

Evaluating and Comparing the Performance of Machine Learning Algorithms

Measures of Performance

- Accuracy (percent correct over all test instances)
- Precision/Recall
- Area under ROC curve

- Precision
 - The ability to retrieve top-ranked documents that are mostly relevant.
- Recall
 - The ability of the search to find *all* of the relevant items in the corpus.

Precision / Recall

- Confusion matrix for a classifier:

	Classified Positive	Classified Negative
Positive Examples	True positives (TP)	False negatives (FN)
Negative Examples	False positives (FP)	True negatives (TN)

Some performance measures

- **Accuracy:** proportion of classifications, over all the N examples, that were correct:

$$\text{accuracy} = \frac{TP + TN}{N}$$

- **Recall (or true positive rate, or “detection rate”):** Proportion of positive examples that were classified correctly:

$$\text{recall} = \frac{TP}{TP + FN}$$

- **Precision :** Proportion of correct positive classifications over all positive classifications:

$$\text{precision} = \frac{TP}{TP + FP}$$

Accuracy

	Classified Positive	Classified Negative
Positive Examples	True positives (TP)	False negatives (FN)
Negative Examples	False positives (FP)	True negatives (TN)

÷

	Classified Positive	Classified Negative
Positive Examples	True positives (TP)	False negatives (FN)
Negative Examples	False positives (FP)	True negatives (TN)

Recall

	Classified Positive	Classified Negative
Positive Examples	True positives (TP)	False negatives (FN)
Negative Examples	False positives (FP)	True negatives (TN)

÷

	Classified Positive	Classified Negative
Positive Examples	True positives (TP)	False negatives (FN)
Negative Examples	False positives (FP)	True negatives (TN)

Precision

	Classified Positive	Classified Negative
Positive Examples	True positives (TP)	False negatives (FN)
Negative Examples	False positives (FP)	True negatives (TN)

÷

	Classified Positive	Classified Negative
Positive Examples	True positives (TP)	False negatives (FN)
Negative Examples	False positives (FP)	True negatives (TN)

Example

<u>Test data</u>	<u>Correct Classification</u>	<u>Model's Classification</u>
X ₁	T	T
X ₂	T	F
X ₃	F	T
X ₄	F	F
X ₅	F	T
X ₆	F	F
X ₇	F	F
X ₈	F	T

Accuracy =

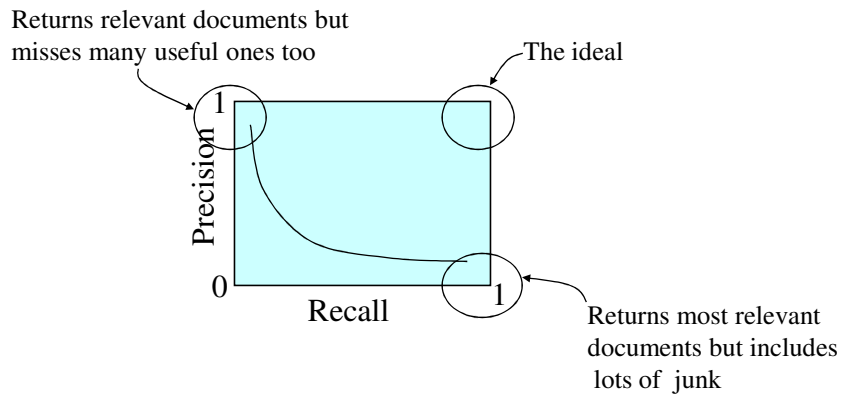
Recall =

Precision =

Interpretation of precision and recall

- Precision and recall are often plotted against one another, especially in “detection” applications (such as spam detection), when positive examples are sparse in the observed data.
- **Recall:** How often did the system correctly identify positive examples when it encountered them?
- **Precision:** How often did the system get positive classifications correct?
- How do these two measures trade off against one another? I.e., when would you care more about one than the other?
- MAP

Trade-off between Recall and Precision



From: www.cs.utexas.edu/~mooney/ir-course/slides/Evaluation.ppt

Computing Recall/Precision Points: Example

n	doc #	relevant
1	588	x
2	576	
3	589	x
4	342	
5	590	x
6	717	
7	984	
8	772	x
9	321	x
10	498	
11	113	
12	628	
13	772	
14	592	x

Let total # of relevant docs = 6
Check each new recall point:

$R=1/6=0.167$; $P=1/1=1$

$R=2/6=0.333$; $P=2/3=0.667$

$R=3/6=0.5$; $P=3/5=0.6$

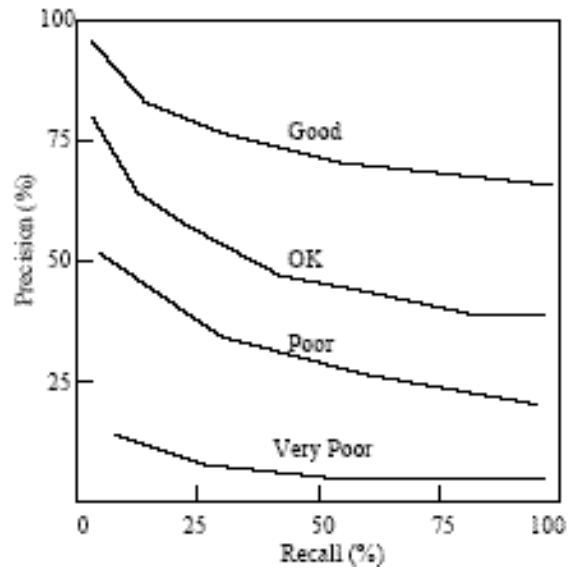
$R=4/6=0.667$; $P=4/8=0.5$

$R=5/6=0.833$; $P=5/9=0.556$

$R=6/6=1.0$; $p=6/14=0.429$

From: www.cs.utexas.edu/~mooney/ir-course/slides/Evaluation.ppt

Precision / Recall Curves



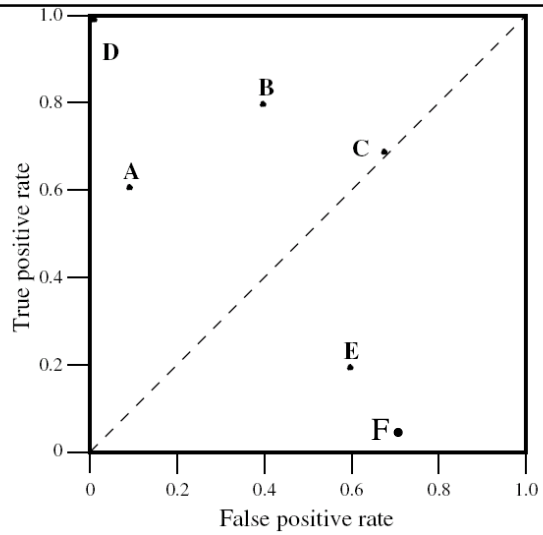
Receiver Operating Characteristic (ROC) Curves

- Alternative to precision/recall curves
- Shows tradeoff between true positive rate and false positive rate.

$$\text{True positive rate} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{False positive rate} = \text{FP}/(\text{TN} + \text{FP})$$

Example ROC diagram for discrete-valued classifiers



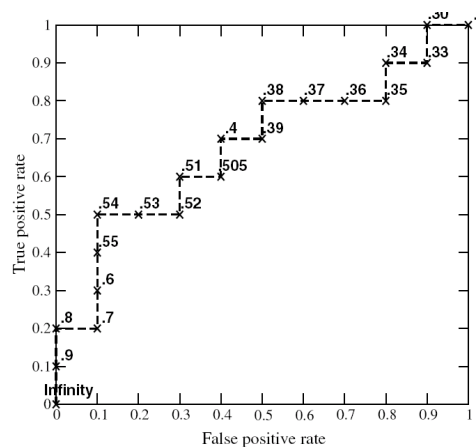
What is the interpretation of TPR = 0, FNR = 0?

How about TPR = 1, FNR = 1?

What is the ideal model? What is random guessing “true” with probability p ? Is F a good classifier?

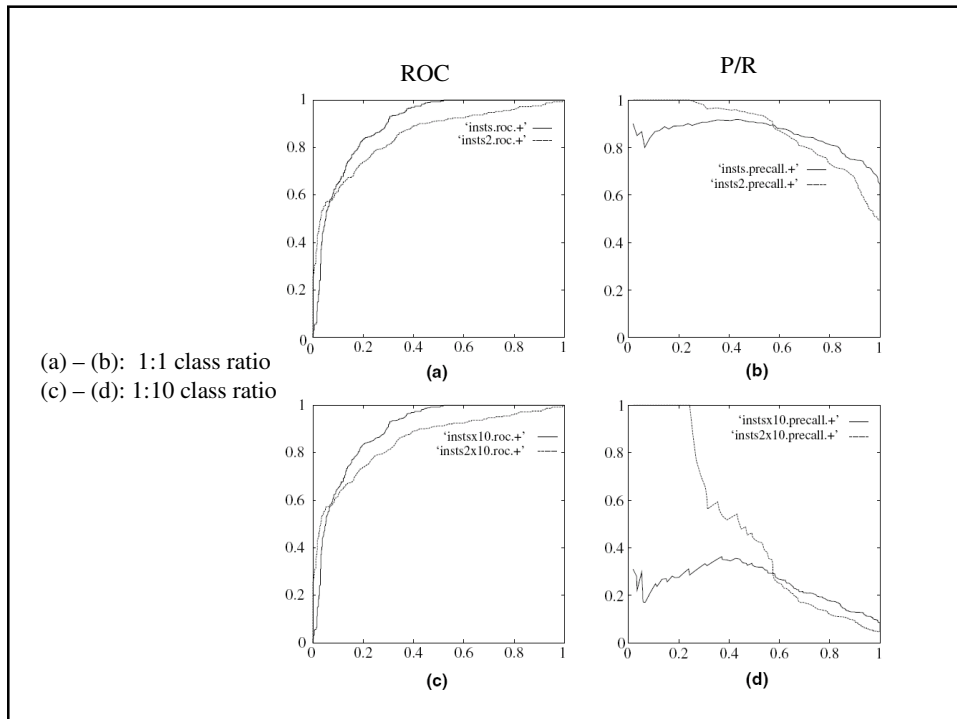
Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

Example ROC diagram for single continuous-valued classifier with varying threshold between ∞ and 0



What is accuracy of classifier at various points?

ROC curve shape is insensitive to “operating conditions”,
i.e., to balance between positive and negative instances or
cost of misclassification.



Area under ROC curve (AUC)

- Summary statistic: Area under ROC curve (AUC) = probability that classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.
- AUC is always between 0 and 1.

Averaging ROC curves

- For cross-validation, can average ROC curves for a classifier using validation sets T_0, \dots, T_k

Statistical Evaluation of Classifiers

Assigned reading: T. Mitchell,
Machine Learning, Ch. 5
(will be on e-reserve at the library)

- **What is this topic about?**
 - Mainly a review of some statistics, statistical hypothesis testing, and its application to machine learning
- **What questions do we want to answer?**
 - What is expected accuracy of a classification algorithm (i.e., over all instances in the instance space)?
 - Given two classification algorithms, which one has better accuracy for a given application, and how much better is it?

- Framework:

- An instance space X .
- A probability distribution D over instances in X .
- A hypothesis h that makes binary classifications on instances $x \in X$.
- Sample $error_S(h)$ on a sample $S \subseteq X$, where

$$error_S(h) = \Pr_{x \in S}[\text{incorrect}(h(x))]$$

- True error $error_D(h)$ of h over x , given distribution D , where

$$error_D(h) = \Pr_{x \in X}[\text{incorrect}(h(x))]$$

Question to be addressed today

How well does $error_S(h)$ estimate $error_D(h)$? I.e., can we put a confidence interval around $error_S(h)$ that gives $N\%$ confidence that $error_D(h)$ is in that interval?

That is, we want to say something like:

“With 95% confidence, $error_D(h)$ lies in the interval $error_S(h) \pm M$.”

- Basics of sampling theory:

We collect a random sample S of n independently drawn examples from distribution D .

We measure sample error $error_S(h)$.

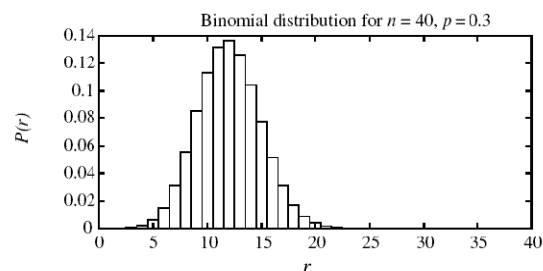
If we repeated this many times on different samples S_1, S_2, \dots, S_k , would get different values for $error_S(h)$.

Thus $error_S(h)$ can be considered a random variable.

What is distribution of $error_S(h)$?

Binomial distribution.

When is a distribution Binomial?



- n identical trials
- Outcome of trial can be one of two possible values, say 0 and 1
- Probability of 1 on a single trial is given by constant p .
- Trials are independent of one another.

- Let p be the probability that $h(\mathbf{x})$ will be an error for a given $\mathbf{x} \in S$. Let r be the number of errors observed over sample S . Let R be a binomially distributed random variable corresponding to r .

Expectation of R :

- Let p be the probability that $h(\mathbf{x})$ will be an error for a given $\mathbf{x} \in S$. Let r be the number of errors observed over sample S . Let R be a binomially distributed random variable corresponding to r .

Expectation of R :

$$E[R] = np$$

- Let p be the probability that $h(\mathbf{x})$ will be an error for a given $\mathbf{x} \in S$. Let r be the number of errors observed over sample S . Let R be a binomially distributed random variable corresponding to r .

Expectation of R :

$$E[R] = np$$

Variance of R :

- Let p be the probability that $h(\mathbf{x})$ will be an error for a given $\mathbf{x} \in S$. Let r be the number of errors observed over sample S . Let R be a binomially distributed random variable corresponding to r .

Expectation of R :

$$E[R] = np$$

Variance of R :

$$\text{Var}[R] = (E[R - E[R]])^2 = E(R^2) - (E(R))^2 = np(1 - p).$$

- Let p be the probability that $h(\mathbf{x})$ will be an error for a given $\mathbf{x} \in S$. Let r be the number of errors observed over sample S . Let R be a binomially distributed random variable corresponding to r .

Expectation of R :

$$E[R] = np$$

Variance of R :

$$\text{Var}[R] = np(1 - p).$$

Standard Deviation of R :

- Let p be the probability that $h(\mathbf{x})$ will be an error for a given $\mathbf{x} \in S$. Let r be the number of errors observed over sample S . Let R be a binomially distributed random variable corresponding to r .

Expectation of R :

$$E[R] = np$$

Variance of R :

$$\text{Var}[R] = np(1 - p).$$

Standard Deviation of R : $\sigma_R = \sqrt{np(1 - p)}$

- We can approximate:

$$p \approx \frac{r}{n} = error_S(h)$$

- So:

$$\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

Example

- Suppose that after training, your classifier h misclassifies 12 out of 72 test examples.
- $error_S(h) = r / n = 12/72 = 0.17$

$$\begin{aligned} \sigma_{error_S(h)} &= \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{(r/n)(1-r/n)}{n}} = \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}} \\ &= \sqrt{\frac{.17 \times .83}{72}} \approx .04 \end{aligned}$$

Confidence intervals

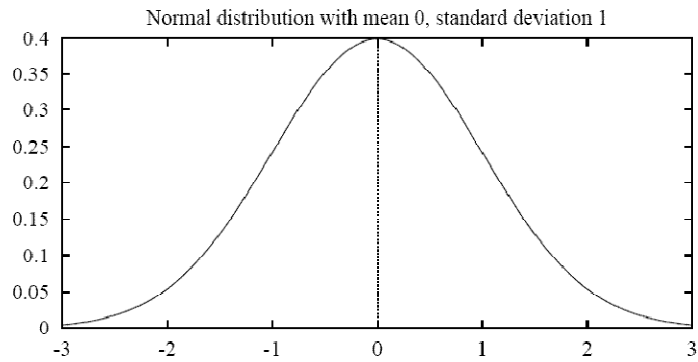
- We want to make a statement like:

“With 95% probability, the true error $error_D(h)$ lies in the interval

$$error_S(h) \pm f(error_S(h), n)$$

Confidence intervals

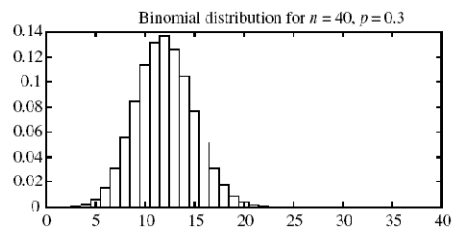
- Now, what is this function f ?
 - We know that $error_S(h)$ is binomially distributed, with mean $error_D(h)$ and standard deviation $\sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$.
 - Need to find the interval centered around value $error_S(h)$ that contains $N\%$ of the total probability under this distribution. This interval will contain $error_D(h)$ $N\%$ of the time.
 - Hard to calculate exactly for Binomial distributions, but a good approximation (at large sample sizes) can be obtained by approximating the Binomial distribution with the Normal (Gaussian) distribution.



Central Limit Theorem: Sum of a large number of independent, identically distributed (iid) random variables follows a distribution that is approximately Normal.

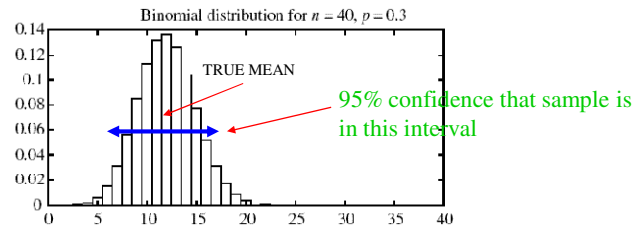
For large n , any binomial distribution is closely approximated by a Normal distribution with the same mean and variance.

- Need to find the interval centered around value $error_S(h)$ that contains 95 % of the total probability under this distribution. This interval will contain $error(h)$ 95 % of the time.



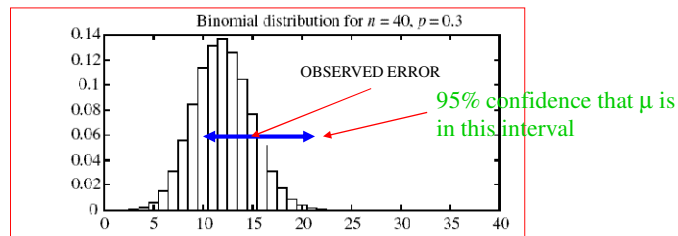
Distribution for $error_S(h)$

- Need to find the interval centered around value $error_S(h)$ that contains 95 % of the total probability under this distribution. This interval will contain $error(h)$ 95 % of the time.



Distribution for $error_S(h)$

- Need to find the interval centered around value $error_S(h)$ that contains 95 % of the total probability under this distribution. This interval will contain $error(h)$ 95 % of the time.



Distribution for $error_S(h)$

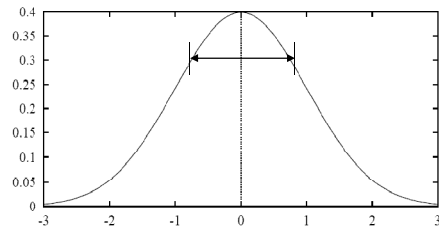
Statistics tables tell us that any new sample $error_S(h)$ has 95% probability of being at most 1.96 standard deviations from μ .

Equivalently, given $error_S(h)$, μ has 95% probability of being at most 1.96 standard deviations from $error_S(h)$.

Thus we can say that, with 95% confidence, the interval

$$error_S(h) \pm 1.96 \sigma$$

contains μ (= true error $error(h)$)

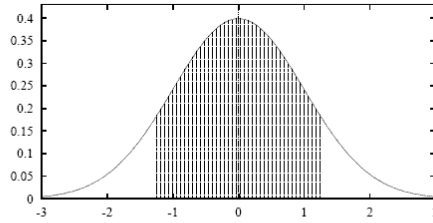


- Statistics tables give size of interval about the mean that contains $N\%$ of the probability mass under the Normal distribution.

- Example table:

Confidence Level $N\%$	50%	68%	80%	90%	95%	98%	99%
z_N	0.67	1.0	1.28	1.64	1.96	2.33	2.58

where z_N is half the width (measured in standard deviations) of the interval about the mean that contains $N\%$ of the total probability mass in the distribution.



80% of area (probability) lies in $\mu \pm 1.28\sigma$

N% of area (probability) lies in $\mu \pm z_N\sigma$

N%:	50%	68%	80%	90%	95%	98%	99%
z_N :	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Now, suppose random variable Y follows a Normal distribution, and we have a measured value y from Y . We can say:

y will fall into interval $\mu \pm z_N\sigma$ N% of the time, or
 μ will fall into interval $y \pm z_N\sigma$ N% of the time

- Now we can derive general expression for N% confidence intervals for discrete-valued hypotheses.

$error_S(h)$ follows Binomial distribution with $\mu = error_D(h)$ and $\sigma = \sqrt{\frac{error_D(h)(1-error_D(h))}{n}}$

- Approximate with Normal distribution with same mean and standard deviation.

N % confidence interval for discrete hypotheses:

With N % confidence, $error_D(h)$ is in the interval

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

Our two approximations ($error_D(h) \approx error_S(h)$ and Binomial distribution \approx Normal distribution) are good as long as $n \geq 30$.

Example

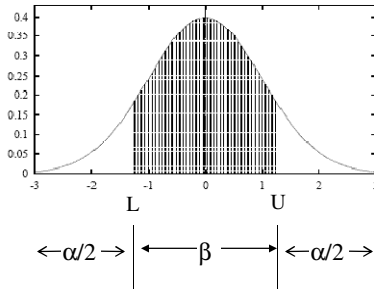
Suppose you test a hypothesis h and find that it commits $r=10$ errors on a sample S of $n=65$ randomly drawn test examples. What is the 90% two-sided confidence interval for the true error rate?

One-sided confidence bounds

In ML, usually care only about upper bound on error, not lower bound. E.g., we want to say,

“With N % confidence, $error_D(h) \leq U$ ”

- Note that Normal distribution is symmetric about mean, so any two-sided confidence interval based on a Normal distribution can be converted to corresponding one-sided interval with twice the confidence.



Let β be the probability that the error is in the interval $[L,U]$. I.e., $\beta = N/100$. Let $\alpha = 1 - \beta$. Then what is probability that the error is above U? $\alpha/2$. So probability that error is has upper bound U and no lower bound is $(1- \alpha/2)$.

In other words, a $100(1 - \alpha)\%$ confidence interval with lower bound L and upper bound U implies a $100(1- \alpha/2)$ confidence interval with upper bound U and no lower bound.

Example

Suppose h 's error rate is measured as $12/40$.

Two-sided 95% confidence interval:

$$\begin{aligned} & error_s(h) \pm z_N \sqrt{\frac{error_s(h)(1-error_s(h))}{n}} \\ & = .3 \pm 1.96 \sqrt{\frac{.3 \times .7}{40}} = .3 \pm .14 \end{aligned}$$

What is corresponding one-sided confidence interval?

$95\% = 100(1 - \alpha)$, so $\alpha = .05$.

$100(1 - \alpha/2) = 97.5\%$ confidence that $error(h) \leq .3 + 0.14 = .44$.

Exercise: Suppose h commits 10 errors over 50 examples.

(a) What is 90% confidence interval (two-sided) for the true error rate?

(b) What is the 95% one-sided interval (i.e., what is the upper bound U such that $error(h) \leq U$ with 95% confidence)?

(c) What is U for the 90% one-sided interval?

Difference in error of two hypotheses

- Given h_1 and h_2 for some discrete-valued target function, and S_1 and S_2 (with n_1 and n_2 examples, respectively) both drawn from distribution D .
- We have $error_{S_1}(h_1)$ and $error_{S_2}(h_2)$.
- We want to estimate

$$d = error_D(h_1) - error_D(h_2).$$

with a confidence interval.

Here's what we do:

1. Define estimator \hat{d} for d :

$$\hat{d} = error_{s_1}(h) - error_{s_2}(h)$$

2. Figure out what probability distribution governs \hat{d} :

For large n , both $error_{s_1}(h_1)$ and $error_{s_2}(h_2)$ approximately follow a Normal distribution. Thus \hat{d} approximately follows a Normal distribution with mean d .

3. What is the variance of \hat{d} ?

$$\begin{aligned} Var[\hat{d}] &= Var[error_{s_1}(h_1)] + Var[error_{s_2}(h_2)] \\ &\approx \frac{error_{s_1}(h_1)(1 - error_{s_1}(h_1))}{n_1} + \frac{error_{s_2}(h_2)(1 - error_{s_2}(h_2))}{n_2} \end{aligned}$$

4. Now give expression for approximate N % confidence intervals:

$$\hat{d} \pm z_N \sqrt{\frac{error_{s_1}(h_1)(1 - error_{s_1}(h_1))}{n_1} + \frac{error_{s_2}(h_2)(1 - error_{s_2}(h_2))}{n_2}}$$

Example: Let S_1 be our 72 test examples, and S_2 be an independent set of 45 examples. Suppose h_1 misclassifies 12 on S_1 and h_2 misclassifies 5 on S_2 .

Based on these observations, what is our estimate \hat{d} of the true difference d in error rate between h_1 and h_2 ?

With what level of confidence can we say that d is within two standard deviations of \hat{d} ?

Comparing Learning Algorithms

- Let L_1 and L_2 be two learning algorithms.
- How to determine if difference in performance of L_1 and L_2 is statistically significant?
- Many approaches to this. Reading describes one method: **paired t test** .

Defining “difference in performance”

- First, what does “difference in performance” mean?

Let

$$d = error_D(L_1(S)) - error_D(L_2(S))$$

where $L(S)$ is the hypothesis output by L on training set S .

We want the expected difference in performance over all training sets S of size n drawn from distribution D :

$$E[d]$$

S drawn from D

Estimating expected difference in performance

- In practice, can't get all samples S of size n over D . So need an estimator.
- Suppose we have sample D_0 of data, and divide it into S_0 (training set) and T_0 (test set). Then our estimator for the difference in performance between L_1 and L_2 is:

$$\hat{d} = error_{T_0}(L_1(S_0)) - error_{T_0}(L_2(S_0))$$

Improving estimator using k -fold cross-validation

1. Partition data D_0 into k disjoint test sets T_1, T_2, \dots, T_k of equal size, where this size is at least 30.

2. For i from 1 to k , do

use T_i for the test set, and the remaining data for training set S_i

- $S_i \leftarrow \{D_0 - T_i\}$
- $h_A \leftarrow L_A(S_i)$
- $h_B \leftarrow L_B(S_i)$
- $\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$

3. Return the value $\bar{\delta}$, where

$$\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^k \delta_i$$

Comparing Learning Algorithms (revisited)

We want to estimate

$$d = error_D(L_1(S)) - error_D(L_2(S))$$

I.e., we want to estimate

$$E[d]$$

S drawn from D

So we do k -fold cross-validation. For each iteration i , we get

$$\delta_i = error_{T_i}(h_1) - error_{T_i}(h_2)$$

and finally: $\bar{\delta} = \frac{1}{k} \sum_{i=1}^k \delta_i$

$\bar{\delta}$ is our estimate of $E[d]$
 S drawn from D

Note that $\bar{\delta}$ is the mean of samples from random variables δ_i .

$\bar{\delta}$ is called a “sample mean”. It is an unbiased estimator of the population mean $E[d]$
 S drawn from D

The sample standard deviation is:

$$\sigma_{sample} = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (\delta_i - \bar{\delta})^2}$$

But the true standard deviation of the sample mean, $\sigma_{\bar{\delta}}$, is unknown.

It is estimated by

$$S_{\bar{\delta}} = \frac{\sigma_{sample}}{\sqrt{n}}$$

This is also called “standard error”. It estimates the standard deviation of the distribution of samples. It reflects how much sampling fluctuation σ_{sample} has.

So we have:

$$S_{\bar{\delta}} = \frac{\sigma_{sample}}{\sqrt{k}} = \frac{\sqrt{\frac{1}{(k-1)} \sum_{i=1}^k (\delta_i - \bar{\delta})^2}}{\sqrt{k}} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^k (\delta_i - \bar{\delta})^2}$$

Approximate N % confidence interval:

$$\bar{\delta} \pm t_{N,k-1} s_{\bar{\delta}}$$

Here $t_{N,k-1}$ is a constant that corresponds to z_N .

When $k \rightarrow \infty$, $t_{N,k-1} \rightarrow z_N$.

Determining this confidence interval is called the “paired t test”.

$t_{N,k}$ table

Confidence level N

	90%	95%	98%	99%
$k = 2$	2.92	4.30	6.96	9.92
$k = 5$	2.02	2.57	3.36	4.03
$k = 10$	1.81	2.23	2.76	3.17
$k = 20$	1.72	2.09	2.53	2.84
$k = 30$	1.70	2.04	2.46	2.75
$k = \infty$	1.64	1.96	2.33	2.58

Summary

1. Assessing the error of a single hypothesis h

A. Two-sided confidence interval:

$$\begin{aligned} & error_s(h) \pm z_N \sigma_s \\ & = error_s(h) \pm z_N \sqrt{\frac{error_s(h)(1 - error_s(h))}{n}} \end{aligned}$$

“With $N\%$ confidence, true error is in this interval”

B. One-sided confidence interval:

If the two-sided confidence interval is for $N = 100(1 - \alpha)\%$ with lower bound L and upper bound U , this implies a one-sided confidence interval of $N_U = 100(1 - \alpha/2)\%$.

“With $N_U\%$ confidence, the true error is smaller than U .”

2. Comparing the error of two hypothesis, h_1 and h_2

A. Confidence interval around observed difference in error rate:

N% confidence interval around

$$\hat{d} = error_{s_1}(h_1) - error_{s_2}(h_2)$$

is

$$\hat{d} \pm z_N \sqrt{\frac{error_{s_1}(h_1)(1 - error_{s_1}(h_1))}{n_1} + \frac{error_{s_2}(h_2)(1 - error_{s_2}(h_2))}{n_2}}$$

“With confidence N%, the true difference in error rate d is in this interval.

B. Hypothesis testing (corrected)

Given $\hat{d} = error_{s_1}(h_1) - error_{s_2}(h_2)$

what is the probability that $d > 0$?

$$\begin{aligned} P(d > 0) &= P(\hat{d} < d + \hat{d}) \\ &= P\left(\hat{d} < \mu_{\hat{d}} + \frac{\hat{d}}{\sigma_{\hat{d}}} \sigma_{\hat{d}}\right) \end{aligned}$$

Look up $\frac{\hat{d}}{\sigma_{\hat{d}}}$ in z_n table to get N .

Find associated one - sided confidence interval N_U .

“With confidence N_U %, we accept the hypothesis that $error(h_1) > error(h_2)$.”

3. Comparing two learning algorithms, L_A and L_B

k -fold cross-validation paired t -test:

$$\bar{d} = \frac{1}{k} \sum_{i=1}^k \delta_i$$

where $\delta_i = \text{error}_{T_i}(h_A^i) - \text{error}_{T_i}(h_B^i)$.

We have confidence interval $\bar{d} \pm t_{N,k-1} s_{\bar{d}}$

where

$$s_{\bar{d}} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^k (\delta_i - \bar{d})^2}$$