ISCIS 27, Paris, October 3-4 2012
Computer and Information Sciences III,
E Gelenbe, R. Lent (Eds.), Springer

Statistical Tests using Hinge/ϵ-Sensitive Loss

Olcay Taner Yıldız[1] and Ethem Alpaydın[2]

[1] Department of Computer Engineering, Işık University, TR-34980, Istanbul, Turkey
olcaytaner@isikun.edu.tr
[2] Department of Computer Engineering, Boğaziçi University, TR-34342, Istanbul, Turkey alpaydin@boun.edu.tr

Abstract. Statistical tests used in the literature to compare algorithms use the misclassification error which is based on the 0/1 loss and square loss for regression. Kernel-based, support vector machine classifiers (regressors) however are trained to minimize the hinge (ϵ-sensitive) loss and hence they should not be assessed or compared in terms of the 0/1 (square loss) but with the loss measure they are trained to minimize. We discuss how the paired t test can use the hinge (ϵ-sensitive) loss and show in our experiments that doing that, we can detect differences that the test on error cannot detect, indicating higher power in distinguishing between the behavior of kernel-based classifiers (regressors). Such tests can be generalized to compare $L > 2$ algorithms.

1 Introduction

Statistical tests in the literature, namely, paired t test, 5×2 cv t test [1], 5×2 cv F test [2], are based on the misclassification error which corresponds to 0/1 loss. Support vector machine classifiers [3] are trained to minimize the *hinge loss* which not only checks whether the decision is on the right side of the boundary but also its position in the margin. Let us say $f(x^t) \in \Re$ is the kernel classifier output for input x^t and $y^t \in \{-1, +1\}$ is the desired output, 0/1 and hinge loss are defined as (Fig. 1):

$$0/1 \text{ loss} = \begin{cases} 0 \text{ if } f(x^t)y^t \geq 1 \\ 1 \text{ otherwise} \end{cases} \tag{1}$$

$$\text{hinge loss} = \begin{cases} 0 & \text{if } f(x^t)y^t \geq 1 \\ 1 - f(x^t)y^t & \text{otherwise} \end{cases} \tag{2}$$

Misclassification error only checks whether the classifier output is on the correct side of the boundary; hinge loss differs in two respects: (1) It also penalizes slightly those instances that are on the correct side but are in the *margin*, that is, ot classified with enough confidence, and (2) the misclassified instances are penalized linearly proportional to how deep they are in the wrong side.

Most of the regression algorithms are trained to minimize the square loss. In support vector regression, we use the ϵ-sensitive loss which tolerate errors up to ϵ and redefines the margin as the ϵ-tube. Let us say $f(x^t) \in \Re$ is the support

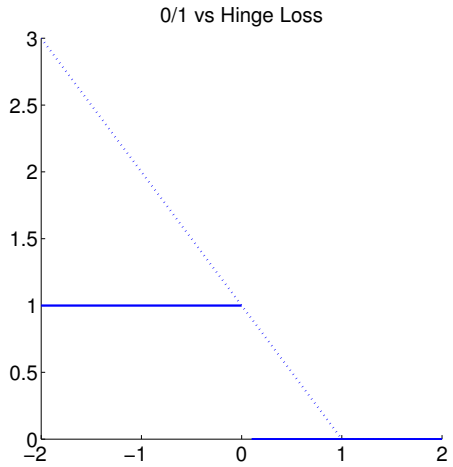

Fig. 1. 0/1 vs. hinge loss as a function of $f(x^t)$ for $y^t = 1$.

vector regression output for input x^t and $y^t \in \Re\}$ is the desired output, square and ϵ-loss are defined as (Fig. 2):

$$\text{square loss} = |y^t - f(x^t)|^2 \tag{3}$$

$$\epsilon\text{-sensitive loss} = \begin{cases} 0 & \text{if } |y^t - f(x^t)| \leq \epsilon \\ |y^t - f(x^t)| - \epsilon & \text{otherwise} \end{cases} \tag{4}$$

ϵ-sensitive loss differs from square loss in two respects: (1) it does not penalize those instances that are in the ϵ-tune, and, (2) the instances are penalized linearly proportional to how much they are far from the correct output. Therefore, ϵ-sensitive loss is more tolerant to noisy instances and thus more robust than the square loss.

Taking the losses used in training into account while comparing the test performance of kernel algorithms would thus enable to better distinguish between their generalization behavior. The two kernel algorithms compared may be using two different kernels or their kernels may be using two different sources of input, etc. and we want to check if there is a significant difference between them, for example, to test whether a new proposed kernel leads to improvement.

In statistical testing, we run both algorithms a number of times on a number of (training, validation) folds and compare the distributions of validation results for statistically significant difference; typically, k-fold cross-validation is used to generate k (training, validation) data set pairs by resampling from a single data set [4].

This paper is organized as follows: In Section 2, we discuss paired t test usually used for error comparison and discuss how it can also use hinge or ϵ-sensitive loss. We give our experimental results in Section 3 and conclude in Section 4.

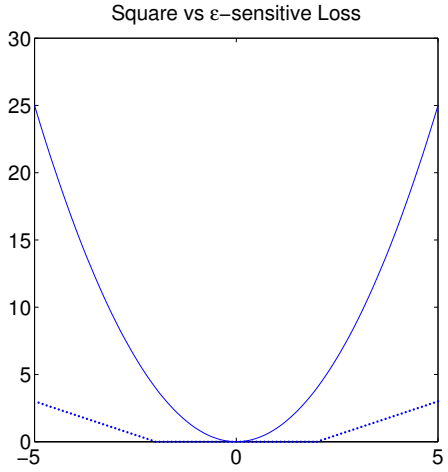

Fig. 2. Square vs ϵ-sensitive loss.

2 Paired t Test for Comparison

Let us say for all folds, $j = 1, \ldots, k$, we train both algorithms on training fold j and test on validation fold j and obtain the performance value $x_{ij}, i = 1, 2$ where x_{ij} is the total loss on the validation set where loss can be calculated using any of Eqs (1-4). It is important that all compared algorithms use the same training, validation data so that the comparison is *paired*. We want to check if the two sets of x_{1j} and x_{2j} can be said to come from the same population or whether they come from two distinct populations.

In the paired t test, we assume that the populations are normal and check if they have the same mean and for this, we test if their paired differences, $d_j = x_{1j} - x_{2j}$, have a mean of zero:

$$H_0 : \mu_d = 0 \text{ vs. } H_1 : \mu_d \neq 0$$

We calculate the average and variance of paired differences

$$m = \sum_{j=1} d_j/k \ , \ s^2 = \sum_j (d_j - m)^2/(k-1)$$

Under the null hypothesis, the statistic

$$t' = \frac{\sqrt{k}m}{s} \tag{5}$$

is t-distributed with $k-1$ degrees of freedom. We reject the null hypothesis that the two algorithms generalize equally well according to whichever loss we use if $|t'| > t_{\alpha/2,k-1}$ with $(1-\alpha)100$ percent confidence.

This test assumes that each x_{ij} is normally distributed. For any of Eqs (1-4), loss on each validation set is independent and identically distributed (but not necessarily normal). From the central limit theorem, we know that when we sum these up, the total loss converges to the normal distribution (unless the data set is small) and hence the normality assumption is tenable. In our experiments, the validation sets are not small, and again using the central limit theorem, we can also claim normality for the hinge values. As we will report later on, for all losses, the samples are indeed found to be normally distributed when tested with a normality test experimentally.

In Figure 3, we see a comparison on a synthetic two-dimensional classification data where we can see the discriminants and the margins. In this case of comparing the linear and Gaussian kernel, the test on error does not reject but the test on hinge loss finds a significant difference. As we see in Figure 3(a) and (b), the two kernels lead to discriminants which are not different in terms of the boundary but they have significantly different margins. In Figure 3(c), we see the histogram of paired differences whose expected value is not so far from 0 and hence, the paired t test on error does not reject the equality of means. If we similarly look at Figure 3(d), we see why the paired t test on hinge loss rejects; all paired differences are positive in Figure 3(d).

3 Experimental Results

We report our experiments on 11 data sets (*australian, breast, credit, cylinder, german, pima, mammographic, satellite47, tictactoe, titanic, transfusion*) for classification and 9 datasets (*abalone, add10, boston, california, concrete, puma8fh, puma8fm, puma8nh, puma8nm*) for regression from the UCI Repository. We used four different support vector machines with *linear, quadratic, cubic* and *Gaussian* kernels. All kernels are normalized for the discriminants to have the same scale. We use 10-fold cv and set $\alpha = 0.05$.

To be able to use the paired t test on hinge or ϵ-sensitive loss, we need to make sure that the assumption of normality holds. For all kernels, we used (the univariate version of the) normality test [5] and counted the percentage of times that the test rejects that the sample comes from a normal population. On each data set, we repeated the 10-fold experiment ten times and the values in Table 1 are hence proportions over $11 \times 10 = 110$ runs. As we see there, the percentage of rejects using the hinge loss compare well with the percentage of rejects using error, indicating that paired t test on hinge loss is as applicable as parametric tests on error. Actually it seems as if the kernel type is a more influential factor in the normality of results than the performance criterion.

On all classification data sets for all kernel types, we do pairwise comparisons using both error and hinge loss and compare the test results. For all 11 data sets and for ten independent runs for each, for all $4 \times 3/2 = 6$ pairwise comparison of four kernels, we have a total of 660 comparisons and the values reported are percentages. In Table 2, we see that the paired t tests on error and hinge loss agree in their decisions in $26.4 + 33.3 = 59.7$ percent of the cases. When they

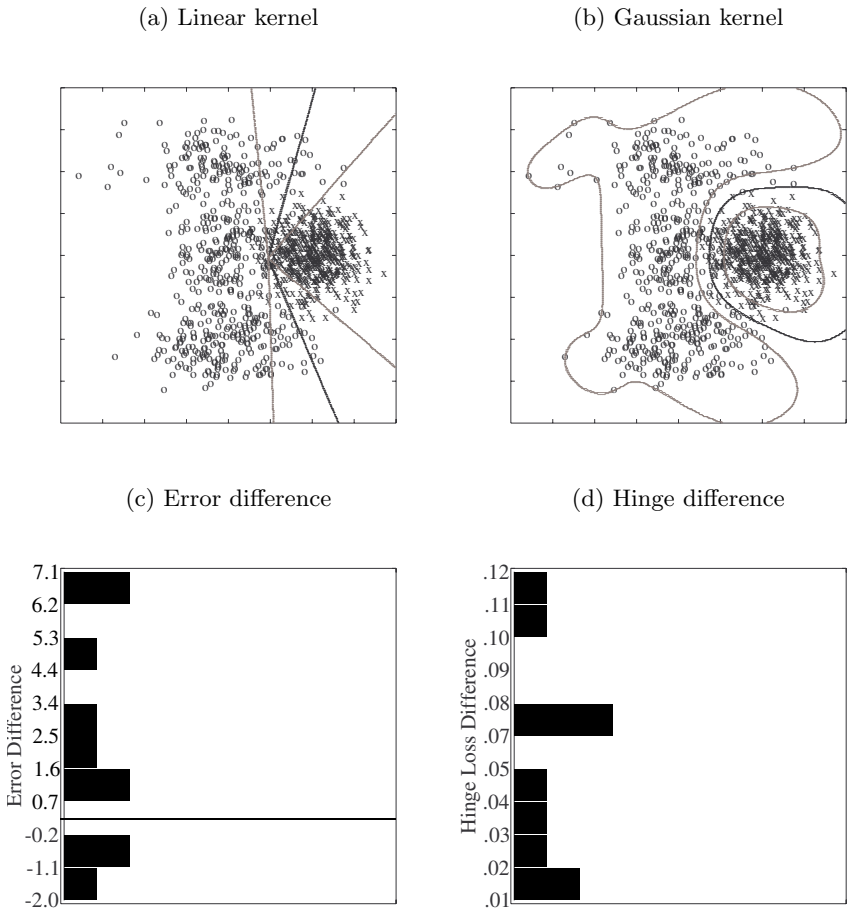

Fig. 3. Comparison of linear and Gaussian kernels on synthetic classification data (the nonlinearity with the linear kernel is due to normalization).

disagree, in 33.6 percent of the cases, the test on hinge loss finds a significant difference and rejects whereas the test on error considers them comparable; the opposite occurs in 6.7 percent of the cases. This shows that in around one-third of the cases, there is a difference between classifiers in terms of hinge loss and this difference information is lost if 0/1 error is used.

As an example where tests on error and hinge loss disagree, in Figure 4, we compare linear and cubic kernels on the *credit* data set. We see that though the classifiers are not significantly different in terms of error, they are significantly different in terms of hinge loss. The paired t test does not reject in terms of error because as we see in Figure 4(a), the two error distributions overlap whereas they

Table 1. Percentage of rejects of normality for 0/1, hinge, square, and ϵ-sensitive losses using different kernels.

	Loss Measure			
Kernel	0/1	Hinge	Square	ϵ-sens.
Linear	0.136	0.109	0.000	0.000
Quadratic	0.009	0.018	0.078	0.067
Cubic	0.009	0.000	0.033	0.022
Gaussian	0.055	0.045	0.067	0.056

Table 2. Percentage of agreement/disagreement of 0/1 and hinge loss.

	Hinge	
0/1	Accept	Reject
Accept	26.4	33.6
Reject	6.7	33.3

do not overlap in terms of hinge loss, as we see in Figure 4(b), and the paired t test on hinge loss rejects the null hypothesis of equality of means.

On all regression data sets for four polynomial kernel types, we did pairwise comparisons using both square and ϵ-sensitive loss and compare the test results. For all 9 data sets and for ten independent runs for each, for all $4 \times 3/2 = 6$ pairwise comparison of three kernels, we have a total of 540 comparisons and the values reported are percentages. In Table 3, we see that the paired t tests on square and ϵ-sensitive loss agree in their decisions in $11.5 + 83.1 = 94.6$ percent of the cases. Contrary to the classification case, they disagree only in 5.4 percent of the cases, indicating that using the paired t test, we can use ϵ-sensitive loss as well as the square loss.

As an example where tests on square and ϵ-sensitive loss disagree, in Figure 5, we compare linear and cubic kernels on the *boston* data set. We see that though the classifiers are significantly different in terms of square loss, they are not significantly different in terms of ϵ-sensitive loss. The paired t test rejects in terms of square loss because as we see in Figure 5(a), the two error distributions do not overlap whereas they overlap in terms of ϵ-sensitive loss, as we see in Figure 5(b), and the paired t test on ϵ-sensitive loss does not reject the null hypothesis of equality of means. For this case, square loss is sensitive to outliers (due to quadratic increase) and therefore rejects the null hypothesis whereas ϵ-sensitive loss is more robust to outliers and fails to reject the null hypothesis.

4 Conclusions and Future Work

Kernel-based, support vector machine classifiers and regressors are trained to minimize the hinge loss and ϵ-sensitive loss respectively. Hence their assessment

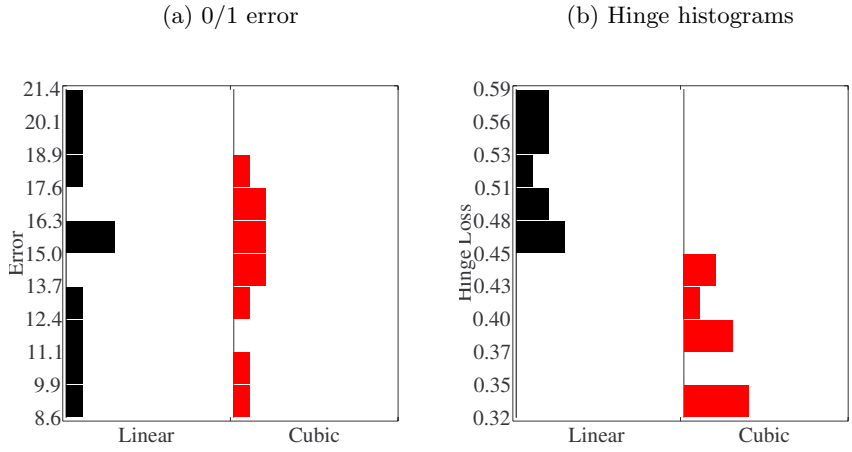

Fig. 4. On *credit* data set, results of comparison of linear and cubic kernels. Paired t test on error does not reject whereas test on hinge loss rejects.

Table 3. Percentage of agreement/disagreement of the paired t test on square and ϵ-sensitive loss.

		ϵ-**sensitive**
Square	Accept	Reject
Accept	11.5	2.4
Reject	3.0	83.1

and comparison should be done using the same measure and not misclassification error or square error. The hinge loss differs from the 0/1 loss and is a more informative measure because (1) it penalizes those instances in the margin—0/1 loss would not penalize them but hinge loss does because they are not classified with enough confidence, and (2) the misclassified instances on the wrong side of the boundary have equal loss of 1 with 0/1 loss, whereas the hinge loss penalizes them linearly proportional to their distance to the boundary thereby taking into account the confidence of the classifier in its decision. Taking these two into account gives more information about the confidence of the underlying classifier and allows us to distinguish between classifiers which are indistinguishable in terms of error. Indeed as we see in our experiments, statistical tests on hinge loss allow finding differences where the tests on error find no significant difference.

A similar rationale can also be put forward for favoring ϵ-sensitive loss over square loss: (1) It tolerates small, insignificant errors and (2) loss increases linearly as opposed to quadratically hence is more robust to outliers.

Here, we only discuss pairwise tests to compare two algorithms, but hinge and ϵ-sensitive loss can also be used to compare $L > 2$ algorithms, for example,

(a) Square loss (b) ϵ-sensitive histograms

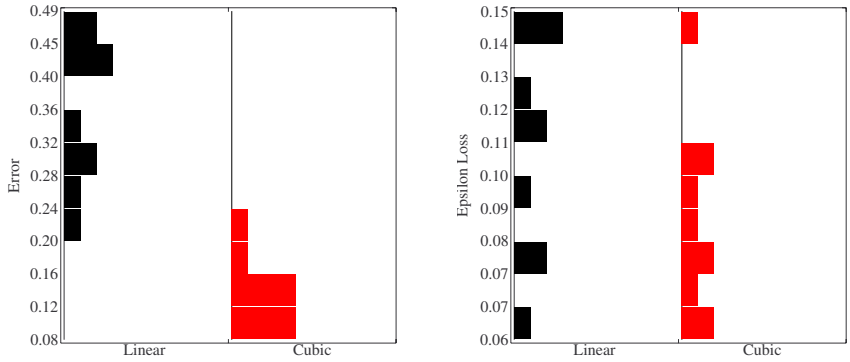

Fig. 5. On *boston* data set, results of comparison of linear and cubic kernels. Paired t test on squared loss rejects whereas test on ϵ-sensitive loss does not reject.

to compare L different kernels. Analysis of variance (ANOVA) can be used to test the equality of the means of $L > 2$ populations. There are also tests that can be used to find cliques of algorithms such that no pairwise test rejects between any two in the clique or ordering them [6]; such tests can also use the hinge loss or ϵ-sensitive loss instead of error. These are possible future directions for research.

Acknowledgments

This work is supported by TÜBİTAK EEEAG 109E186 and BAP 5701.

References

1. Dietterich, T.G.: Approximate statistical tests for comparing supervised classification learning classifiers. Neural Computation **10** (1998) 1895–1923
2. Alpaydın, E.: Combined 5×2 cv F test for comparing supervised classification learning classifiers. Neural Computation **11** (1999) 1975–1982
3. Vapnik, V.: The Nature of Statistical Learning Theory. Springer Verlag, New York (1995)
4. Alpaydın, E.: Introduction to Machine Learning. 2 edn. The MIT Press (2010)
5. Mardia, K.V.: Measures of multivariate skewness and kurtosis with applications. Biometrika **57** (1970) 519–530
6. Yıldız, O.T., Alpaydın, E.: Ordering and finding the best of $K > 2$ supervised learning algorithms. IEEE Transactions on Pattern Analysis Machine Intelligence **28**(3) (2006) 392–402