

Evaluating Hypotheses

Based on “Machine Learning”, T. Mitchell, McGRAW Hill, 1997, ch. 5

Acknowledgement:

The present slides are an adaptation of slides drawn by T. Mitchell

Main Questions in Evaluating Hypotheses

1. How can we **estimate the accuracy of a learned hypothesis** h over the whole space of instances \mathcal{D} , given its observed accuracy over limited data?
2. How can we estimate the probability that a hypothesis h_1 performs is more accurate than another hypothesis h_2 over \mathcal{D} ?
3. If available data is limited, how can we use this data for both training and **comparing the relative accuracy** of two learned hypothesis?

Statistics Prespective

(See Appendix for ◦ Details)

Problem: Given a property observed over some random sample \mathcal{D} of the population X , estimate the proportion of X that exhibits that property.

- Sample error, true error
- Estimators
 - Binomial distribution, Normal distribution
 - Confidence intervals
 - Paired t tests

1. Two Definitions of Error

The **sample error** of hypothesis h with respect to the 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.

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

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

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

Problems in Estimating $error_{\mathcal{D}}(h)$

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

1. If S is training set, then $error_S(h)$ is optimistically biased, because h was learned using S . Therefore, for unbiased estimate, h and S must be chosen independently.
2. Even with unbiased S (i.e., $bias = 0$), the *variance* of $error_S(h) - error_{\mathcal{D}}(h)$ may be not null.

Calculating Confidence Intervals for $error_S(h)$: Preview/Example

Question:

If hypothesis h misclassifies 12 of the 40 examples in S , what can we conclude about $error_{\mathcal{D}}(h)$?

Answer:

If the examples are drawn independently of h and of each other,
then with approximately 95% probability, $error_{\mathcal{D}}(h)$ lies in interval $0.30 \pm (1.96 \times 0.14)$.

$$(error_S(h) = 0.30, z_N = 1.96, \text{ and } 0.14 \approx \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}})$$

Calculating Confidence Intervals for Discrete-valued Hypotheses: A general approach

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

Calculating Confidence Intervals for $error_S(h)$: Proof Idea

- we run the experiment with different randomly drawn S (of size n), therefore $error_S(h)$ is a random variable; we will use $error_S(h)$ to estimate $error_{\mathcal{D}}(h)$.
- probability of observing r misclassified examples follows the Binomial distribution:

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

- for n sufficiently large, the Normal distribution approximates the Binomial distribution (see next slide);
- $N\%$ of the area defined by the Binomial distribution lies in the interval $\mu \pm z_N \sigma$,
with μ and σ respectively the mean and the std. deviation.

Normal Distribution Approximates $error_S(h)$

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

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

- mean $\mu_{error_S(h)} = error_{\mathcal{D}}(h)$
- standard deviation $\sigma_{error_S(h)} \approx \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$

Calculating Confidence Intervals for $error_S(h)$: Full Proof Details

If

- S contains n examples, drawn independently of h and each other
- $n \geq 30$
- $error_S(h)$ is not too close to 0 or 1
(recommended: $n \times error_S(h) \times (1 - error_S(h)) \geq 5$)

then with **approximately N% probability**, $error_S(h)$ lies in the interval

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

Equivalently, $error_{\mathcal{D}}(h)$ lies in interval $error_S(h) \pm z_N \sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$

which is **approximately** $error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$

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

2. Estimate the Difference Between Two 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.
 \hat{d} is approximately Normally distributed:

$$\mu_{\hat{d}} = d$$

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

4. Find the **confidence interval** (L, U) :

N% of probability mass falls in the interval $\mu_{\hat{d}} \pm z_N \sigma_{\hat{d}}$

Difference Between Two Hypotheses: An Example

Suppose $error_{S_1}(h_1) = .30$ and $error_{S_2}(h_2) = .20$.

Question: What is the estimated probability of $error_{\mathcal{D}}(h_1) > error_{\mathcal{D}}(h_2)$?

Answer:

Notation:

$$\hat{d} = error_{S_1}(h_1) - error_{S_2}(h_2) = 0.10$$

$$d = error_{\mathcal{D}}(h_1) > error_{\mathcal{D}}(h_2)$$

Calculation:

$$P(d > 0, \hat{d} = .10) = P(\hat{d} < d + 0.10) = P(\hat{d} < \mu_{\hat{d}} + 0.10)$$

$$\sigma_{\hat{d}} = 0.061, \text{ and } 0.10 \approx 1.64 \times \sigma_{\hat{d}}$$

$$z_{90} = 1.64$$

$$\text{Conclusion: (using one-sided conf. interv.) } P(\hat{d} < \mu_{\hat{d}} + 0.10) = 95\%$$

Therefore, with a 95% confidence, $error_{\mathcal{D}}(h_1) > error_{\mathcal{D}}(h_2)$

3. Comparing learning algorithms L_A and L_B

We would like to estimate the **true error** between the output of L_A and L_B :

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

where $L(S)$ is the hypothesis output by learner L using the training set S drawn according to distribution \mathcal{D} .

When only **limited data** D_0 is available, we will produce an **estimation** of

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

- partition D_0 into training set S_0 and test set T_0 , and measure

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

- better, repeat this many times and average the results (next slide)
- use the t paired test to get an (approximate) confidence interval

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 ,
 - use T_i for the test set, and the remaining data for training set S_i
 - $S_i \leftarrow \{D_0 - T_i\}$
 - $h_A \leftarrow L_A(S_i)$
 - $h_B \leftarrow L_B(S_i)$
 - $\delta_i \leftarrow \text{error}_{T_i}(h_A) - \text{error}_{T_i}(h_B)$
3. Return the value $\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^k \delta_i$

Note: We'd like to **use the paired t test** on $\bar{\delta}$ to obtain a confidence interval. This is not really correct, because the training sets in this algorithm are not independent (they overlap!). But even this approximation is better than no comparison.

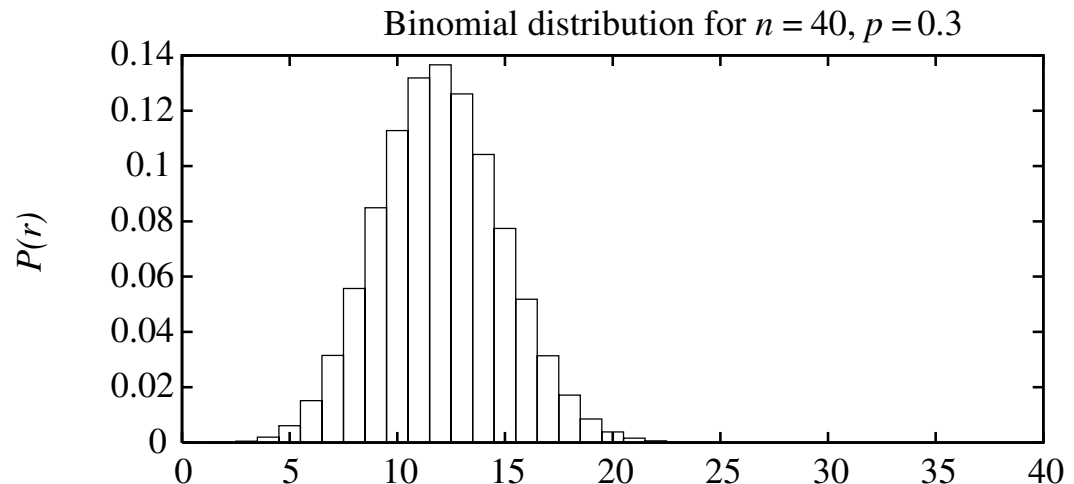
APPENDIX: Statistics Issues

- Binomial distribution, Normal distribution
- Confidence intervals
- Paired t tests

Binomial Probability Distribution

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

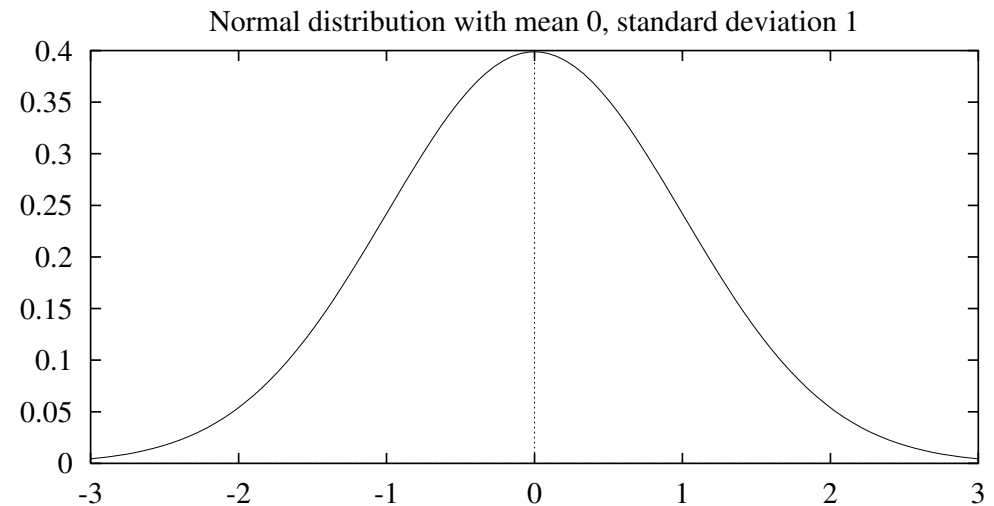
$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$



- 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 $\sigma_X \equiv \sqrt{E[(X - E[X])^2]} = \sqrt{np(1 - p)}$
- For large n , the Normal distribution **approximates** very closely the Binomial distribution.

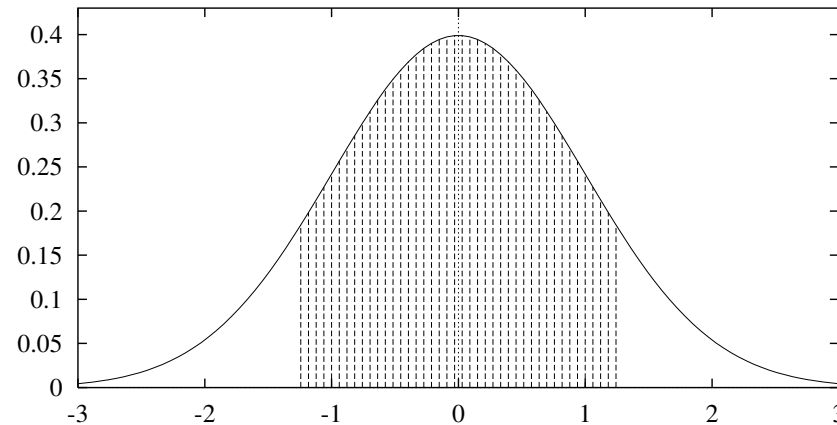
Normal Probability Distribution

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



- Expected, or **mean** value of X , is $E[X] = \mu$
- **Variance** of X is $Var(X) = \sigma^2$
- **Standard deviation** of X is $\sigma_X = \sigma$
- The probability that X falls into the **interval** (a, b) is $\int_a^b p(x)dx$

Normal Probability Distribution (I)

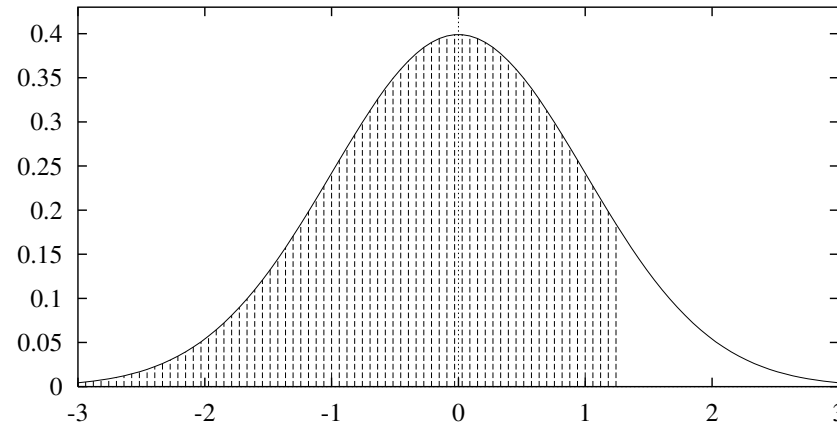


$N\%$ of area (probability) lies in $\mu \pm z_N\sigma$

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

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

Normal Probability Distribution (II)



$N\% + \frac{1}{2}(100-N\%)$ of area (probability) lies in $(-\infty; \mu + z_N\sigma)$

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

Example: 90% of area (probability) lies in the “one-sided” interval $(-\infty; \mu + 1.28\sigma)$

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

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

Note: δ_i is approximately Normally distributed.

3. Return the value $\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^k \delta_i$

$N\%$ confidence interval estimate for $d = \text{error}_{\mathcal{D}}(h_A) - \text{error}_{\mathcal{D}}(h_B)$ is:

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

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