

Comparing Algorithms

If we try different algorithms with different parameters, we want to select the best one.

We can select the algorithm/parameters that performs best on 10-fold CV.

We also want to be confident that the best algorithm/parameters have been selected.

Let A_1 and A_2 be two algorithms (or perhaps the same algorithm with different parameters).

What is the difference between f_{a1} and f_{a2} ?

Poor Test: Difference of Means Test

Using 10-fold CV, calculate u_1 and $(s_1)^2$ for A_1 and u_2 and $(s_2)^2$ for A_2 .

$$s^2 = \frac{(s_1)^2 + (s_2)^2}{2} \quad t = s \sqrt{\frac{1}{10} + \frac{1}{10}}$$

The 95% confidence interval is $u_1 - u_2 \pm 2.101t$. If this interval is > 0 , then conclude that f_{a1} has higher error. If < 0 , then conclude lower error.

Good Test: Paired-Difference t test

This test uses the differences between folds.

Apply identical 10-fold CV to A_1 and A_2 .

Calc. Δ_{1i} and Δ_{2i} , $1 \leq i \leq 10$, for A_1 and A_2 .

$$\Delta_i = \Delta_{1i} - \Delta_{2i} \quad u = \frac{\sum_{i=1}^k \Delta_i}{k}$$

$$s^2 = \frac{\sum_{i=1}^k (\Delta_i - u)^2}{k - 1} \quad t = \frac{s}{\sqrt{10}}$$

Confidence interval is $u \pm 2.262t$.

Comparing More than Two Algorithms

We would like to answer the question:

Is the true error of A lower than A_1, A_2, \dots ?

Difficulty: If A has lower error than A_i with 95% confidence (0.05 significance) pairwise, that does not imply that A is better than all A_i with 95% confidence. The significances should be summed, so if there are 10 other algorithms, we end up with 50% confidence.

Answer: To give the answer “A is best” with 95% confidence, the sum of the significances of A vs. A_i should be no more than 0.05.

Too many algorithms spoil the dataset.

It is not possible to test an unlimited number of algorithms on a limited dataset, and then select the best algorithm with high confidence.

The question of which algorithm has the lowest holdout error or CV error can be answered by repeated holdout or CV.

Example

I show performance of the evaluation methods.

I generated a large LCD dataset (100,000 exs.). About half the examples are digit 0 (positive). The rest are distributed among the other digits.

I generated a smaller LCD dataset (1,000 examples) by sampling with replacement from the large dataset. This ensures that true error for the small dataset = error on the large dataset.

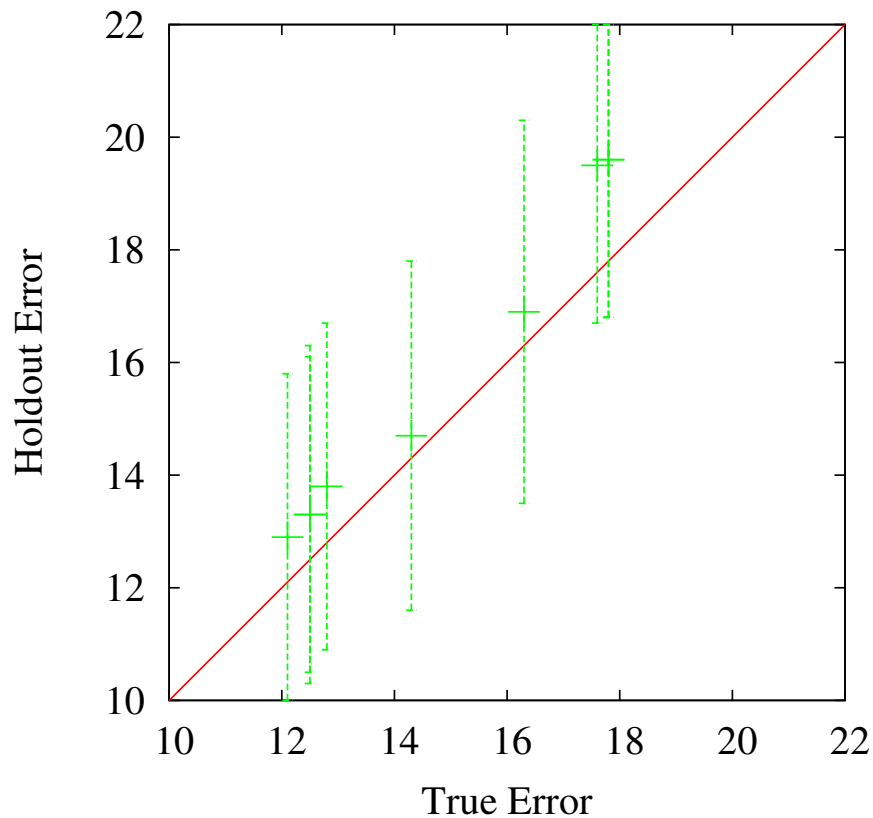
I generate and test a variety of SVMs.

Example True Error

C	σ^2	training error	true error
4	4	0.6%	17.6%
4	16	6.8%	12.8%
4	64	11.3%	12.5%
16	4	0.3%	17.8%
16	16	3.7%	14.3%
16	64	8.7%	12.1%
64	4	0.3%	17.8%
64	16	0.7%	16.3%
64	64	6.9%	12.5%

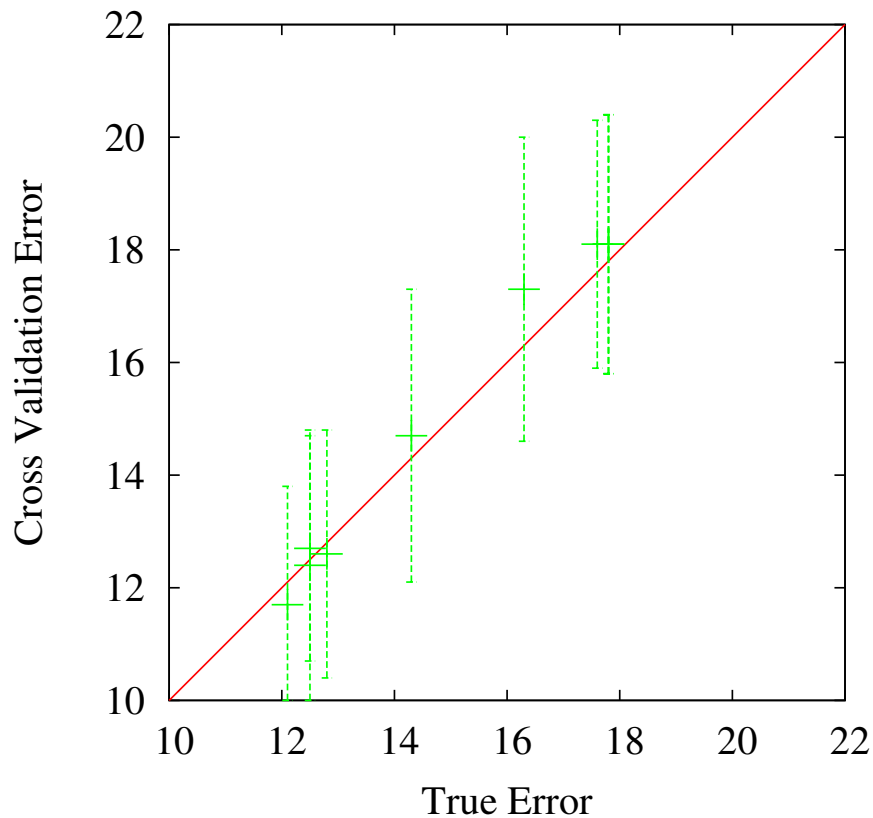
Example Holdout

C	σ^2	holdout error	95% interval
4	4	19.5%	± 2.8
4	16	13.8%	± 2.9
4	64	13.3%	± 2.8
16	4	19.6%	± 2.8
16	16	14.7%	± 3.1
16	64	12.9%	± 2.9
64	4	19.6%	± 2.8
64	16	16.9%	± 3.4
64	64	13.3%	± 3.0



Example 10-Fold CV

C	σ^2	CV error	95% interval
4	4	18.1%	± 2.2
4	16	12.6%	± 2.2
4	64	12.7%	± 2.0
16	4	18.1%	± 2.3
16	16	14.7%	± 2.6
16	64	11.7%	± 2.1
64	4	18.1%	± 2.3
64	16	17.3%	± 2.7
64	64	12.4%	± 2.4



Example Difference of Means

Each SVM minus SVM for $C = 16$ and $\sigma^2 = 64$.

C	σ^2	difference in means	95% interval
4	4	6.3%	± 2.8
4	16	0.9%	± 2.8
4	64	0.9%	± 2.6
16	4	6.4%	± 2.9
16	16	3.0%	± 3.1
64	4	6.4%	± 2.9
64	16	5.6%	± 3.2
64	64	0.6%	± 2.9

Example Paired-Difference t Test

Each SVM minus SVM for $C = 16$ and $\sigma^2 = 64$.

C	σ^2	difference in means	95% interval
4	4	6.3%	± 1.1
4	16	0.9%	± 0.4
4	64	0.9%	± 0.5
16	4	6.4%	± 1.1
16	16	3.0%	± 1.1
64	4	6.4%	± 1.1
64	16	5.6%	± 1.6
64	64	0.6%	± 0.7

