Outline: Evaluating Hypotheses

[Read Ch. 5] Tom Mitchell: Machine Learning

[Recommended exercises: 5.2, 5.3, 5.4]

Error of a Sample versus true error

Confidence intervals for observed hypothesis error

Estimators

Binomial distribution, Normal distribution, Central Limit Theorem

Paired t tests

Comparing learning methods

Sampling

Consider learning the target function "people who plan to purchase new skis this year," given a sample of training data collected by surveying people as they arrive at a ski resort.

instance space X == space of all people (denoted by x). Each individual x may be described by features like e.g.: age, occupation, how many times they skied last year, etc.

Some (unknown) distribution D specifies for each person x the probability that x will be encountered as the next person arriving at the ski resort.

The target function *f: X -> {0,1}* classifies each x according to whether or not they plan to purchase skis this year.

We are interested in following two questions:

1. Given a hypothesis h and a data sample S containing n examples drawn at random according to the distribution D, what is the best estimate of the accuracy of h over future instances drawn from the same distribution?

2. What is the probable error in this accuracy estimate?

Two Definitions of Error

The true error of hypothesis h with respect to target function f and distribution D is the probability that h will misclassify an instance drawn at random according to 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 that h misclassifies:

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

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

Q: How well does $error_S(h)$ estimate $error_D(h)$?

Estimate a binomial p **IS SAME** as estimate the $error_D(h)$

Estimate p from random sample of coin tosses	Estimate $error_D(h)$ from random sample of instances
single toss	draw single random instance i from D.
probability p that a coin toss yields head up	i is misclassified by h
r heads over n tosses = r/n	$error_S(h)$

General setting for Binom. distr.

1. There is a base, or underlying, experiment (e.g., toss of the coin) whose outcome can be described by a random variable, say Y. The random variable Y can take on only two possible values (e.g., $Y = 1$ if heads, $Y = 0$ if tails).
2. The probability that $Y = 1$ on any single trial of the underlying experiment is given by some constant p, independent of the outcome of any other experiment. The probability that $Y = 0$ is therefore $(1 - p)$. Typically, p is not known in advance, and the problem is to estimate it.
3. A series of n independent trials of the underlying experiment is performed (e.g., n independent coin tosses), producing the sequence of independent, identically distributed random variables $Y_1, Y_2, \ldots, Y_n$. Let R denote the number of trials for which $Y_i = 1$ in this series of n experiments:
$$R \equiv \sum_{i=1}^{n} Y_i$$
4. The probability that the random variable R will take on a specific value r (e.g., the probability of observing exactly r heads) is given by the Binomial distribution
$$\Pr(R = r) = \binom{n}{r} p^r (1-p)^{n-r}$$

   A plot of this probability distribution is shown on slide 11.

Problems when Estimating Error

1. Bias:
   If ***s*** is training set, $error_S(h)$ is optimistically biased
   $$bias = E[error_S(h)] - error_D(h)$$
   For unbiased estimate, ***h*** and ***S*** must be chosen independently
2. Variance:
   Even with unbiased ***S***, $error_S(h)$ may still vary from $error_D(h)$

Example

Hypothesis h misclassifies 12 of the 40 examples in *S*

$$error_S(h) = \frac{12}{40} = 0.30$$

What is $error_D(h)$?

Estimators

Experiment:

1. *choose sample* S *of size* n *according to distribution* D
2. *measure* $\text{error}_S(h)$

$error_S(h)$ is a random variable (why?)

(i.e., the result of a random experiment)

$error_S(h)$ is an unbiased estimator for $error_D(h)$

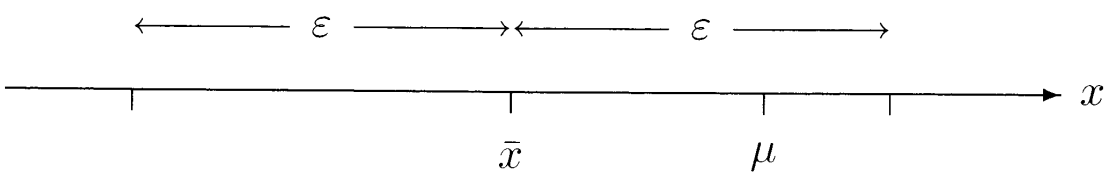
given the observed $error_S(h)$ what can we conclude about $error_D(h)$?

Confidence Intervals

If

S contains *n* examples, drawn independently of *h* and each other,

and *n* >= *30*



then

with approximately 95% probability,

$error_D(h)$ 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, and *n* >= *30*

then

with approximately N% probability, $error_D(h)$ lies in interval

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

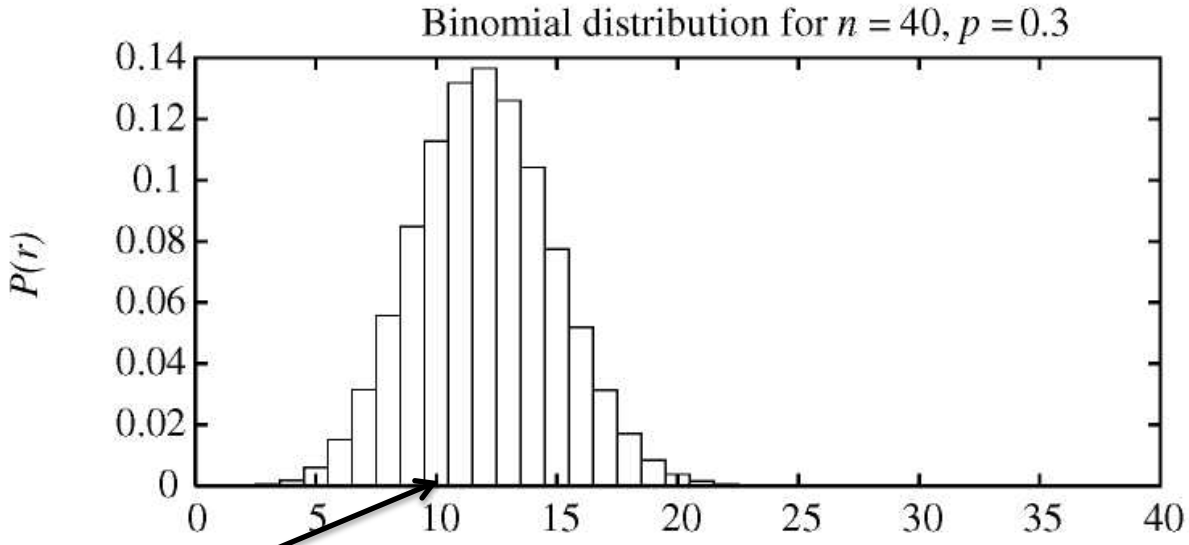
(Table 5.1)

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

$error_S(h)$ is a (binomial) Random Variable

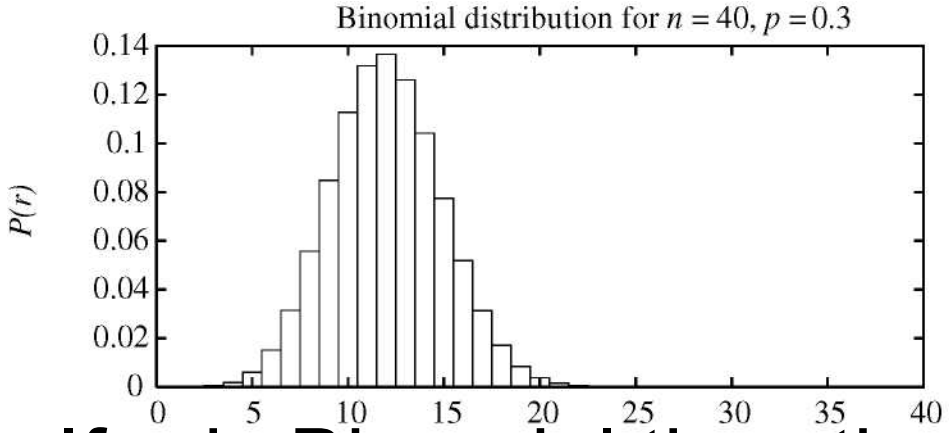
Rerun the experiment with different randomly drawn S (of size n)

Probability of observing r misclassified examples:



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

Binomial Probability Distribution



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

If X is Binomial then the probability $Pr(X=r)$ of r heads in n coin flips, is given by $P(r)$

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

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

Variance of X is :

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

Standard deviation of X, σ_X, is :

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

Now: Normal Distribution approximates Binomial distr. !!!

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

$\mu_{errorS}(h) = error_D(h)$

and standard deviation:

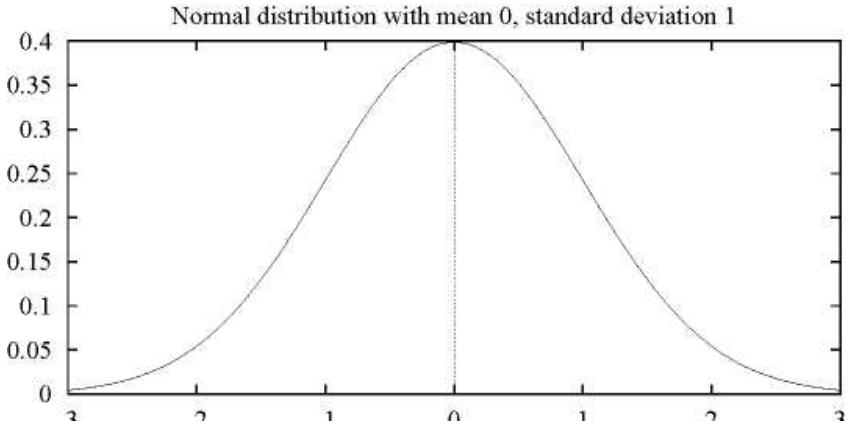
$$\sigma_{error_S(h)} = \sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

approximate this by a **Normal distribution** with estimated mean and variance:

$\mu_{errorS}(h)$

and estimated standard deviation:

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



Normal Probability Distribution

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

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

=> $E[X] = \mu$

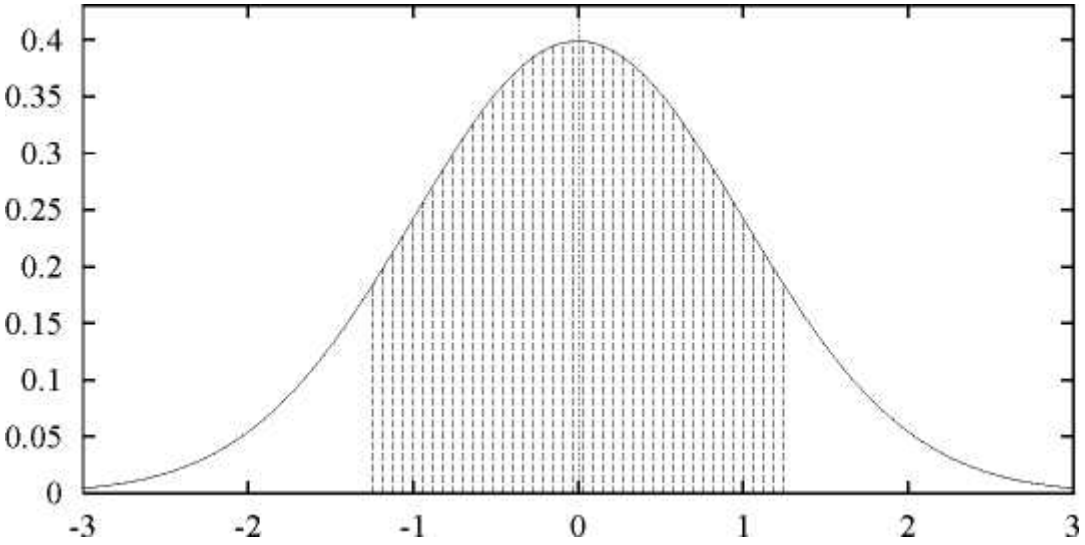
Variance of *X* is

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

Standard deviation of *X*, σ_X, is

$\Rightarrow \sigma_X = \sigma$

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
and *n* > *30*

then

with approximately 95% prob., $error_S(h)$ lies in the interval

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

equivalently, $error_D(h)$ lies in interval

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

which is approximately

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

Central Limit Theorem

Consider a set of independent, identically distributed random variables $Y_1,..,Y_n$ all governed by an arbitrary probability distribution with mean μ and finite variance σ^2. Define the sample mean, $\bar{Y} \equiv \frac{1}{n} \sum_{i=1}^{n} Y_i$

As $n \rightarrow oo$, the distribution governing $\bar{Y}$ approaches a Normal distribution, with mean μ and variance σ^2/n.

Calculating Confidence Intervals

1. Pick parameter p to estimate:
   e.g. $error_D(h)$
2. Choose an estimator:
   e.g. $error_S(h)$
3. Determine probability distribution that governs the estimator:
   $error_S(h)$ is governed by Binomial Distribution, which in turn is approximated by a Normal Distribution (if n >= 30)
4. Find the interval (θ_L, θ_U), such that $N\%$ of probability mass falls in this interval:
   i.e. use table of z_N values

Difference Between Hypotheses

Test h_1 on sample S_1, test h_2 on S_2 (Section 5.5)

1. Pick parameter to estimate

$$d \equiv error_D(h_1) - error_D(h_2)$$

2. Choose an estimator

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

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

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

Paired *t* test to compare h_A, h_B

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

For i from 1 to k, do

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

Return the value δ, where

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

N% confidence interval estimate for d:

$$\bar{\delta} \pm t_{N,k-1}\, s_{\bar{\delta}} \qquad (5.17)$$

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

Note: δ_i are approximately Normally distributed

	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

Comparing learning algorithms L_A and L_B

What we'd like to estimate:

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

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 D.

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

We 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

1. 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.
2. For i from 1 to k, do

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

   $S_i \leftarrow \{D_0 - T_i\}$

   $h_A \leftarrow L_A(S_i)$

   $h_B \leftarrow L_B(S_i)$

   $\delta_i \leftarrow error_{Ti}(h_A) - error_{Ti}(h_B)$
3. Return the value $\overline{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$

Comparing learning algorithms L_A and L_B

Notice we'd like to use the paired t test on δ 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

$$E_{S \subset D_0}[error_D(L_A(S)) - error_D(L_B(S))] \quad (5.16)$$

instead of

$$E_{S \subset D}[error_D(L_A(S)) - error_D(L_B(S))] \quad (5.14)$$

but even this approximation is better than no comparison

Summary

Statistical theory provides a basis for estimating the true error ($error_D(h)$) of a hypothesis h, based on its observed error ($error_S(h)$) over a sample S of data. For example, if A is a discrete-valued hypothesis and the data sample S contains n > 30 examples drawn independently of h and of one another, then the N% confidence interval for $error_D(h)$ is approximately

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

where values for z_N are given in Table 5.1.

In general, the problem of estimating confidence intervals is approached by identifying the parameter to be estimated (e.g., $error_D(h)$) and an estimator (e.g., $error_S(h)$) for this quantity. Because the estimator is a random variable (e.g., $error_S(h)$ depends on the random sample S, it can be characterized by the probability distribution that governs its value. Confidence intervals can then be calculated by determining the interval that contains the desired probability mass under this distribution.

One possible cause of errors in estimating hypothesis accuracy is estimation bias. If Y is an estimator for some parameter p, the estimation bias of Y is the difference between p and the expected value of Y. For example, if S is the training data used to formulate hypothesis h, then $error_S(h)$ gives an optimistically biased estimate of the true error $error_D(h)$

Summary

A second cause of estimation error is variance in the estimate. Even with an unbiased estimator, the observed value of the estimator is likely to vary from one experiment to another. The variance σ^2 of the distribution governing the estimator characterizes how widely this estimate is likely to vary from the correct value. This variance decreases as the size of the data sample is increased.

Comparing the relative effectiveness of two learning algorithms is an estimation problem that is relatively easy when data and time are unlimited, but more difficult when these resources are limited. One possible approach described in this chapter is to run the learning algorithms on different subsets of the available data, testing the learned hypotheses on the remaining data, then averaging the results of these experiments.

In most cases considered here, deriving confidence intervals involves making a number of assumptions and approximations. For example, the above confidence interval for $error_D(h)$ involved approximating a Binomial distribution by a Normal distribution, approximating the variance of this distribution, and assuming instances are generated by a fixed, unchanging probability distribution. While intervals based on such approximations are only approximate confidence intervals, they nevertheless provide useful guidance for designing and interpreting experimental results in machine learning.

EXERCISES (5 out of 8)

1. Suppose you test a hypothesis h and find that it commits $r = 300$ errors on a sample S of $n = 1000$ randomly drawn test examples. What is the standard deviation is $error_S(h)$? How does this compare to the standard deviation in the example at the end of Section 5.3.4?
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 $Error_D(h)$?
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 $error_D(h) <= U$ with 95% confidence)? What is the 90% one-sided interval?
4. You are about to test a hypothesis h whose $error_D(h)$ 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. Give general 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.
6. Explain why 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).
7. proof E[X] for binomial == np
8. Prove: for binomial Var(X) = np(1-p)