

Evaluating Hypotheses

Key Questions

What is the error rate of a hypotheses?

Is one hypothesis better than another?

How should a dataset be used for training and testing?

Statistics answers the first two questions, and machine learning has standard methods for the third question.

Question	Answer
Accuracy	Confidence Intervals
Comparison	Paired-Difference t Test
Train/Test	k -fold Cross Validation

Error Definitions

X is the set of all possible instances (examples).

Y is the set of all possible outcomes (classes).

(\mathbf{x}, y) is a labeled example, $\mathbf{x} \in X$ and $y \in Y$.

D is a probability distribution on labeled exs.

S is a data sample, a set of labeled examples.

S is randomly drawn according to D .

A hypothesis h is a function from X to Y .

The *sample classification error* (also called 0-1 loss) is:

$$error_S(h) = (1/|S|) \sum_{(\mathbf{x}, y) \in S} (\text{if } h(\mathbf{x}) \neq y \text{ then } 1 \text{ else } 0)$$

The *true classification error* is:

$$error_D(h) = \Pr_{(\mathbf{x}, y) \in D} (h(\mathbf{x}) \neq y)$$

Cost-Sensitive Classification

Sometimes, different mistakes have different costs. Recall the decision on turning onto an road. A cost matrix shows the tradeoffs.

	actual turn	actual wait
decide turn	0	100
decide wait	1	0

The cost matrix shows the difference of being wrong. Cost can more than money (time, health, safety, convenience, unhappiness, freedom, ...).

One measure is simply summing the mistakes' costs. Many other measures quantify different types of errors.

More Error Definitions

A *confusion matrix* shows the number and kind of mistakes. For example, this result from Weka:

```
  a  b  c  <-- classified as
49  1  0  |  a = Iris-setosa
 0 47  3  |  b = Iris-versicolor
 0  2 48  |  c = Iris-virginica
```

shows that 3 versicolor exs. were misclassified as virginica.

Recall is the rate that a class is correctly classified. E.g., the recall of setosa above is 49/50.

Precision is the rate that a prediction of that class is correct. E.g., the precision of setosa is 49/49.

More Error Definitions

The *F-measure* combines recall and precision:

$$F = 2 * \text{recall} * \text{precision} / (\text{recall} + \text{precision})$$

In the two-class pos/neg case, *sensitivity* is recall of positive and *specificity* is recall of negative. Also,

false positive = positive prediction is incorrect

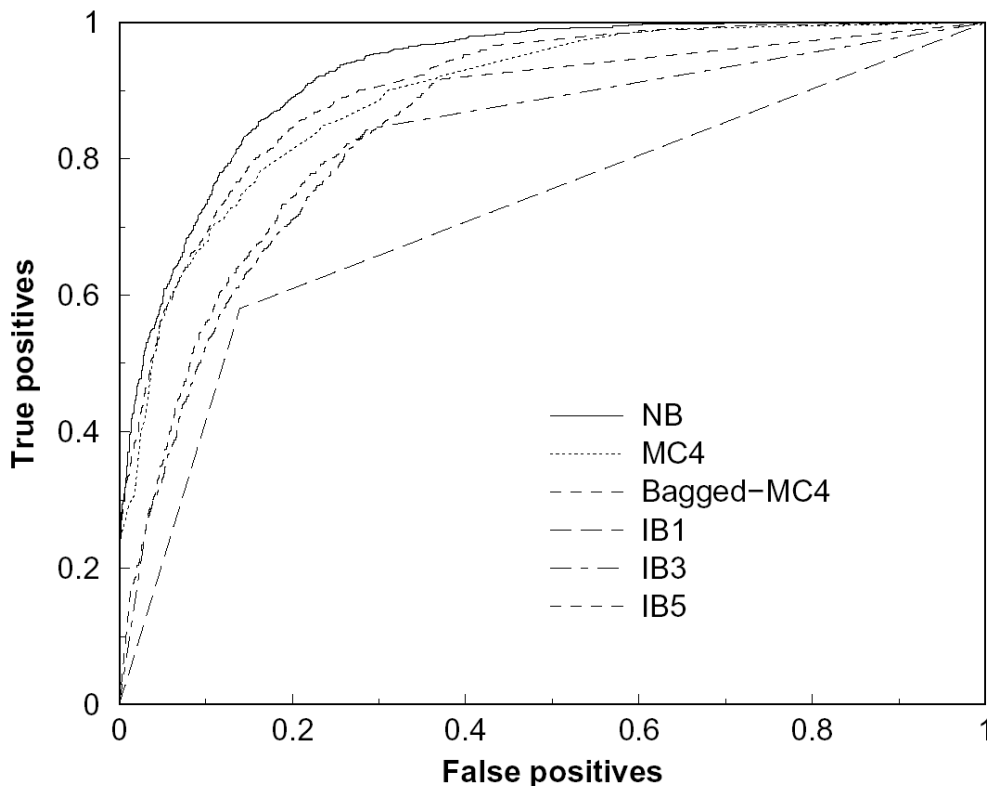
false positive rate = rate positive predictions are incorrect

false negative = negative prediction is incorrect

false negative rate = rate neg. predictions are incorrect

If sensitivity and specificity can vary (e.g., bias), we can create an *ROC curve*. This plots (1 – specificity, sensitivity) pairs. One measure is the area under the curve.

ROC Curve Example



Training Set and Test Set

Wrong

If S is used to train h , then $error_S(h)$ is the the *resubstitution error* and will almost always badly underestimate $error_D(h)$.

Right

Ideally, S should be independent of h .

In practice, a dataset will be partitioned into a training set T and a test set S .

This assumes that the dataset was sampled according to distribution D . Almost all machine learning depends on this assumption.

Error Rate

The first question “What is $error_D(h)$?” is modified to “How well does $error_S(h)$ estimate $error_D(h)$?”.

Assuming that the sample size n is large enough, the following is a 95% confidence interval:

$$error_S(h) \pm 1.96 \sqrt{\frac{error_S(h) (1 - error_S(h))}{n - 1}}$$

If $5 \leq n error_S(h) (1 - error_S(h))$, then this is an acceptable approximation.

Example: If $error_S(h) = 0.06$ and $n = 100$, then this gives 0.06 ± 0.047 as the 95% confidence interval.

Statistical Definitions

A *random variable* is the outcome of an experiment.

The *probability distribution* of a random variable Y specifies the probability $\Pr(Y = y_i)$ that Y has value y_i , for each possible y_i . A continuous prob. dist. specifies $\Pr(y_i \leq Y \leq y_j)$, or equivalently a density fn. $p(y_i)$.

The *expected value* or *mean* of Y is:

$$\mu = E[Y] = \sum_i y_i \Pr(Y = y_i) \text{ for discrete case}$$

$$\mu = E[Y] = \int_{-\infty}^{+\infty} y p(y) dy \text{ for continuous case}$$

For y_1, \dots, y_n , the mean is estimated by: $u = \frac{1}{n} \sum_{i=1}^n y_i$

More Statistical Definitions

The *variance* of a random variable Y is:

$$\sigma^2 = \text{Var}(Y) = E[(Y - \mu)^2]$$

For a sample $\{y_1, \dots, y_n\}$, the variance is estimated by:

$$s^2 = \frac{\sum_{i=1}^n (y_i - u)^2}{n - 1} = \frac{(\sum_{i=1}^n y_i^2) - nu^2}{n - 1}$$

The *standard deviation* of Y is $\sigma = \sqrt{\text{Var}(Y)}$. The estimate for the standard deviation is s .

For example, for k errors on n examples:

$$\text{error}_S(h) = u = k/n$$

$$s^2 = \frac{k - nu^2}{n - 1} = \frac{n(k/n - u^2)}{n - 1} = \frac{n(u - u^2)}{n - 1} = \frac{nu(1 - u)}{n - 1}$$

More Statistical Definitions

Binomial distribution: The probability of observing r successes in n trials when p is the probability of success.

$$\Pr(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

Normal distribution: This is the familiar bell-shaped probability distribution governed by the density function:

$$p(y) = \frac{e^{-(y-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

The *Central Limit Theorem* states that the average of n i.i.d. numbers converges to a normal dist. as n increases. The *Law of Large Numbers* states that the average of n i.i.d. numbers converges to the mean as n increases.

Hypothesis Testing

Often, we want to compare 2 hyps. (or 2 learning algs.).

Question: Does the error of h_1 differ from h_2 ?

Poor Test: Difference of Means Test

$$z = \frac{\text{error}_S(h_1) - \text{error}_S(h_2)}{\frac{\text{error}_S(h_1)(1-\text{error}_S(h_1))}{n-1} + \frac{\text{error}_S(h_2)(1-\text{error}_S(h_2))}{n-1}}$$

If $z \geq 1.96$, answer “ h_1 is worse than h_2 ”.

If $z \leq -1.96$, answer “ h_1 is better than h_2 ”.

The significance is 0.05, i.e., the probability of being wrong is 5% (or less).

If $|z| < 1.96$, then answer “I don’t know”.

Question: Does the error of h_1 differ from h_2 ?

Good Test: Paired-Difference Test. For each $(\mathbf{x}_i, y_i) \in S$:

$$\text{let } \delta_i = \begin{cases} 1 & \text{if } h_1(x_i) = y_i \neq h_2(x_i) \\ 0 & \text{if } h_1(x_i) = h_2(x_i) \\ -1 & \text{if } h_1(x_i) \neq y_i = h_2(x_i) \end{cases}$$

Calculate:

$$u = \frac{\sum_{i=1}^n \delta_i}{n} \quad s^2 = \frac{\sum_{i=1}^n (\delta_i - u)^2}{n - 1} \quad z = \frac{u}{s/\sqrt{n}}$$

If $z \geq 1.96$, answer “ h_1 is better than h_2 ”.

If $z \leq -1.96$, answer “ h_1 is worse than h_2 ”.

The significance is 0.05.

If $|z| < 1.96$, then answer “I don’t know”.

Question: Does the error of h_1 differ from h_2 ?

Better Test: Paired Difference t Test:

Rather than 1.96, a better critical value comes from the t distribution. The critical value depends on n and the significance. Here are some for 0.05 significance.

n	Critical Value	n	Critical Value
10	2.262	60	2.001
20	2.093	70	1.995
30	2.045	80	1.990
40	2.023	90	1.987
50	2.010	100	1.984

Question: Does the error of h_1 differ from h_2 ?

Best Test: Binomial Test (Signs Test):

Use δ_i from paired-difference test. Calculate:

$$n' = \sum_{i=1}^n |\delta_i| \quad m = \frac{n' + \sum_{i=1}^n \delta_i}{2}$$

$$p = \begin{cases} 2^{-n'} \sum_{i=m}^{n'} \binom{n'}{i} & \text{if } m > n'/2 \\ 2^{-n'} \sum_{i=0}^m \binom{n'}{i} & \text{if } m < n'/2 \end{cases}$$

$p \leq 0.025$ and $m > n'/2$, $\rightarrow h_2$ is better

If $p \leq 0.025$ and $m < n'/2 \rightarrow h_1$ is better

$p > 0.025$ or $m = n'/2 \rightarrow$ “I don’t know”

The significance is 0.05.

Algorithm Testing

Question: Does alg. A differ from alg. B ?

Good Test: Holdout

Partition the dataset into T and S .

Run algorithms A and B on T to get h_1 and h_2 .

Use S and some test for h_1 vs. h_2 .

If h_1 is better, answer “algorithm A is better”.

If h_2 is better, answer “algorithm B is better”.

Else answer “I don’t know”.

Advantages: Use $error_S(A)$ to est. true error.

Disadvantages: Requires a large dataset.

Misleading answer if alg. has high variance.

Question: Does alg. A differ from alg. B ?

Better Test: Stratified k -Fold Cross-Validation

Randomly split dataset into k almost-equal-size subsets S_i .

Each subset has almost-equal-proportions of classes.

For i from 1 to k ,

Training set T_i includes all subsets except S_i .

Run both algorithms on T_i .

Obtain $error_{S_i}(A)$ and $error_{S_i}(B)$

Calculate:

$$u = \frac{\sum_{i=1}^k error_{S_i}(A) - error_{S_i}(B)}{k}$$

$$s^2 = \frac{\sum_{i=1}^k (error_{S_i}(A) - error_{S_i}(B) - u)^2}{k - 1}$$

$$t = \frac{u}{s/\sqrt{k}}$$

For the critical value, use t distribution with $k - 1$ degrees of freedom and 0.05 significance. For $k = 10$, the critical value is 2.262

Advantages: Uses whole dataset for train and test.

Can use $error_{S_i}(A)$ to estimate true error.

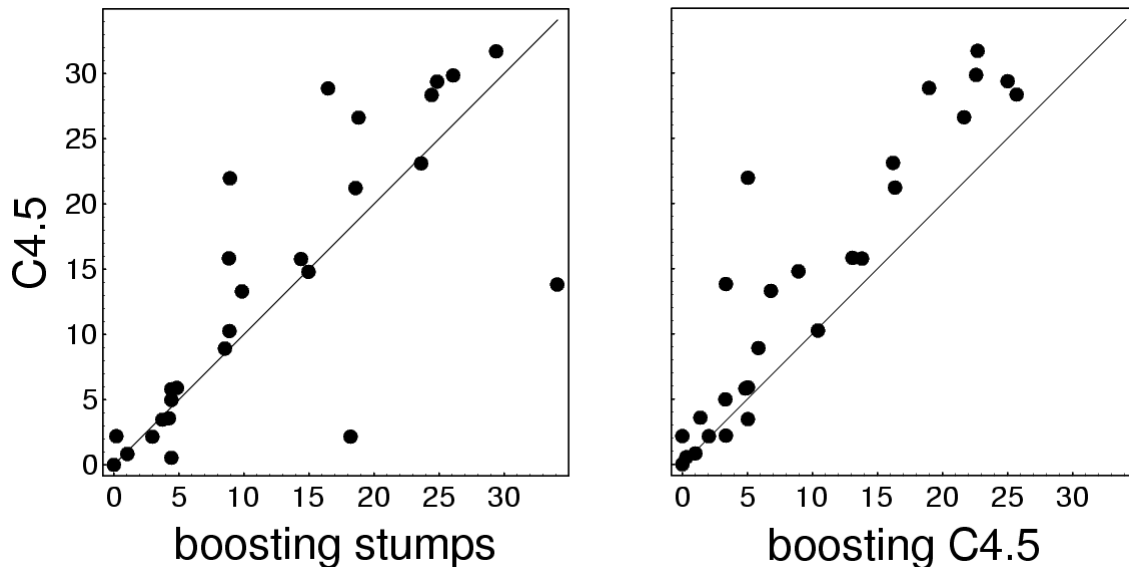
Variance of algorithm is detected.

Disadvantages: Needs more time than holdout.

See Dietterich paper for more discussion.

Comparison Graphs

A *comparison graph* plots the error rates of two algorithms on multiple datasets. Each point is an error rate pair on a dataset.



Problems with Comparing Algorithms

When comparing multiple algorithms, the *sum* of the significances should be less than a small fraction, strictly speaking, to be confident in all the conclusions. It is better to keep score rather than believe any single result.

No free lunch theorem: No algorithm will perform best on all datasets. We hope though to find good algorithms (or improve them) on “typical” datasets.