Evaluating Hypotheses

[Read Ch. 5]
[Recommended exercises: 5.2, 5.3, 5.4]

- Sample error, true error
- Confidence intervals for observed hypothesis error
- Estimators
- Binomial distribution, Normal distribution, Central Limit Theorem
- Paired t tests
- Comparing learning methods

Two Definitions of Error

The **true error** of hypothesis h with respect to target function f and distribution $\mathcal{D}$ is the probability that h will misclassify an instance drawn at random according to $\mathcal{D}$.

$$error_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[f(x) \neq h(x)]$$

The **sample error** of h with respect to target function f and data sample S is the proportion of examples h misclassifies

$$error_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x))$$

Where $\delta(f(x) \neq h(x))$ is 1 if $f(x) \neq h(x)$, and 0 otherwise.

How well does $error_S(h)$ estimate $error_{\mathcal{D}}(h)$?

Problems Estimating Error

1. *Bias:* If S is training set, $error_S(h)$ is optimistically biased

$$bias \equiv E[error_S(h)] - error_{\mathcal{D}}(h)$$

   For unbiased estimate, h and S must be chosen independently

2. *Variance:* Even with unbiased S, $error_S(h)$ may still *vary* from $error_{\mathcal{D}}(h)$. The smaller the test-set, the larger the probability of a large variance.

Example

Hypothesis h misclassifies 12 of the 40 examples in S

$$error_S(h) = \frac{12}{40} = .30$$

What is $error_{\mathcal{D}}(h)$?

Estimators

Experiment:

1. choose sample S of size n according to distribution $\mathcal{D}$

2. measure $error_S(h)$

$error_S(h)$ is a random variable (i.e., result of an experiment)

$error_S(h)$ is an unbiased *estimator* for $error_{\mathcal{D}}(h)$

Given observed $error_S(h)$ what can we conclude about $error_{\mathcal{D}}(h)$?

Confidence Intervals

- **IF** S contains n examples, drawn independently of h and each other
- $n \geq 30$
- **THEN** With approximately 95% probability, $error_{\mathcal{D}}(h)$ lies in interval

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

Confidence Intervals

- **IF** S contains n examples, drawn independently of h and each other
- $n \geq 30$
- **THEN** with approximately N% probability, $error_{\mathcal{D}}(h)$ lies in interval

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

WHERE

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

- at least 30 examples
- $error_S(h)$ not too close to 0 or 1
- or

$$n \times error_S(h) \times (1 - error_S(h)) \geq 5$$

Example

data sample S, n $= 40$

$r = 12$, number of error h commit over S

i.e. $error_S(h) = \frac{12}{40} = 0.3$

95% confidence interval estimate for

$$error_D(h) \in [0.3 \pm (1.96 \times \sqrt{\frac{0.3*0.7}{40}})]$$

$$error_D(h) \in [0.3 \pm 0.14]$$

Example 2

Same example, different confidence interval

data sample S, n = 40

$r = 12$, number of error h commit over S

i.e. $error_S(h) = \frac{12}{40} = 0.3$

98% confidence interval estimate for

$$error_D(h) \in [0.3 \pm (2.33 \times \sqrt{\frac{0.3*0.7}{40}})]$$

$$error_D(h) \in [0.3 \pm 0.1631]$$

Example 3

Same example, different sample size and error

data sample S, n = 1000

$r = 300$, number of error h commit over S

i.e. $error_S(h) = \frac{300}{1000} = 0.3$

95% confidence interval estimate for

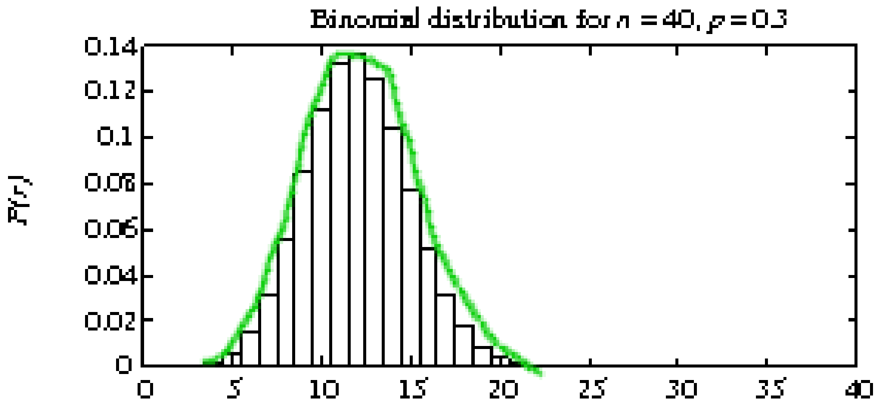
$$error_D(h) \in [0.3 \pm (1.96 \times \sqrt{\frac{0.3*0.7}{1000}})]$$

$$error_D(h) \in [0.3 \pm 0.028403098]$$

$error_S(h)$ is a Random Variable

Rerun the experiment with different randomly drawn S (of size n)

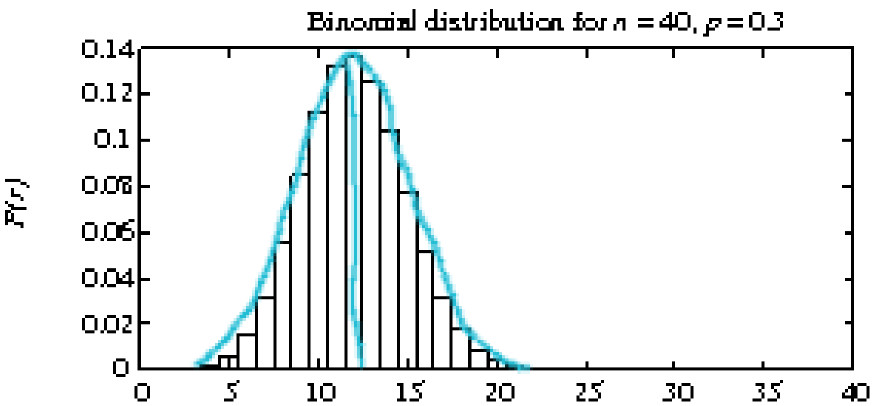
Probability of observing r misclassified examples:



$$P(r) = \frac{n!}{r!(n-r)!}\ error_{\mathcal{D}}(h)^r (1 - error_{\mathcal{D}}(h))^{n-r}$$

$$error_{\mathcal{D}}(h) = P$$

Binomial Probability Distribution



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

Probability $P(r)$ of r heads in n coin flips, if $p = \mathrm{Pr}(heads)$

- Expected, or mean value of X, $E[X]$, is

$$E[X] \equiv \sum_{i=0}^{n} iP(i) = np$$

- <u>Variance</u> of X is

$$Var(X) \equiv E[(X - E[X])^2] = np(1-p)$$

- <u>Standard deviation</u> of X, σ_X, is

$$\sigma_X \equiv \sqrt{E[(X - E[X])^2]} = \sqrt{\frac{np(1-p)}{n^2}}$$

Normal Distribution Approximates Binomial

$error_S(h)$ follows a *Binomial* distribution, with

- mean $\mu_{error_S(h)} = error_{\mathcal{D}}(h)$
- standard deviation $\sigma_{error_S(h)}$

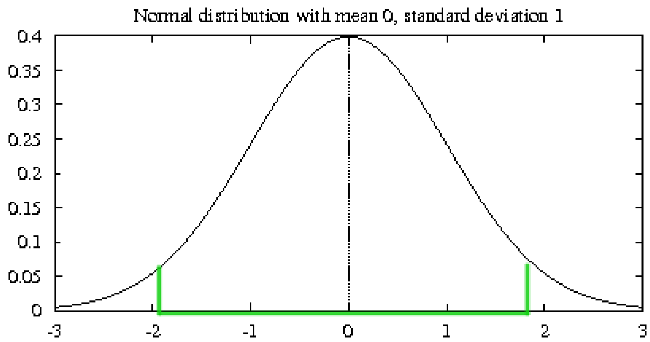
$$\sigma_{error_S(h)} = \sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

Approximate this by a *Normal* distribution with

- mean $\mu_{error_S(h)} = error_{\mathcal{D}}(h)$
- standard deviation $\sigma_{error_S(h)}$

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

Normal Probability Distribution



$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

The probability that X will fall into the interval (a, b) is given by

$$\int_a^b p(x)dx$$

- Expected, or mean value of X, $E[X]$, is

$$E[X] = \mu$$

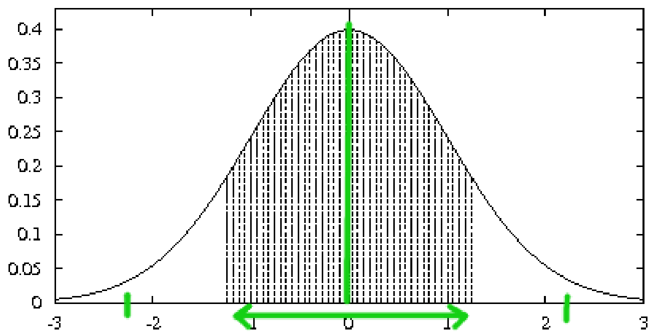
- Variance of X is

$$Var(X) = \sigma^2$$

- Standard deviation of X, σ_X, is

$$\sigma_X = \sigma$$

Normal Probability 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

Confidence Intervals, More Correctly

- **IF** S contains n examples, drawn independently of h and each other
- $n \geq 30$
- **THEN** with approximately 95% probability, $error_S(h)$ lies in interval

$$error_{\mathcal{D}}(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

equivalently, $error_{\mathcal{D}}(h)$ lies in interval

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

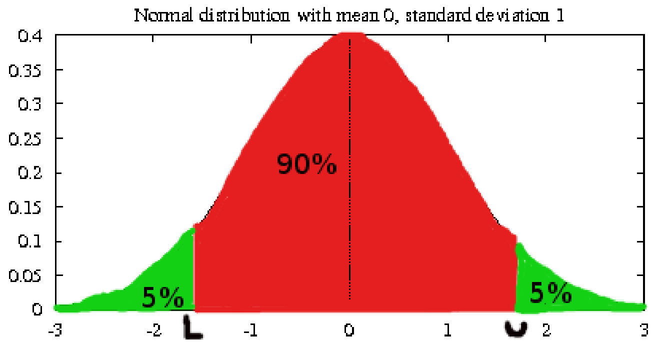
which is approximately

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

A General Approach For Calculating Confidence Intervals

1. Pick parameter p to estimate
   - $error_{\mathcal{D}}(h)$
2. Choose an estimator (unbiased, low variance)
   - $error_S(h)$
3. Determine probability distribution that governs estimator
   - $error_S(h)$ governed by Binomial distribution, approximated by Normal when $n \geq 30$
4. Find interval (L, U) such that N% of probability mass falls in the interval
   - Use table of z_N values

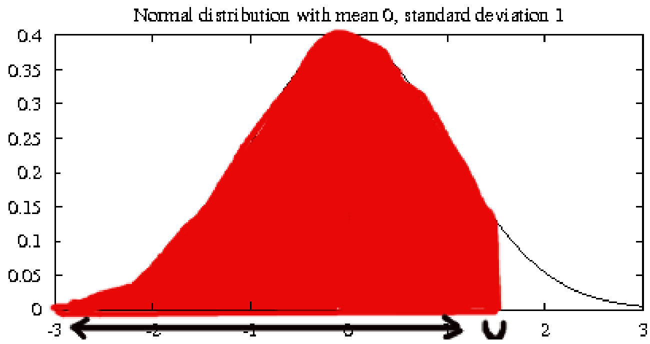
Two-sided bounds and One-sided bounds



$$P(x \in [L, U]) = N\%$$

$$P(x \notin [L, U]) = (100 \ - N)\%$$

$$P(x \leq U]) = (N + \frac{100-N}{2})\%$$



An example

$$error_S(h) = 0.3, n = 40$$

$$?u \text{ such that } p(x \leq u) = 97.5\%$$

$$\text{or } N \text{ such that } N + \frac{100-N}{2} = 97.5$$

$$\rightarrow N = 95$$

$$u = 0.30 \pm Z_{95}\sqrt{\frac{(0.3)(1-0.3)}{40}}, Z_n = 1.96$$

$$u = 0.30 + 0.14 = 0.44$$

Different Hypotheses, Which One Is The Best?

Test h_1 on sample S_1, test h_2 on S_2

1. Pick parameter to estimate

$$d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2) = \text{true error}$$

2. Choose an estimator (unbiased)

$$\hat{d} \equiv error_{S_1}(h_1) - error_{S_2}(h_2)$$

3. Determine probability distribution that governs estimator

$$\sigma_{\hat{d}} \approx \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}}$$

4. Find interval (L, U) such that N% of probability mass falls in the interval

$$\hat{d} \pm z_N \sqrt{\frac{error_{S_1}(h_1)(1 - error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1 - error_{S_2}(h_2))}{n_2}}$$

Example

$$h_1, S_1, n1 = 100$$

Thus, $error_{S1}(h1) = 0.3$

AND

$$h_2, S_2, n2 = 100$$

Thus, $error_{S2}(h2) = 0.2$ Given $\hat{\delta} = 0.1$

Is $error_D(h1) > error_D(h2)$?

or if $d = error_D(h1) - error_D(h2)$

What is the probability that $d > 0$, given we observed $\hat{d} = 0.1$

probability $\hat{d} < d + 0.1$

probability $\widehat{d} \leftarrow$ one sided interval

$\mu_{\hat{d}} + z_N \sigma_{\hat{d}}$ with $\sigma_{\hat{d}} = 0.061$ (see Eq. 5.12)

! Z_N such that $0.1 = Z_N 0.061$

$Z_N \approx 1.64$

Thus, two-sided confidence level = 90%

one-sided confidence level = $90\% + \frac{100\% - 90\%}{2} = 95\%$

Paired t test to compare h_A,h_B

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

2. For i from 1 to k, do

$$\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$$

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

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

N% confidence interval estimate for d:

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

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

Note δ_i approximately Normally distributed

Comparing learning algorithms L_A and L_B

What we'd like to estimate:

$$E_{S \subset \mathcal{D}}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

where $L(S)$ is the hypothesis output by learner L using training set S

i.e., the expected difference in true error between hypotheses output by learners L_A and L_B, when trained using randomly selected training sets S drawn according to distribution $\mathcal{D}$.

But, given limited data D_0, what is a good estimator?

- could partition D_0 into training set S and test set T_0, and measure

$$error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0))$$

- even better, repeat this many times and average the results (next slide)

Comparing learning algorithms L_A and L_B

(a) Partition data D_0 into k disjoint test sets $T_1, T_2, \ldots, T_k$ of equal size, where this size is at least 30.

(b) For i from 1 to k, do

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

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

(c) Return the value $\bar{\delta}$, where

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

(d) $\bar{\delta}$ is an estimator of $E_{S \subset D_o}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$

Comparing learning algorithms L_A and L_B

Notice we'd like to use the paired t test on $\bar{\delta}$ to obtain a confidence interval

but not really correct, because the training sets in this algorithm are not independent (they overlap!)

more correct to view algorithm as producing an estimate of

$$E_{S \subset D_0}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

instead of

$$E_{S \subset \mathcal{D}}[error_{\mathcal{D}}(L_A(S)) - error_{\mathcal{D}}(L_B(S))]$$

but even this approximation is better than no comparison

Confidence levels

	Confidence level N			
	90%	95%	98%	99%
$\nu = 2$	2.92	4.30	6.96	9.92
$\nu = 5$	2.02	2.57	3.36	4.03
$\nu = 10$	1.81	2.23	2.76	3.17
$\nu = 20$	1.72	2.09	2.53	2.84
$\nu = 30$	1.70	2.04	2.46	2.75
$\nu = 120$	1.66	1.98	2.36	2.62
$\nu = \infty$	1.64	1.96	2.33	2.58

TABLE 5.6
Values of $t_{N,\nu}$ for two-sided confidence intervals.

As $\nu \to \infty$, $t_{N,\nu}$ approaches z_N.

Summary

- p = probability coin
- $p(toss) = p(head)$
- r = number of heads over sample of size n
- $\rightarrow \frac{1}{n}$
- Estimating p

- $error_{\mathcal{D}}(h)$ = probability h misclassifies random instance
- ratio of misclassifications by h over n random instances
- r = number of heads over sample of size n
- $\rightarrow error_S(h)$
- Estimating $error_{\mathcal{D}}(h)$

Confidence Interval

- Describes the uncertainty associated with an estimate

- It is the interval within which the true value is expected to fall with a certain probability

An example

- h tested on 40 samples of S and $r = 12$ errors

- approx. prob of 95%

- $error_{\mathcal{D}}(h) \in 0.3 \pm 0.14$

- Why, how to compute interval?

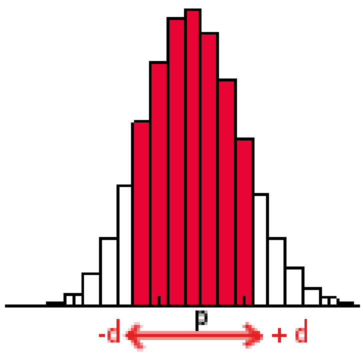
- We know $error_{\mathcal{D}}(h)$ random variable according to Binomial probability distribution

- SD mean $= error_{\mathcal{D}}(h) = P$

- SD mean $= \sqrt{\frac{P(1-P)}{n}}$

- $P(error_S(h) \in H) = 90\%$

- $P(error_{\mathcal{D}}(h) \in error_S(h) \pm d) = 90\%$



Exercises

EXERCISFS

5.1. Sur po:c you test a hypothesis h and find that it commits $r = 300$ errors on a sample S of $n =$ 1CCO ranuomly urawn test examples. What is the standard deviation in $error_S(h)$? How does this compare to the standard deviation in the example at the end of Section 5._.4?

5.2. Consider a learned hypothesis, h, for some boolean concept. When h is tested on a set of 100 examples, it classifies 83 correctly. What is the standard deviation and the 95% confidence interval for the true error rate for $Error_{\mathcal{D}}(h)$?

5.3. Suppose hypothesis h commits $r = 10$ errors over a sample of $n = 65$ independently drawn examples. What is the 90% confidence interval (two-sided) for the true error rate? What is the 95% one-sided interval (i.e., what is the upper bound U such that $error_{\mathcal{D}}(h) \leq U$ with 95% confidence)? What is the 90% one-sided interval?

5.4. You are about to test a hypothesis h whose $error_{\mathcal{D}}(h)$ is known to be in the range between 0.2 and 0.6. What is the minimum number of examples you must collect to assure that the width of the two-sided 95% confidence interval will be smaller than 0.1?

5.5. Give genera expressions for the upper and lower one-sided $N\%$ confidence intervals for the difference in errors between two hypotheses tested on different samples of data. Hint: modify the expression given in Section 5.5.

5.6. Explain wh the confidence interval estimate given in Equation (5.17) applies to estimating the quantity in Equation (5.16), and not the quantity in Equation (5.14).