Performance Evaluation and Experimental Comparisons for Classifiers

Prof. Richard Zanibbi

Performance Evaluation

Goal

We wish to determine the number and type of errors our classifier makes

Problem

Often the feature space (i.e. the input space) is vast; impractical to obtain a data set with labels for all possible inputs

Compromise (solution?...)

Estimate errors using a labeled sample (ideally, a *representative* sample)

The Counting Estimator of the Error Rate

Definition

For a labelled test data set *Z*, this the percentage of inputs from *Z* that are misclassified (#errors / |Z|)

Question

Does the counting estimator provide a complete picture of the errors made by a classifier?

A More General Error Rate Formulation

$$Error(D) = \frac{1}{|Z|}\sum_{j=1}^{|Z|}\{1 - I(l(z_j), s_j)\}, \quad z_j \in Z$$

where $I(a,b) = \begin{cases} 1, & \text{if a=b} \\ 0, & \text{otherwise} \end{cases}$
is an indicator function, and $l(z_j)$ returns the label (true class) for test sample $z_j \in Z$

*The indicator function can be replaced by one returning values in [0,1], to smooth (reduce variation in) the error estimates (e.g. using proximity of input to closest instance in the correct class)

Confusion Matrix for a Binary Classifier (Kuncheva, 2004)

	$D(\mathbf{x})$	
True Class	ω_1	ω_2
ω_1	7	0
ω_2	1	7

Our test set Z has 15 instances

One error (confusion) is made: a class 1 instance is confused for a class 2 instance

Larger Example: Letter Recognition (Kuncheva, 2004)

TABLE 1.1 The “H”-Row in the Confusion Matrix for the Letter Data Set Obtained from a Linear Classifier Trained on 10,000 Points.

“H” mistaken for:	A	B	C	D	E	F	G	H	I	J	K	L	M
No of times:	2	12	0	27	0	2	1	**165**	0	0	26	0	1
“H” mistaken for:	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
No of times:	31	37	4	8	17	1	1	13	3	1	27	0	0

Full Table: 26 x 26 entries

R·I·T

The “Reject” Option

Purpose

Avoid errors on ambiguous inputs through rejection (i.e. ‘skip’ the input). One approach: threshold discriminant function scores (e.g. estimated probabilities), reject inputs with max discriminant function score below the threhold

Confusion matrix: adding rejection: size (c+1) x c

Trade-off

Rejection avoids error, but often has its own cost (e.g. human inspection of OCR results, medical diagnosis)

Reject Rate

Reject Rate

Percentage of inputs rejected

Reporting

Initial recognition results should be reported with *no* rejection. Rejection may then be used, but with parameters and the rejection rate reported along with error estimates. A binary classification example:

- No rejection: error rate of 10%
- Discriminant scores < 0.5 : 30% reject rate, 2% error rate
- Discriminant scores < 0.9 : 70% reject rate, 0% error rate

R·I·T

Using Available Labeled Data: Training, Test and Validation Set Creation

Using Available Data

Labeled Data

Expensive to produce, as it often involves people (e.g. image labeling)

Available Data

Is finite; we want a large sample to learn model parameters accurately, but also want a large sample to estimate errors accurately

Common Division of Available Data into (Disjoint) Sets

Training Set

To learn model parameters

Testing Set

To estimate error rates

Validation Set

"Pseudo" test set used during training; stop training when improvements on training set do not lead to improvements on validation set (avoid overtraining)

R·I·T

Methods for Data Use

Resubstitution (avoid!)

Use all data for training and testing: optimistic error estimate

Hold-Out Method

Randomly split data into two sets. Use one half as training, the other as testing (pessimistic estimate)

- Can split into 3 sets, to produce validation set
- Data shuffle: split data randomly L times, and average the results

Methods for Data Use (Cont'd)

Cross-Validation

Randomly partition the data into K sets. Treat each partition as a test set, using the remaining data for training, then average the K error estimates.

- Leave-one-out: K=N (the number of samples), we "test" on each sample individually

Error Distribution

For hold-out and cross-validation, obtain an error rate *distribution* that characterizes the stability of the estimates (e.g. variance in error across samples)

Experimental Comparison of Classifiers

Factors to Consider for Classifier Comparisons

Choice of test set

Different sets can rank classifiers differently, even though they have the same accuracy over the population (over all possible inputs)

- Dangerous to draw conclusions from a single experiment, esp. if data size is small

Choice of training set

Some classifiers are *instable*: small changes in training set can cause significant changes in accuracy

- must account for variation with respect to training data

Factors, Cont'd

Randomization in Learning Algorithms

Some learning algorithms involve randomization (e.g. initial weights in a neural network, use genetic algorithm to modify parameters)

- For a fixed training set, the classifier may perform differently! Need multiple training runs to obtain a complete picture (distribution)

Ambiguity and Mislabeling Data

In complex data, often ambiguous patterns that have more than one acceptable interpretation, or errors in labeling (human error)

Guidelines for Comparing Classifiers (Kuncheva pp. 24-25)

1. Fix the training and testing procedure before starting an experiment. Give enough detail in papers so that other researchers can replicate your experiment
2. Include controls ("baseline" versions of classifiers) along with more sophisticated versions (e.g. see earlier binary classifier with "reject" example)
3. Use available information to largest extent possible, e.g. best possible (fair) initializations
4. Make sure the test set has not been seen during training
5. Report the run-time and space complexity of algorithms (e.g. big 'O'), actual running times and space usage

R·I·T

Experimental Comparisons: Hypothesis Testing

The Best Performance on a Test Set

....does not imply best performance over the entire feature space

Example

Two classifiers run on a test set have accuracies 96% and 98%. Can we claim that the error distributions for these are *significantly different?*

Testing the Null Hypothesis

Null Hypothesis

That the distributions in question (accuracies) do *not* differ in a statistically significant fashion (i.e. insufficient evidence)

Hypothesis Tests

Depending on the distribution types, there are a tests intended to determine whether we can *reject* the null hypothesis at a given *significance level (p,* the probability that we incorrectly reject the null hypothesis, e.g. $p < 0.05$ or $p < 0.01$)

Example Tests

chi-square, t-test, f-test, ANOVA, McNemar test, etc.

R·I·T