Statistical Tests using Hinge/ϵ-Sensitive Loss

Olcay Taner Yıldız Ethem Alpaydın

Department of Computer Engineering, Işık University, TR-34980, Istanbul, Turkey

Department of Computer Engineering, Boğaziçi University, TR-34342, Istanbul, Turkey

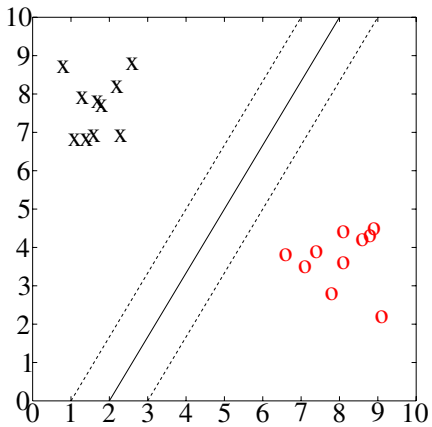
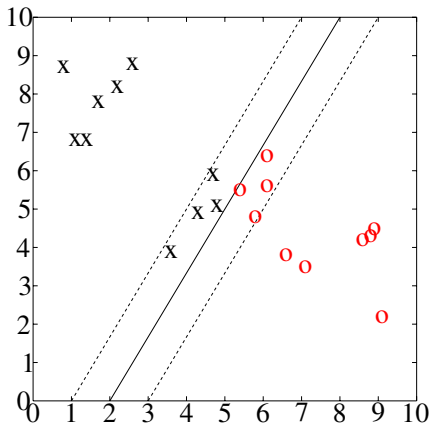
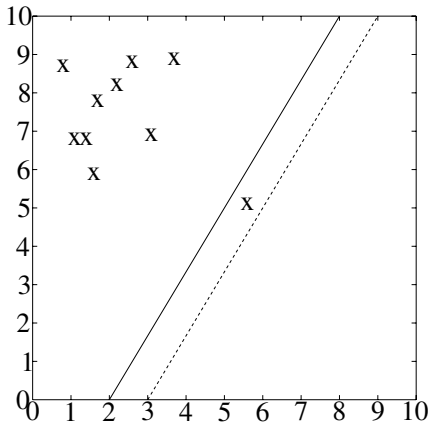
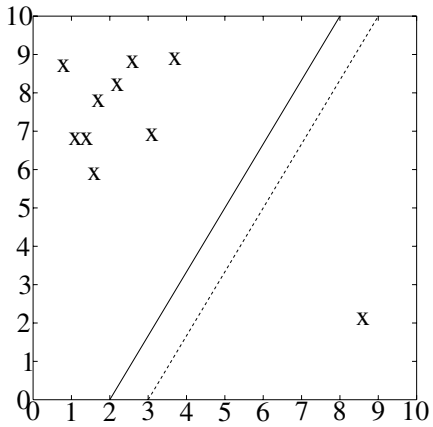
ISCIS 2012

Outline

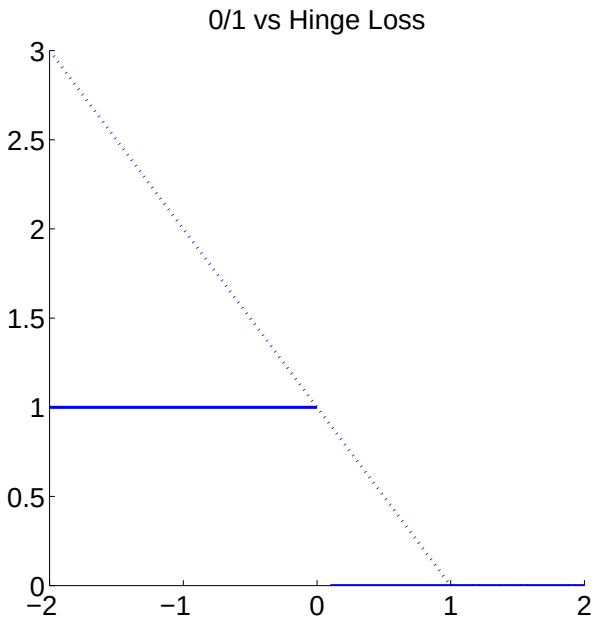
Motivation

- Statistical tests based on the misclassification error which corresponds to 0/1 loss.
- Support vector machine (SVM) classifiers trained to minimize the *hinge loss*.
- SVM's should not be compared in terms of the 0/1 but with the *hinge loss* they are trained to minimize.

Motivation

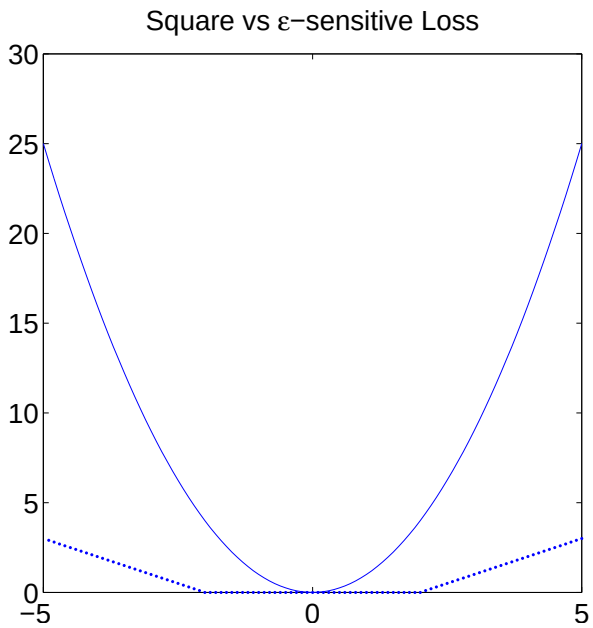
0/1 vs. hinge loss



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

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

Square loss vs. ϵ-sensitive loss



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

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

SVM comparison based on Hinge/ϵ-Sensitive Loss

- Two kernel algorithms compared may be using two different kernels or their kernels are using two different sources of input.
- Check if there is a significant difference between kernel algorithms.

Paired t Test

- For all folds, $j = 1, \ldots, k$, we train both algorithms on training fold j and test on validation fold j.
- We 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 equations above.
- As the null hypothesis, we test if their paired differences, $d_j = x_{1j} - x_{2j}$, have a mean of zero:

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

Paired t Test

- The average and variance of paired differences

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

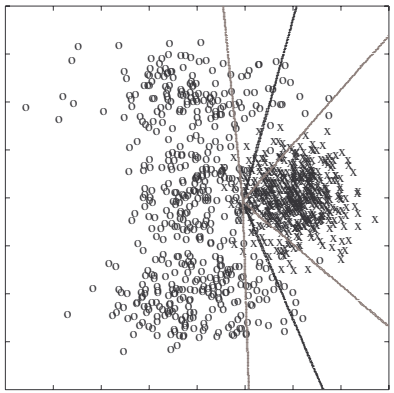
- Under the null hypothesis, the statistic

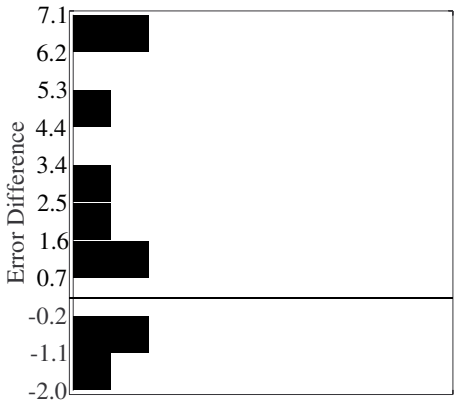
$$t' = \frac{\sqrt{k}m}{s} \tag{1}$$

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.

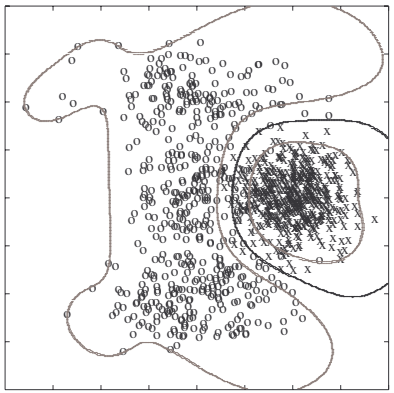
Comparison of linear/Gaussian kernels

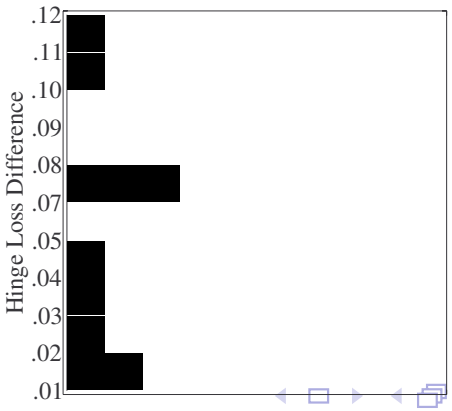
(a) Linear kernel





(b) Gaussian kernel





General Setup

- 11 classification data sets (*australian, breast, credit, cylinder, german, pima, mammographic, satellite47, tictactoe, titanic, transfusion*)
- 9 regression datasets (*abalone, add10, boston, california, concrete, puma8fh, puma8fm, puma8nh, puma8nm*)
- Four different support vector machines with *linear, quadratic, cubic* and *Gaussian* kernels.
- All kernels are normalized.
- We use 10-fold cv and set $\alpha = 0.05$ for paired t test.

Setup

- We used the univariate version of the normality test (Mardia, 1970) 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.

Results

Table : 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

Setup

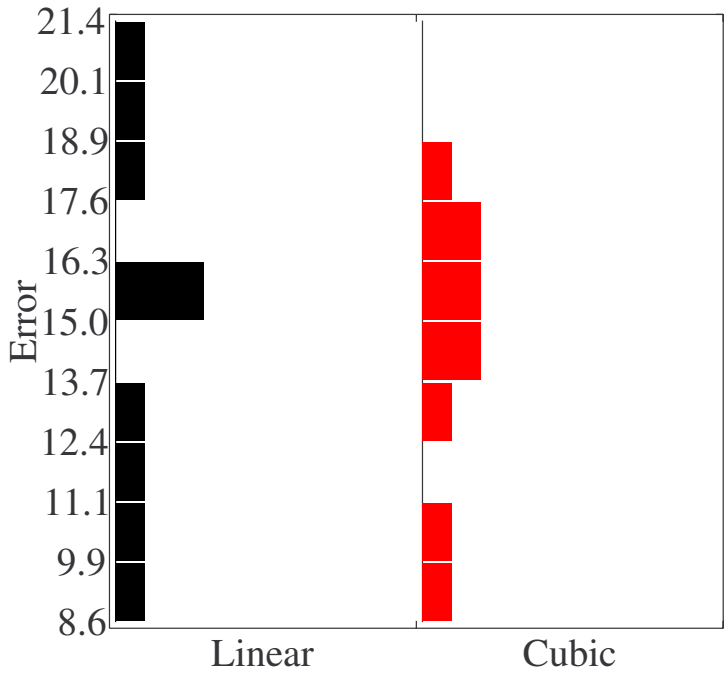
- On all classification data sets for all kernel types, we do pairwise comparisons using both error and hinge loss and compare the test results.
- For all 11 data sets and for ten independent runs for each, for all $4 \times 3/2 = 6$ pairwise comparison of four kernels, we have a total of 660 comparisons.

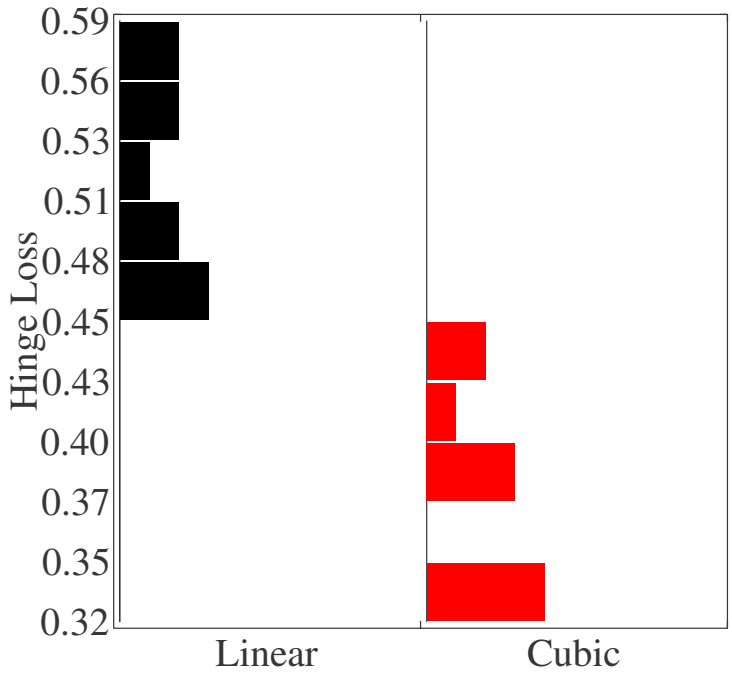
Results

Table : 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

Case Study: Linear and cubic kernels on *credit*





Setup

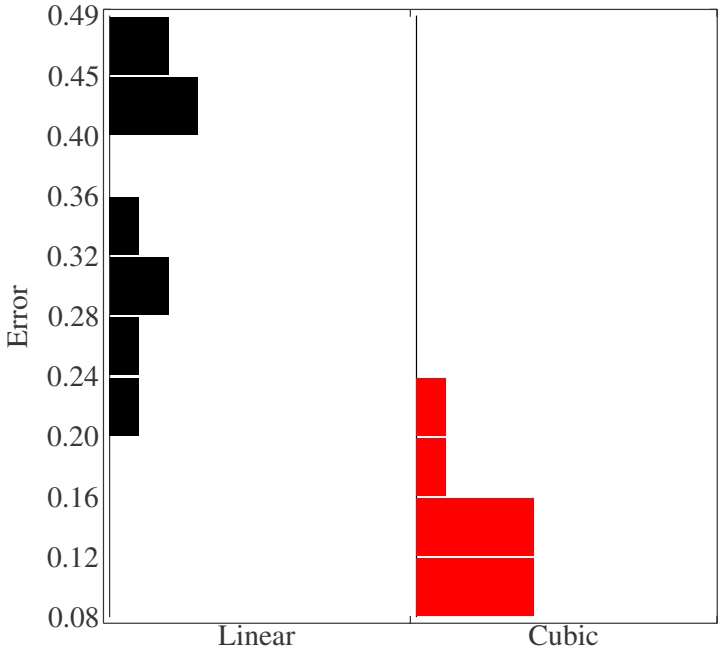
- On all regression data sets for four polynomial kernel types, we did pairwise comparisons using both square and ϵ-sensitive loss and compare the test results.
- For all 9 data sets and for ten independent runs for each, for all $4 \times 3/2 = 6$ pairwise comparison of three kernels, we have a total of 540 comparisons.

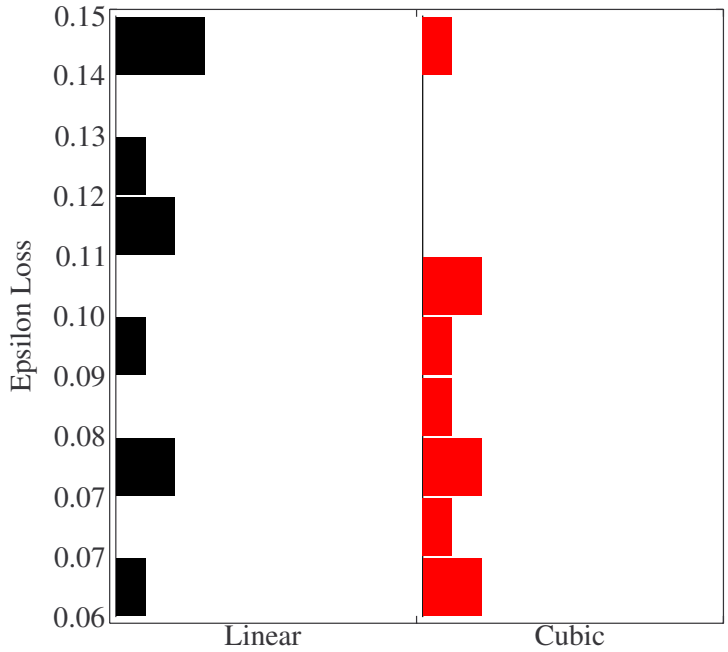
Results

Table : 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

Case Study: Linear and cubic kernels on *boston*





Summary

- SVM classifiers and regressors are trained to minimize the hinge loss and ϵ-sensitive loss respectively.
- Their assessment and comparison should be done using the same measure and not misclassification error or square error.

Hinge Loss vs. 0/1 loss

- Hinge loss penalizes those instances in the margin, because they are not classified with enough confidence.
- Hinge loss penalizes misclassified instances linearly proportional to their distance to the boundary.

- 0/1 loss would not penalize them.
- The misclassified instances on the wrong side of the boundary have equal loss of 1 with 0/1 loss.

ϵ-sensitive loss vs. square loss

- ϵ-sensitive loss tolerates small, insignificant errors.
- ϵ-sensitive loss increases linearly as opposed to quadratically hence is more robust to outliers.

Future Work

- Compare $L > 2$ algorithms using hinge and ϵ-sensitive loss.
- Find cliques of algorithms or ordering algorithms (Yildiz and Alpaydin, 2006) using hinge and ϵ-sensitive loss.

Questions?