begin by splitting the data randomly into S disjoint sets of nearly equal size. Let the size of the s-th group be m_s and assume that $\lfloor N/S \rfloor \leq m_s \leq \lfloor N/S \rfloor + 1$ for all s. Let $F_{(N-m_s)}(z)$ be the empirical estimate of $F(z)$ based on $(N - m_s)$ observations outside group s and let $F_{m_s}(z)$ be the empirical estimate of $F(z)$ based on m_s observations in group s. Then a general form of the S-fold cross-validated estimate of T_N is given by

$$CV_{S-fold}(\theta) = \sum_{s=1}^{S} \frac{m_s}{N} \int L\left(z, F_{(N-m_s)}(z)\right) dF_{m_s}(z). \tag{2}$$

Let $\hat{f}^{(-m_S)}(x;\theta)$ be the regression estimate based on the $(N - m_s)$ observation not in group s. Then the least squares S-fold cross-validated estimate of T_N is given by $CV_{S-fold}(\theta) = \sum_{s=1}^{S}(m_s/N)\sum_{k=1}^{m_s}(1/m_s)\left(y_k - \hat{f}^{(-m_s)}(x_k;\theta)\right)^2$.

The cross-validation score function can be written as a function of $(S+1)$ means and estimates a location-scale parameter of the corresponding s-samples. Let $u = L(v)$ be a function of a random variable v. In the S-fold cross-validation case, a realization of the random variable v is given by $v_k = \left(y_k - \hat{f}^{(-m_S)}(x_k;\theta)\right)$ $k = 1, ..., m_s \ \forall s$ with

$$\begin{aligned} CV_{S-fold}(\theta) &= \sum_{s=1}^{S} \frac{m_S}{N}\left[\frac{1}{m_s}\sum_{k=1}^{m_s} L(v_k)\right] = \sum_{s=1}^{S} \frac{m_S}{N}\left[\frac{1}{m_s}\sum_{k=1}^{m_s} u_k\right] \\ &= \hat{\mu}\left[\hat{\mu}_1(u_{11}, ..., u_{1m_1}), ..., \hat{\mu}_S(u_{S1}, ..., u_{Sm_S})\right], \end{aligned} \tag{3}$$

where u_{sj} denotes the j-th element of the s-th group, $\hat{\mu}_s(u_{s1}, ..., u_{sm_1})$ denotes the sample mean of the s-th group and $\hat{\mu}$ is the mean of all the sample group means. Consider only the random sample of the s-th group and let $F_{m_s}(u)$ be the empirical distribution function. Then $F_{m_s}(u)$ depends in a complicated way on the noise distribution $F(e)$, the θ values and the loss function $L(\cdot)$. In practice $F(e)$ is unknown (except for the assumption of symmetry around 0). Whatever the loss function would be (L_2 or L_1), the distribution $F_{m_s}(u)$ is always concentrated on the positive axis with an asymmetric distribution. Another important consequence is the difference in the tail behavior of the distribution $F_{m_s}(u)$. L_2 creates a more heavy-tailed distribution than the L_1 loss function.

3 Robust $S-$fold cross-validation score function

A classical cross-validation score function with L_2 or L_1 works well in situations where many assumptions (such as $e_k \sim N\left(0, \sigma^2\right)$, $E\left[e_k^2\right] < \infty$ and no outliers)

are valid. These assumptions are commonly made, but are at best approximations to reality. There exist a large variety of approaches towards the robustness problem. The approach based on influence functions [2] will be used here. Rather than assuming $F(e)$ to be known, it may be more realistic to assume a mixture noise model for representing the quantity of contamination. The distribution function for this noise model may be written as $F_\epsilon(e) = (1-\epsilon)F(e) + \epsilon H(e)$, where ϵ and $F(e)$ are given and $H(e)$ is an arbitrary (unknown) distribution, both $F(e)$ and $H(e)$ being symmetric around 0. An important remark is that the regression estimate $\hat{f}^{(-m_s)}(x;\theta)$ must be constructed via a robust method, for example the weighted LS-SVM [6].

In order to understand why certain location-scale estimators behave the way they do, it is necessary to look at the various measures of robustness. The effect of one outlier on the location-scale estimator can be described by the influence function (IF) which (roughly speaking) formalizes the bias caused by one outlier. Another measure of robustness is how much contaminated data a location-scale estimator can tolerate before it becomes useless. This aspect is covered by the breakdown point of the location-scale estimator.

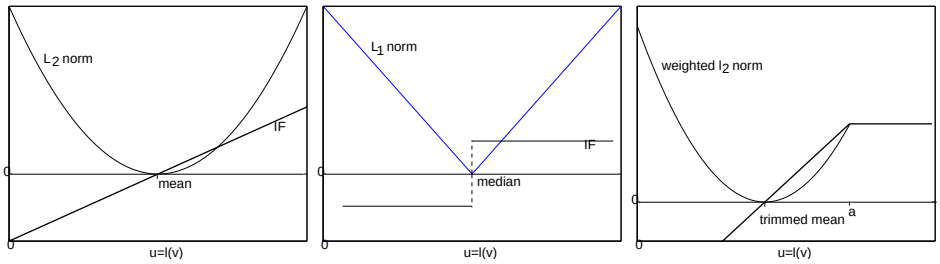

Fig. 1. Different norms with corresponding influence function (IF): (left) L_2; (Middle) L_1; (Right) weighted L_2.

The influence function of the S-fold crossvalidation score function based on the sample mean is sketched in Fig.1. We see that the IF is unbounded in $\mathbb{R}$. This means that an added observation at a large distance from the location-scale parameter (mean, median or trimmed mean) gives a large value in absolute sense for the IF. For example the breakdown point of the mean is 0%. One of the more robust location-scale estimators is the median, a special case of an L-estimator. Although the median is much more robust (breakdown point is 50 %) than the mean, its asymptotic efficiency is low. But in the asymmetric distribution case, the mean and the median estimate not the same quantity.

A compromise between mean and median (trade-off between robustness and asymptotic efficiency) is the trimmed mean defined as $\hat{\mu}_{(\beta_1,\beta_2)} = \frac{1}{a}\sum_{i=g_1+1}^{N-g_2} u_{(i)}$, where β_1 (the trimming proportion at the left) and β_2 (the trimming proportion at the right) are selected so that $g_1 = \lfloor N\beta_1 \rfloor$ and $g_2 = \lfloor N\beta_2 \rfloor$ and $a = N - g_1 - g_2$. The trimmed mean is a linear combination of the order statistics given

zero weight to g_1 and g_2, extreme observations at each end and equal weight $1/(N-g_1-g_2)$ to the $(N-g_1-g_2)$ central observations. Remark that $F_N(u)$ is asymmetric and has only a tail at the right, so we set $\beta_1 = 0$ and β_2 will be estimated from the data. The IF is sketched in figure 1.c. The S-fold cross-validation score function can be written as

$$CV_{S-fold}^{Robust}(\theta) = \hat{\mu}_{(0,\beta_2)}\left[\hat{\mu}_{(0,\beta_{2,1})}\left(u_{1(1)},...,u_{1(m_1)}\right),...,\hat{\mu}_{(0,\beta_{2,S})}\left(u_{S(1)},...,u_{S(m_S)}\right)\right] \tag{4}$$

with ordering $u_{(1)} \leq ... \leq u_{(N)}$. It estimates a location-scale parameter of the s-samples, where $\hat{\mu}_{(0,\beta_{2,s})}\left(u_{s(1)},...,u_{s(m_1)}\right)$ is the sample trimmed mean of the s-th group, and $\hat{\mu}_{(0,\beta_2)}$ is the trimmed mean of all the sample group trimmed means. We select $\beta_{2,s}$ to minimize the standard deviation of the trimmed mean based on a random sample.

More sophisticated location-scale estimators (M-estimators) have been proposed in [3].With appropriate tuning constant, M-estimates are extremely efficient; their breakdown point can be made very high. However, the computation of M-estimates, for asymmetric distributions, requires rather complex iterative algorithms and its convergence cannot be guaranteed in some important cases [4].

4 Repeated robust *S*-fold cross-validation score function

We propose now the following procedure. Permute and split repeatedly the random values $u_1,...,u_N$ - e.g. r times - into S groups as discussed. Calculate the S-fold cross-validation score function for each split and finally take the average of the r estimates

$$\text{Repeated_}CV_{S-fold}^{Robust}(\theta) = \frac{1}{r}\sum_{j=1}^{r} CV_{S-fold,j}^{Robust}(\theta). \tag{5}$$

The sampling distributions of the estimates, based on mean or trimmed means, are asymptotically Gaussian. The sampling distributions of both standard as robust cross-validations are shown in figure 2. The repeated S-fold cross-validation score function has about the same bias as the S-fold cross-validation score function, but the average of r estimates is less variable than for one estimate.

5 Illustrative example

In a simulation example we use here LS-SVMs and weighted LS-SVMs with RBF kernel as the non-linear function estimator. LS-SVMs are reformulations to the standard SVMs [8]. The cost function is a regularized least squares function with equality constraints. The solution of this set of equations is found by solving a linear Karush-Kuhn-Tucker system, closely related to regularization networks, Gaussian processes and kernel ridge regression. The solution can be computed efficiently by iterative methods like the Conjugate Gradient algorithm.

Simulations were carried out to show the differences among the two criteria $\text{CV}_{S-fold}^{L_2}$ and $\text{CV}_{S-fold}^{Robust}$. The data $(x_1,y_1),...,(x_{200},y_{200})$ are generated from

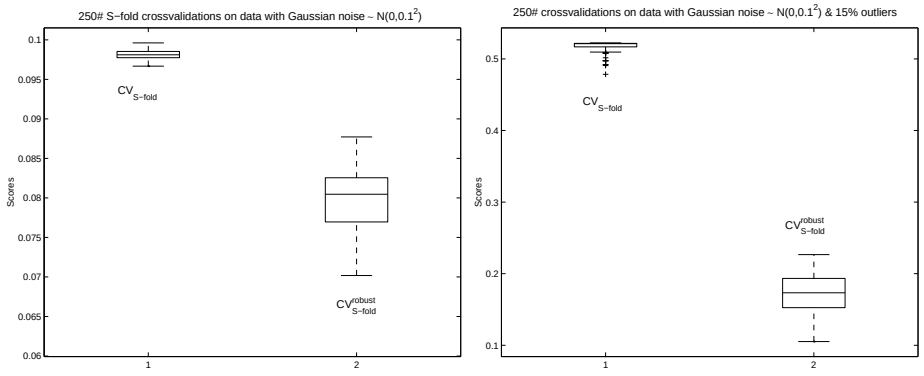

Fig. 2. Sampling distribution of the crossvalidation score function: (Left) Gaussian noise on $\sin(x)/x$; (Right) Gaussian noise & 15% outliers. Significant improvements are made by using the robust CV score function.

the nonparametric regression curve $y_k = f(x_k) + e_k$, $k = 1, ..., 200$, where the true curve is $f(x_k) = \sin(x)/x$, the observation errors $e_k \sim^{i.i.d.} N\left(0, 0.1^2\right)$ and $e_k \sim^{i.i.d.} F_\epsilon(e)$. Define the mixture distribution $F_\epsilon(e) = (1-\varepsilon) F(e) + \varepsilon \Delta_e$ which yields observations from F with high probability $(1-\varepsilon)$ and from the point e with small probability ε. We take for $F(e) = N(0, 0.1^2)$, $\Delta_e = N(0, 1.5^2)$ and $\varepsilon = (1 - \Phi(1.04)) = 0.15$, where $\Phi(\cdot)$ is the standard normal distribution function. A summary of the results is given in Table 1.

	$e \sim \mathcal{N}(0, 0.1^2)$			$e \sim (1-\epsilon)\mathcal{N}(0, 0.1^2) + \epsilon(\mathcal{N}(0, 1.5^2))$		
criteria	$\|\|.\|\|_\infty$	$\|\|.\|\|_1$	$\|\|.\|\|_2$	$\|\|.\|\|_\infty$	$\|\|.\|\|_1$	$\|\|.\|\|_2$
CV_{S-fold}	0.0470	0.0162	$3.95.10^{-4}$	0.148	0.0435	0.0030
CV_{S-fold}^{Robust}	0.0529	0.0178	$4.76.10^{-4}$	0.089	0.0280	0.0012

Table 1. The performance on a **test** set of a weighted LS-SVM with RBF kernel and tuned hyperparameters using different criteria. Significant improvements are obtained by the robust CV score method in the case of outliers.

6 Conclusion

Cross-validation methods are frequently applied for selecting hyperparameters in neural network methods, usually by using L_2 or L_1 norms. However, due to the asymmetric and non-Gaussian nature of the score function, better location-scale parameters can be used to estimate the performance. In this paper we have introduced a robust cross-validation score function method which applies concepts of influence function to the cross-validation methodology. Simulation results suggest that this method can be very effective and promising, especially with non-Gaussian noise distributions and outliers on data. The proposed method has a good robustness/efficiency trade-off such that it performs sufficiently well where

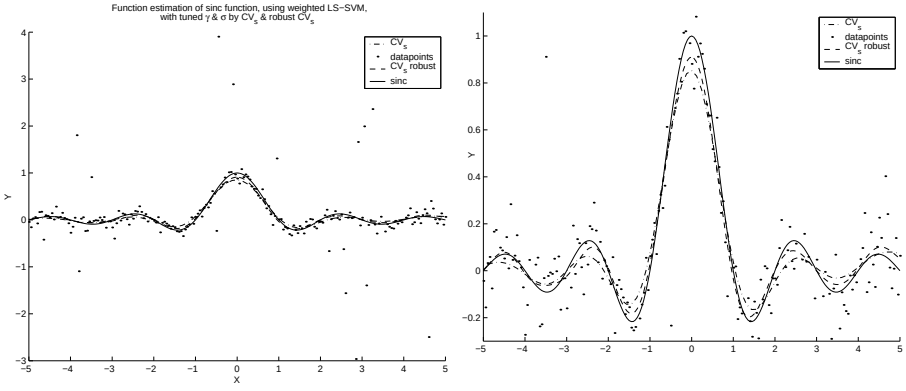

Fig. 3. Result of the weighted LS-SVM where the hyperparameters are tuned by S-fold cross-validation and robust S-fold cross-validation. The robust CV method outperforms the other method. (Left) 15% outliers; (Right) zoom on the estimate.

L_2 performs optimally. In addition the robust method will improve in many situations where the classical methods will fail.

Acknowledgements. This research work was carried out at the ESAT laboratory and the Interdisciplinary Center of Neural Networks ICNN of the Katholieke Universiteit Leuven, in the framework of the Belgian Program on Interuniversity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture (IUAP P4-02 & IUAP P4-24), the Concerted Action Project MEFISTO of the Flemish Community and the FWO project G.0407.02. JS is a postdoctoral researcher with the National Fund for Scientific Research FWO - Flanders.

References

1. Craven P. and Wahba G. (1979). Smoothing noisy data with spline functions. *Numer. Math.,* 31, 377-390.
2. Hampel F.R. (1974). The influence curve and its role in robust estimation. *J.Am.Stat.Assoc.* 69, 383-393.
3. Huber P.J. (1964). Robust estimation of a location parameter. *Ann.Math. Stat.* 35, 73-103
4. Marazzi A. and Ruffieux C. (1996). Implementing M-estimators of the Gamma Distribution. Lecture Notes in Statistics, 109, Springer Verlag, Heidelberg.
5. Stone M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Royal Statist. Soc. Ser. B* 36 111-147.
6. Suykens J.A.K., Vandewalle J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters.* Vol. 9, 293-300
7. Suykens J.A.K., De Brabanter J., Lukas L., Vandewalle J. (2002). "Weighted least squares support vector machines : robustness and sparse approximation", *Neurocomputing*, in press.
8. Vapnik V. (1995) *The nature of statistical learning theory,* Springer-Verlag.
9. Yang Y. and Zheng Z. (1992). "Asymptotic properties for cross-validation nearest neighbour median estimates in non-parametric regression: the L_1-view", *Probability and statistics*, 242-257.