Universität Augsburg University

Evaluating Hypotheses

Estimating Hypothesis Accuracy

- Questions:
  - Given the observed accuracy of a hypothesis over a limited sample of data, how well does this estimate its accuracy over additional samples?
  - Given that one hypothesis outperforms another over some sample of data, how probable is it that this hypothesis is more accurate in general?
  - When data is limited what is the best way to use this data to both learn a hypothesis and estimate its accuracy?
- Why?
  - Understand whether to use a hypothesis
  - Evaluating hypotheses is an integral component of many learning methods

Outline

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

Definitions

- X = space of possible instances x
- D = unknown probability distribution that defines the probability of encountering each instance in X.
  - Says nothing about concept values
  - Only occurrence probability

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*.

$$error_D(h) \equiv \Pr_{x \in D}\left[f(x) \neq h(x) \right]$$

Error rate of the hypothesis over the entire unknown distribution *D*

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

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

Error rate of the hypothesis over "some" sample set *S*

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 $error_D(h)$?

This is all we can measure

This is was we can expect when applying the hypothesis to future examples

Problems Estimating Error

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

$$bias \equiv E\left[error_S(h)\right] - error_D(h)$$

   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*

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

What is $error_D(h)$?

Estimators

Experiment:

1. Choose sample S of size n according to distribution D (i.e., independent of one another and of h)
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 $error_D(h)$

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

Confidence Intervals (1)

We show the results here first!

If

- S contains n examples, drawn independently of h and each other
- $n \geq 30$

Then

- With approximately 95% probability, $error_D(h)$ lies in interval

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

Example:

- n=40; $error_S(h)$ = 12 / 40 = 0,30
- If we repeatedly draw a new sample set S_{new} containing 40 new samples, we would find that for approx. 95% of these experiments, the calculated interval would contain the true error.

➔ 95% confidence interval estimate for $error_D(h)$

➔ 0,30 ± 0,14

Confidence Intervals (2)

We show the results here first!

If

- S contains n examples, drawn independently of h and each other
- $n \geq 30$

Then

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

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

where

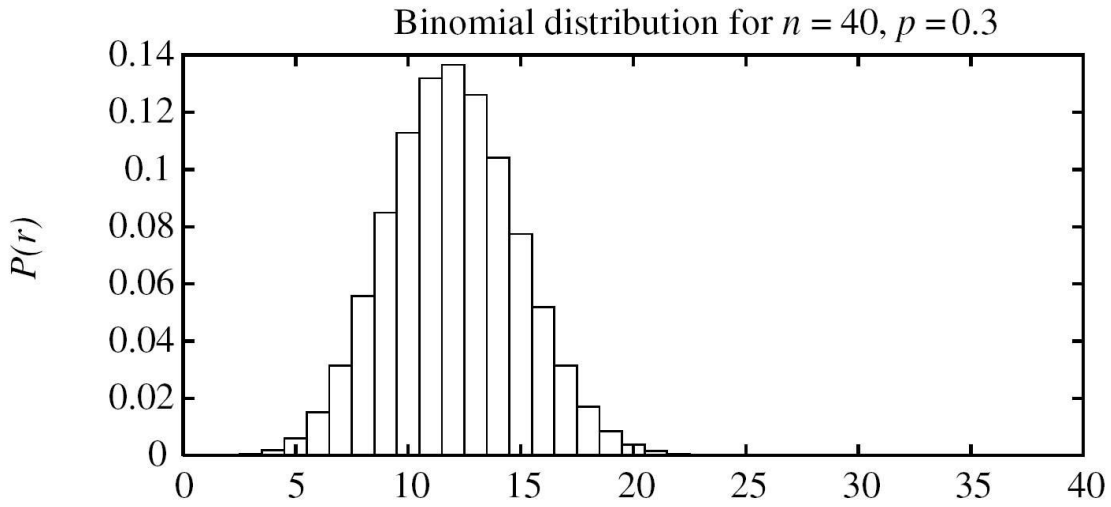
N%:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Confidence Intervals (3)

- Applies to discrete-valued hypotheses only
- Assumes *S* is drawn at random from *D*
- Data *S* is independent of the hypothesis being tested

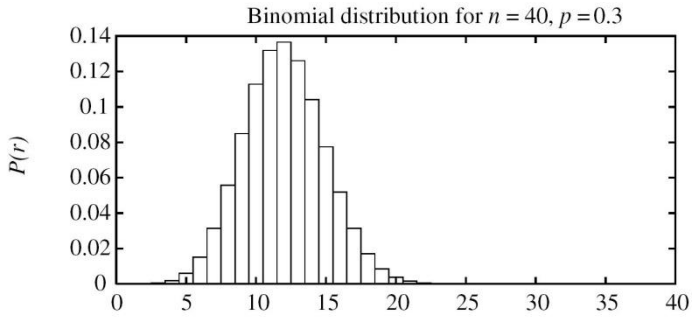
$error_S(h)$ is a Random Variable

- *Random variable* = name of an experiment with probabilistic outcome
- Rerun the experiment with different randomly drawn *S* (of size *n*)
- Probability of observing *r* misclassified examples:



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

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 $p = \Pr(heads)$

- 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

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

Application in our Context

- r = # of misclassifications
- n = total # of classifications
- X = random number of absolute number of misclassifications
- $p = \frac{r}{n}$ = relative frequency of misclassifications

- New random number Y:

$$Y := \frac{X}{n} \equiv error_S(h)$$

Express everything with $Y := \frac{X}{n}$

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

$$E[Y] \equiv E\left[\frac{X}{n}\right] = \frac{1}{n}E[X] = p = \sum_{i=0}^{n} \frac{i}{n} P(i)$$

- Variance of Y is

$$Var(Y) \equiv E[(Y - E[Y])^2]$$

$$= E\left[\left(\frac{X}{n} - E\left[\frac{X}{n}\right]\right)^2\right] = E\left[\left(\frac{1}{n}X - \frac{1}{n}E[X]\right)^2\right]$$

$$= E\left[\left(\frac{1}{n}(X - E[X])\right)^2\right] = E\left[\frac{1}{n^2}(X - E[X])^2\right]$$

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

- Standard deviation of X, σ_Y , is

$$\sigma_Y = \sqrt{\frac{p(1-p)}{n}}$$

Normal Distribution Approximates Binomial (1)

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

- mean $\mu_{error_S(h)} = error_D(h)$
- standard deviation

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

Step 1: Approximate $error_D(h)$ by $error_S(h)$

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

Normal Distribution Approximates Binomial (2)

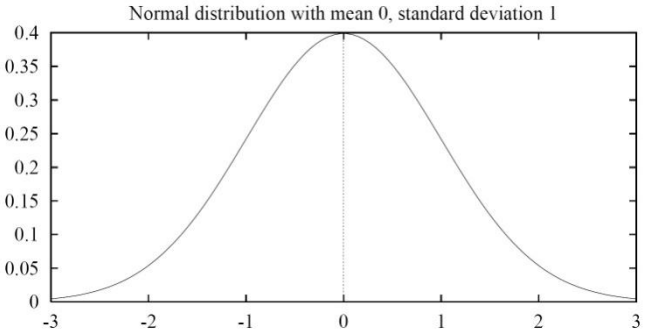
Step 2:

For sufficiently large values of *n* (e.g., for *np*(1-*p*) ≥ 5 or n>= 30) approximate this by a *Normal* distribution with

- mean $\mu_{error_S(h)} = error_D(h)$
- standard deviation

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

Normal Probability Distribution (1)



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

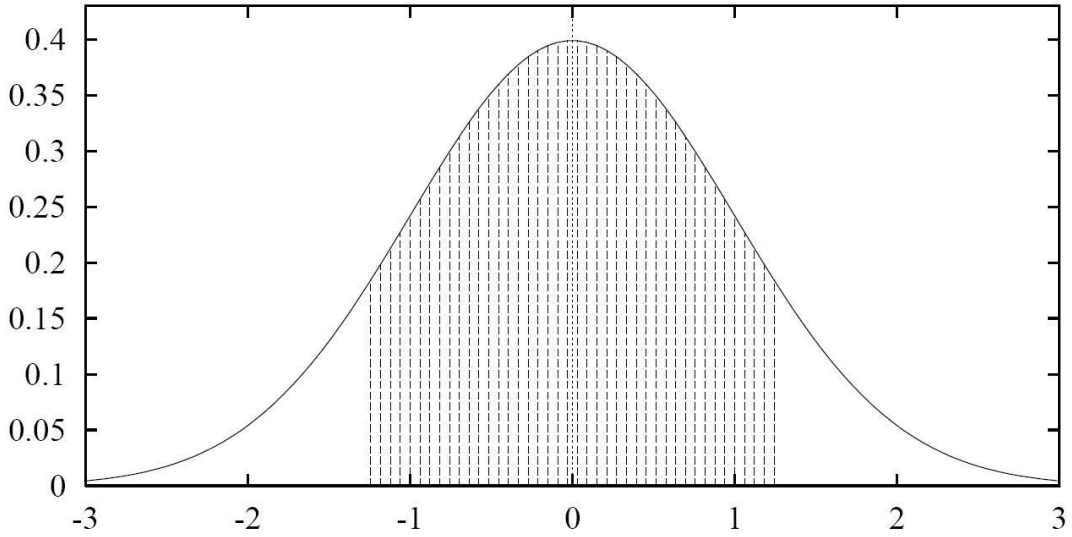
The probability that X will fall into the interval (a, b) is given by

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

- 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 $\sigma_X = \sigma$.

Normal Probability Distribution (2)



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

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

equivalently, $error_D(h)$ lies in interval

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

which is approximately (for $np(1-p) \geq 5$)

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

General Approach (1): Central Limit Theorem

Universität Augsburg University

Consider a set of independent, identically distributed (=iid) random variables $Y_1, \ldots, Y_n$, all governed by an arbitrary probability distribution with mean μ and finite variance σ^2. Define the sample mean,

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

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

➔General approach to deriving confidence intervals

Whenever we define an estimator that is the mean of some sample, the distribution governing this estimator can be approximated by a Normal distribution for sufficiently large *n*. If we also now the variance for this approx. Normal distribution, we can compute confidence intervals.

General Approach (2): Summary Calculating Confidence Intervals

Universität Augsburg University

1. Pick parameter p to estimate
   - $error_D(h)$
2. Choose an estimator
   - $error_S(h)$

Optimal: minimal variance + unbiased estimator

3. Determine probability distribution that governs estimator
   - Special case: If $error_S(h)$ governed by Binomial distribution, then 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

Example: Difference Between Hypotheses

Test h_1 on sample S_1, test h_2 on S_2

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
   - $\sigma_{\hat{d}} \equiv \sqrt{\frac{error_{S_1}(h_1)(1-error_{S_1}(h_1))}{n_1} + \frac{error_{S_2}(h_2)(1-error_{S_2}(h_2))}{n_2}}$
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* Tests (1/2)

- Paired test = test where hypotheses are evaluated over identical samples.
- Paired tests typically produce tighter confidence intervals because any difference in observed errors in a paired test are due to differences between the hypotheses.

- Estimation problem:
  - Given are the observed values of a set of i.i.d. random variables Y_1, Y_2,…, Y_k
  - Estimate the mean of the probability distribution governing these Y_i
  - The estimator is the sample mean

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

  - The individual Y_i follow (approximately) a Normal distribution

Paired *t* Tests (2/2)

$$\Rightarrow \qquad \mu \in \left[\bar{Y} - t_{N,k-1} s_{\bar{Y}}, \bar{Y} + t_{N,k-1} s_{\bar{Y}}\right]$$

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

The *t* distribution is used instead of the normal distribution whenever the standard deviation is estimated. With very few degrees of freedom, the *t* distribution is very leptokurtic. With 100 or more degrees of freedom, the *t* distribution is almost indistinguishable from the normal distribution. As the degrees of freedom increases, the *t* distribution approaches the normal distribution.

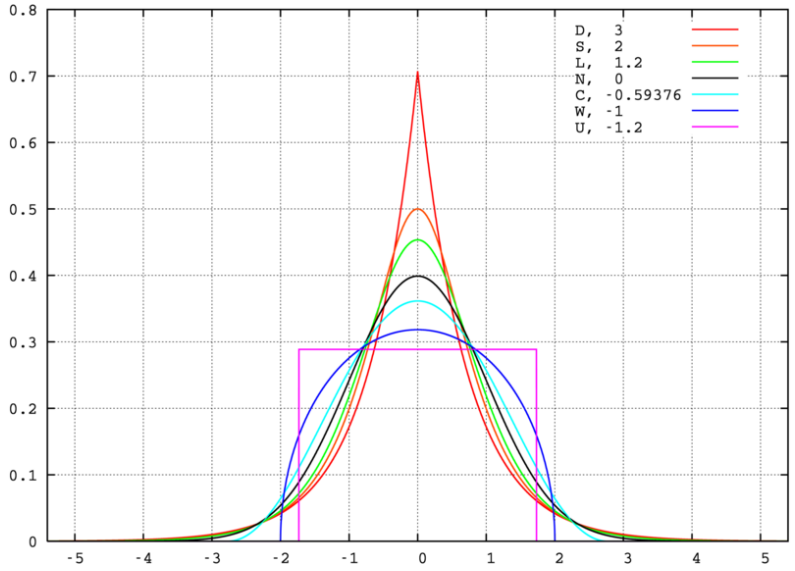


Figure from Wikipedia; "A distribution with positive excess kurtosis is called **leptokurtic.** A high kurtosis distribution has a sharper *peak* and longer

Student‘s Distribution

k	t für zweiseitigen Vertrauensbereich N=1-α							
	0,5	0,75	0,8	0,9	0,95	0,98	0,99	0,998
	t für einseitigen Vertrauensbereich N=1-α / 2							
	0,75	0,875	0,90	0,95	0,975	0,99	0,995	0,999
1	1,000	2,414	3,078	6,314	12,706	31,821	63,657	318,309
2	0,816	1,604	1,886	2,920	4,303	6,965	9,925	22,327
3	0,765	1,423	1,638	2,353	3,182	4,541	5,841	10,215
4	0,741	1,344	1,533	2,132	2,776	3,747	4,604	7,173
5	0,727	1,301	1,476	2,015	2,571	3,365	4,032	5,893
6	0,718	1,273	1,440	1,943	2,447	3,143	3,707	5,208
7	0,711	1,254	1,415	1,895	2,365	2,998	3,499	4,785
8	0,706	1,240	1,397	1,860	2,306	2,896	3,355	4,501
9	0,703	1,230	1,383	1,833	2,262	2,821	3,250	4,297
10	0,700	1,221	1,372	1,812	2,228	2,764	3,169	4,144
11	0,697	1,214	1,363	1,796	2,201	2,718	3,106	4,025
12	0,695	1,209	1,356	1,782	2,179	2,681	3,055	3,930
13	0,694	1,204	1,350	1,771	2,160	2,650	3,012	3,852
14	0,692	1,200	1,345	1,761	2,145	2,624	2,977	3,787
15	0,691	1,197	1,341	1,753	2,131	2,602	2,947	3,733
16	0,690	1,194	1,337	1,746	2,120	2,583	2,921	3,686
17	0,689	1,191	1,333	1,740	2,110	2,567	2,898	3,646
18	0,688	1,189	1,330	1,734	2,101	2,552	2,878	3,610
19	0,688	1,187	1,328	1,729	2,093	2,539	2,861	3,579

Source: Wikipedia

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*:

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

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

Note δ_i approximately Normally distributed

Comparing Learning Algorithms L_A and L_B

Comparing learning algorithms L_A and L_B (1)

What we'd like to estimate:

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

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?

- Could partition D_0 into training set S_0 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 (2)

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

   - $S_i \leftarrow \{D_0 - T_i\}$
   - $h_A \leftarrow L_A(S_i)$
   - $h_B \leftarrow L_B(S_i)$
   - $\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$
3. *Return the value $\bar{\delta}$ where*

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

Comparing learning algorithms L_A and L_B (3)

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

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

instead of

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

but even this approximation is better than no comparison