Error Rate Estimation

What is the error rate/loss of a NN/SVM?

Let f_a be the approximating function, and L be the loss function. The true risk $R[f_a]$ is the loss on the probability distribution P.

$$R[f_a] = \int L(f_a(\mathbf{x}), y)\, P(\mathbf{x}, y)\, d\mathbf{x}\, dy$$

The empirical risk $R_{\text{emp}}[f_a, S]$ is the loss on a sample S.

$$R_{\text{emp}}[f_a, S] = \sum_{(\mathbf{x},y)\in S} L(f_a(\mathbf{x}), y)/|S|$$

The problem of error rate estimation is to estimate or bound $R[f_a]$ from $R_{\text{emp}}[f_a, S]$.

It is assumed that S is drawn from P.

Estimates will be probabilistic due to sampling variations.

VC dimension analysis estimates loss from the training set. E.g., an SVM's expected error rate is bounded by |support vectors|/|training examples|.

Empirical statistics estimates loss from test sets.

What is an Error?

An error is a misclassification. A confusion matrix represented different types of errors:

Prediction	True Class 1	2	3
1	50	0	0
2	0	48	5
3	0	2	45

Correct prediction are on the diagonal. Off-diagonal values show one class being misclassified as another.

For the two-class positive/negative examples case, there are four possibilities:

Prediction	True Class: Positive	Negative
Positive	True Positives	False Positives
Negative	False Negatives	True Negatives

For example, it might be important to reduce false negatives (likely increasing false positives). As a result, a variety of different ratios and terminology have been developed.

Error	(FP + FN)/ (TP + FP + FN + TN)
Accuracy	(TP + TN)/ (TP + FP + FN + TN)
Sensitivity	TP/(TP + FN)
False Negative Rate	FN/(TP + FN)
Specificity	TN/(TN + FP)
False Positive Rate	FP/(TN + FP)
+ Predictive Value	TP/(TP + FP)
− Predictive Value	TN/(TN + FN)

Risk

For an error, the risk is the difference in cost/ utility between begin correct and making the error, quantified by a risk matrix.

	True Class		
Prediction	1	2	3
1	0	1	1
2	2	0	1
3	5	3	0

E.g., this matrix states that 5 is the risk when class 1 is misclassified as class 3.

Assigning Risks

Error rate corresponds to risk of 1 per error.

Assigning risk values is not as simple as assigning money amounts.

Losing \$1000 might be more than 1000 times worse than losing \$1.

Non-monetary factors (health, safety, convenience, freedom, happiness, ...) must be mapped to the same scale.

Different people will have different values.

Apparent Error Estimates

Let risk = 1 for mistakes and S = training set.

The apparent error is the error rate on the training set $R_{\mathrm{emp}}[f_a, S]$.

With unlimited examples and a limited classifier, apparent error converges to true error.

Overfitting makes apparent error too optimistic.

The nearest neighbor algorithm has no apparent error, but true error can be double the noise.

Example of Bad Estimates

Problem: Distinguishing 0 from other LCD digits

Dataset: 10% noise, ≈50% are positive.

Training set has 1000 examples. Test set has 100,000.

A SVM was trained using a Gaussian kernel with $\sigma^2 = 0.5$ and $C = 1$.

Error	
Apparent	Test
23.4%	28.6%
23.6%	28.5%
24.7%	7.1%
24.2%	28.1%
22.9%	28.4%

I'm not sure what happened on the third trial, but the two errors are not close in any trial.