

# Evaluating Hypotheses

---

[Read Ch. 5]

- 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  $error_{\mathcal{D}}(h)$

# 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

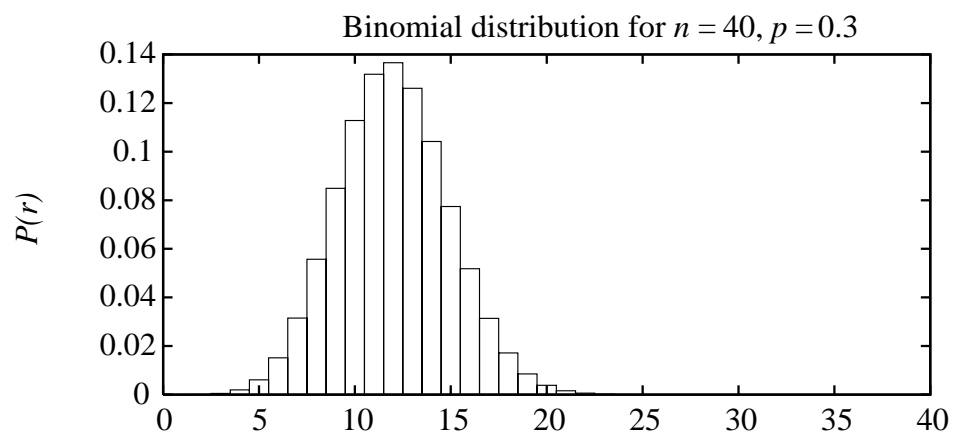
$N\%$ :	50%	68%	80%	90%	95%	98%	99%
$z_N$ :	0.67	1.00	1.28	1.64	1.96	2.33	2.58

# $error_S(h)$ is a Random Variable

---

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

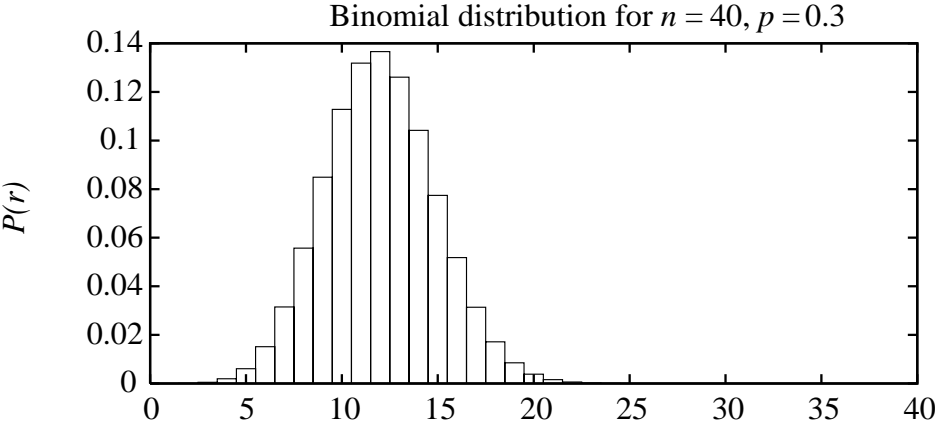
Probability of observing  $r$  misclassified examples:



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

# 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  $p = \Pr(\text{heads})$

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

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

- Variance of  $X$  is

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

- Standard deviation of  $X$ ,  $\sigma_X$ , is

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

# Normal Distribution Approximates Binomial

---

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

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

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

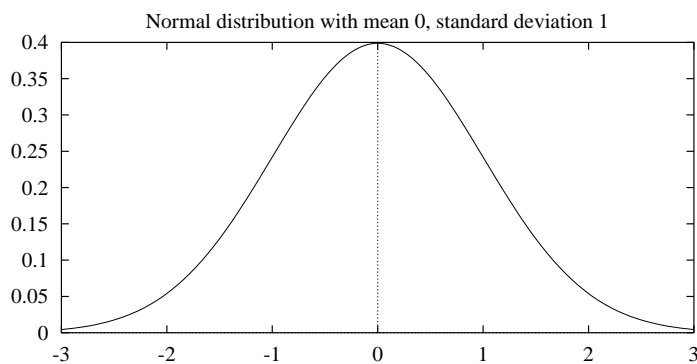
Approximate this by a *Normal* distribution with

- 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

---



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

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

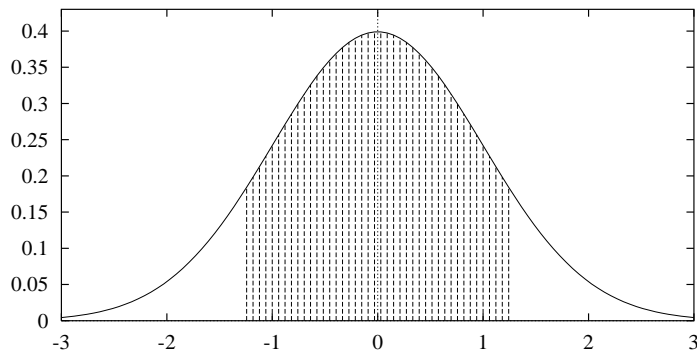
$$Var(X) = \sigma^2$$

- Standard deviation of  $X$ ,  $\sigma_X$ , 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

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

# Central Limit Theorem

---

Consider a set of independent, identically distributed random variables  $Y_1 \dots Y_n$ , all governed by an arbitrary probability distribution with mean  $\mu$  and finite variance  $\sigma^2$ . Define the sample mean,

$$\bar{Y} \equiv \frac{1}{n} \sum_{i=1}^n Y_i$$

**Central Limit Theorem.** As  $n \rightarrow \infty$ , the distribution governing  $\bar{Y}$  approaches a Normal distribution, with mean  $\mu$  and variance  $\frac{\sigma^2}{n}$ .

# Calculating Confidence Intervals

---

1. Pick parameter  $p$  to estimate
  - $error_{\mathcal{D}}(h)$
2. Choose an estimator
  - $error_S(h)$
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

# Difference Between Hypotheses

---

Test  $h_1$  on sample  $S_1$ , test  $h_2$  on  $S_2$

1. Pick parameter to estimate

$$d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$$

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 $h_A, h_B$

---

1. Partition data 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

$$\delta_i \leftarrow \text{error}_{T_i}(h_A) - \text{error}_{T_i}(h_B)$$

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

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

$N\%$  confidence interval estimate for  $d$ :

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

*Note  $\delta_i$  approximately Normally distributed*

# Comparing learning algorithms $L_A$ and $L_B$

---

What we'd like to estimate:

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

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 training set  $T_0$ , and measure

$$\text{error}_{T_0}(L_A(S_0)) - \text{error}_{T_0}(L_B(S_0))$$

- 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  $T_1, T_2, \dots, T_k$  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 test set  $S_i$*

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

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

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

# Comparing learning algorithms $L_A$ 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

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

instead of

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

but even this approximation is better than no comparison