Machine Learning

Evaluating Hypotheses

Artificial Intelligence & Computer Vision Lab
School of Computer Science and Engineering
Seoul National University

Overview

- Motivation
- Estimating Hypothesis Accuracy
- Basics of Sampling Theory
- A General Approach for Driving Confidence Intervals
- Difference in Error of Two Hypotheses
- Comparing Learning Algorithms

Motivation

- The importance of evaluating hypotheses
  - To understand whether to use the hypothesis
  - An integral component of learning algorithm Ex) Post-pruning in DT
- Two difficulties
  - Bias in Estimate

    → Test examples chosen independently of the training examples
  - Variance in Estimate

    → Larger set of test examples
- Subjects in this chapter
  - Evaluating learned hypotheses
  - Comparing the accuracy of two hypotheses
  - Comparing the accuracy of two learning algorithms

Estimating Hypothesis Accuracy

- Notations

  X **:** space of all possible instances   ***D*** **:** probability distribution of ***X***
  S **:** sample drawn from ***D***   ***f*** **:** target function   ***h*** **:** hypothesis

- Sample error and true error
  - Sample Error : the fraction of *S* that it misclassifies

$$error_S(h) = \frac{1}{n}\sum_{x \in S} \delta\big(f(x), h(x)\big)$$

  - True Error : the misclassification probability a randomly drawn instance from *D*

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

"While we want to know $error_D(h)$, *we can measure only* $error_S(h)$*."*

→ *"How good an estimate of* $error_D(h)$ *is provided by* $error_S(h)$*?"*

Estimating Hypothesis Accuracy (cont.)

- Confidence interval for discrete-valued hypothesis

$$|S| = n, \quad n \geq 30, \quad error_S(h) = \frac{r}{n}$$

Confidence interval : $error_S(h) \pm z_N \sqrt{\frac{error_S(h)\left(1 - error_S(h)\right)}{n}}$

<u>Requirements</u>

❶ Discrete-valued hypothesis

❷ *S* drawn randomly from *D*

❸ The data independent of hypothesis

<u>Recommendation</u>

$n \geq 30$ **& $error_S(h)$ *is not too close to 0 or 1***

or

$$n \; error_S(h)\left(1 - error_S(h)\right) \geq 5$$

Basics of Sampling Theory

- Error estimation and estimating binomial poportions
  - The probability that h misclassifies $error_{S_i}(h)$

    Repeated experiment $\rightarrow$ Random variable $error_{S_i}(h)$

    $error_{S_i}(h)$ ~ Binomial Distribution

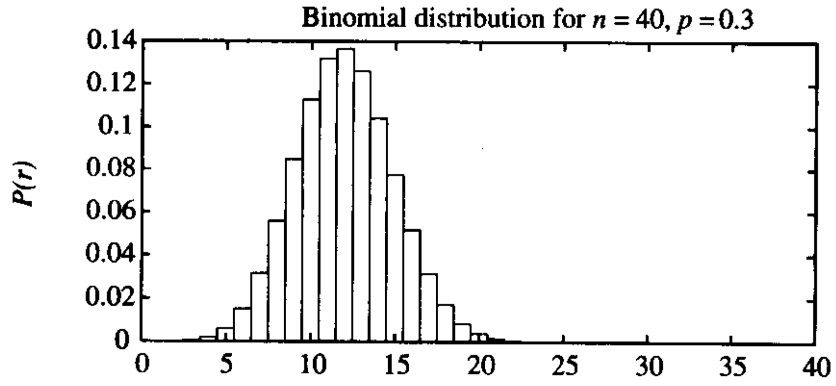
Coin toss example

- Toss a worn and bent coin ***n*** times
- The probability ***p***
- Heads turn up ***r*** times

"The probability of observing *r* heads"

≈ "The probability that *h* misclassifies"

$$p \approx error_D(h), \quad \hat{p} = \frac{r}{n} \approx error_S(h)$$



Basics of Sampling Theory (cont.)

Binomial distribution

- ***Y*** **: A random variable which can take on two values (Ex)** ***0*** **or** ***1***
- ***p*** **: The probability that on any single trial** ***Y=1***
- $Y_1, Y_2, \cdots, Y_n$ **: The sequence of i.i.d random variables** ***Y***
- $R \equiv \sum_{i=1}^{n} Y_i$ **: The number of trials for which** $Y_i = 1$ **in** ***n*** **independent experiments**

The probability that ***R*** **will take on a specific value** ***r***

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

Basics of Sampling Theory (cont.)

- Mean and variance

  Expected Value (or Mean)

  $$E[Y] \equiv \sum_{i=1}^{n} y_i \Pr(Y = y_i)$$

  Variance

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

  "How far the random variable is expected to vary from its mean value"

  Standard Deviation

  $$\sigma_Y \equiv \sqrt{Var[Y]} = \sqrt{E\left[(Y - E[Y])^2\right]}$$

※ In case of binomial distribution

Expected Value (or Mean)

$$E[Y] = np$$

Variance

$$Var[Y] = np(1 - p)$$

Standard Deviation

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

Basics of Sampling Theory (cont.)

- Estimators, bias, and variance
  - An estimator estimates the true value we do not know

    [Example] $error_S(h)$ estimates the true error $error_D(h)$
  - Estimation bias $\left(E[Y]-p\right)$

    : The difference between the expected value of estimator and the true value
  - Unbiased estimator : Y such that $E[Y]-p=0$
    - **As n grows larger,** $E[Y]\rightarrow p$
    - $error_S(h)$ **is the unbiased estimator of** $error_D(h)$
  - Variance : The smaller, the better

$$\sigma_{error_S(h)} = \sigma_{r/n} = \frac{\sigma_r}{n} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{error_S(h)\left(1-error_S(h)\right)}{n}}$$

Two quick remarks

❶ **S and h chosen independently**

❷ **Don’t be confused by “Estimation Bias” and “Inductive Bias”**

Basics of Sampling Theory (cont.)

- Example
  - 12 errors on a sample of 40 randomly drawn test examples

$$\hat{p} = error_S(h) = \frac{r}{n} = \frac{12}{40} = 0.3$$

$$\sigma_r^2 = np(1-p) \cong n\hat{p}(1-\hat{p}) = 40 \times 0.3 \times (1-0.3) = 8.4$$

$$\sigma_r = \sqrt{8.4} \cong 2.9$$

$$\sigma_{error_S(h)} = \sigma_{r/n} = \frac{\sigma_r}{n} = \frac{2.9}{40} = 0.07$$

Basics of Sampling Theory (cont.)

- Example
  - 300 errors on a sample of 1000 randomly drawn test examples

$$\hat{p} = error_S(h) = \frac{r}{n} = \frac{30}{1000} = 0.3$$

$$\sigma_r^2 = np(1-p) \cong n\hat{p}(1-\hat{p}) = 1000 \times 0.3 \times (1-0.3) = 210$$

$$\sigma_r = \sqrt{210} \cong 14.5$$

$$\sigma_{error_S(h)} = \sigma_{r/n} = \frac{\sigma_r}{n} = \frac{14.5}{1000} = 0.0145$$

  - As $\sigma_{error_S(h)}$ gets smaller, the confidence interval gets narrower with same probability

Basics of Sampling Theory (cont.)

- Normal distribution
  - A bell shaped distribution specified by its mean μ and standard deviation σ
  - Central limit theorem (See Section 5.4.1)

"Binomial distribution can be approximated by normal distribution"

Normal distribution

- *X* : A random variable $X \in (-\infty, +\infty)$

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

Cumulative distribution $\Pr[a \le X \le b] = \int_a^b p(x)dx$

Expected value, variance, and standard deviation

$$E[X] = \mu \qquad Var[X] = \sigma^2 \qquad \sigma_X = \sigma$$

Basics of Sampling Theory (cont.)

- Normal distribution
  - Table about the Standard Normal distribution $(\mu = 0, \ \sigma = 1)$; Table 5.1
  - The size of the interval about the mean that contains *N%* of the probability

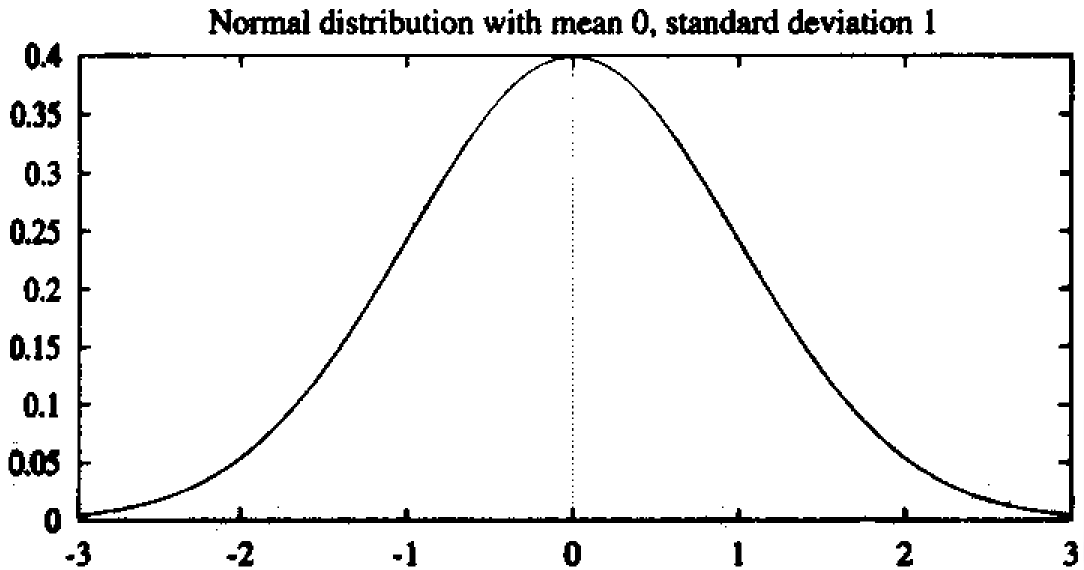


Table 5.1

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

Basics of Sampling Theory (cont.)

- Confidence intervals
  - $N\%$ confidence interval

    : An interval that is expected with probability $N\,\%$ to contain p
  - Confidence interval for μ and y : $y \pm z_N \sigma,\ \mu \pm z_N \sigma$

Obtaining Confidence Intervals for $error_D(h)$

❶ $error_S(h) \sim$ **Binomial distribution where** $\mu = error_D(h),\ \sigma = \sqrt{\dfrac{error_S(h)\left(1 - error_S(h)\right)}{n}}$

❷ **For large *n*, this binomial distribution is approximated by a normal distribution**

❸ **Find the N% confidence interval for estimating** μ **of a Normal distribution**

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)\left(1 - error_S(h)\right)}{n}} \qquad \left(n \geq 30 \quad \text{or} \quad np(1-p) \geq 5\right)$$

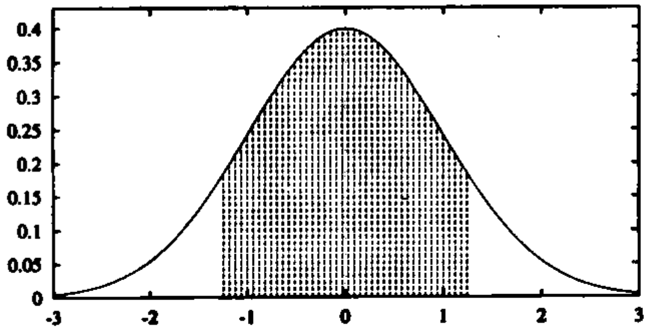
- Two approximations involved
  - $error_D(h)$ approximated by $error_S(h)$
  - Binomial distribution approximated by normal distribution

Basics of Sampling Theory (cont.)

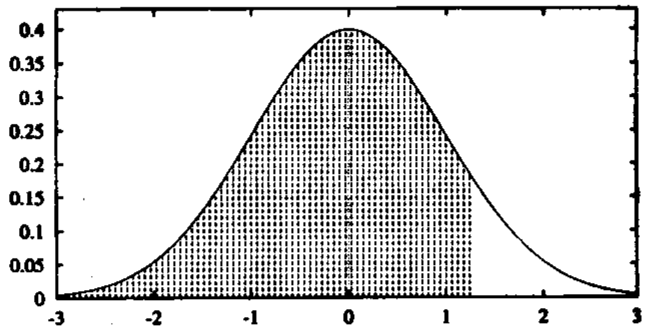
- Two-sided and one-sided bounds
  - Two-sided bound specifies both lower and upper bound
  - One-sided bound specifies either of them

"What is the probability that $error_D(h)$ is at most U?" → One-sided bound

$\begin{cases} 100(1-\alpha)\% & \textbf{Confidence Interval} \\ 100(1-\alpha/2)\% & \textbf{Confidence Interval} \end{cases}$	α : **The probability that the correct value lies outside the interval**



(a)



(b)

Basics of Sampling Theory (cont.)

- Example
  - 12 errors on a sample of 40 randomly drawn test examples

$$error_S(h) = 0.3$$

$$\sigma_{error_S(h)} = 0.07$$

(Two-sided) *95%* confidence interval $(\alpha = 0.05)$

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)\left(1 - error_S(h)\right)}{n}} = 0.3 \pm 1.96 \times 0.07 = 0.3 \pm 0.14$$

(One-sided) *97.5%* confidence interval $(\alpha = 0.05)$

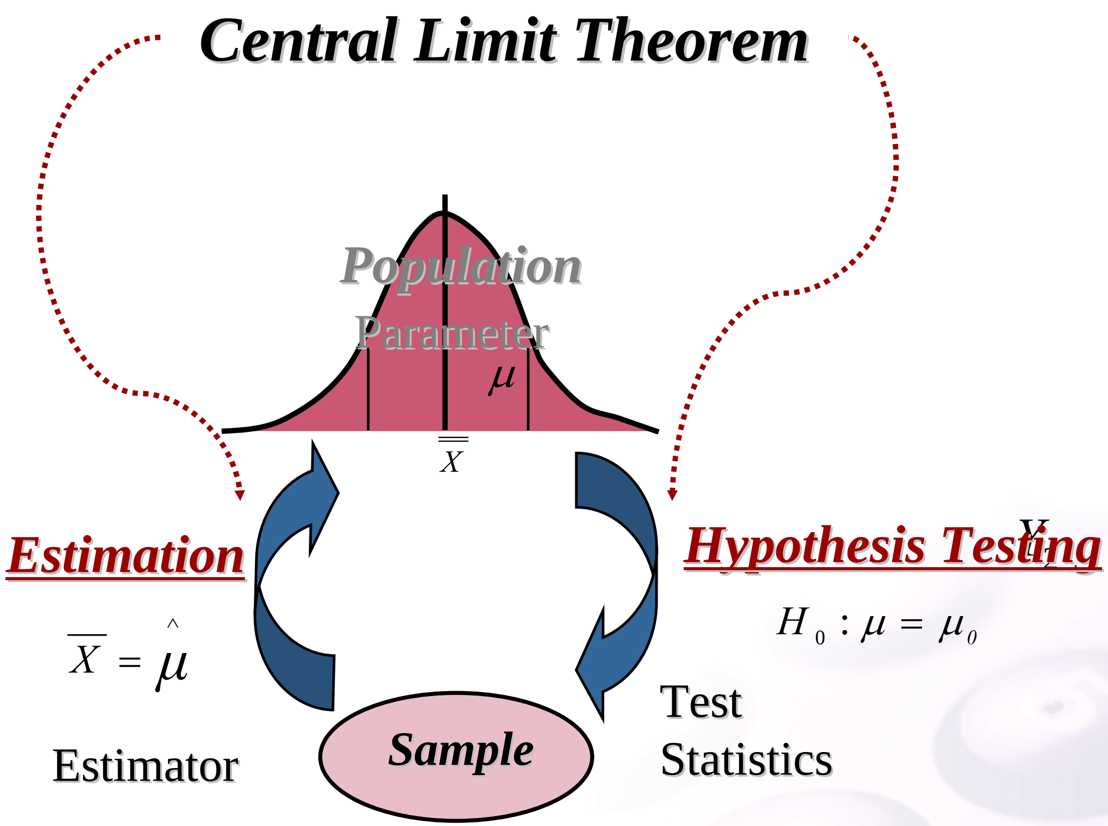
$error_D(h)$ **is at most** $0.3+0.14=0.44$

No assertion about the lower bound!

Next…

- A General approach for driving confidence intervals
  - Central Limit Theorem
- Difference in errors of two hypotheses
  - Hypothesis testing
- Comparing learning algorithms
  - Paired t-tests
  - Practical considerations

Estimation and Hypothesis Testing



Deriving Confidence Intervals

- General process estimating parameter P

(1) Identify the underlying population parameter *p* **:** $error_D(h)$

(2) Define the estimator *Y* **:** $error_S(h)$

: minimum variance, unbiased estimator desirable

(3) Determine the probability distribution D_Y of *Y*

: mean(μ) and variance(σ^2) of Y

(4) Determine the *N*% confidence interval from D_Y

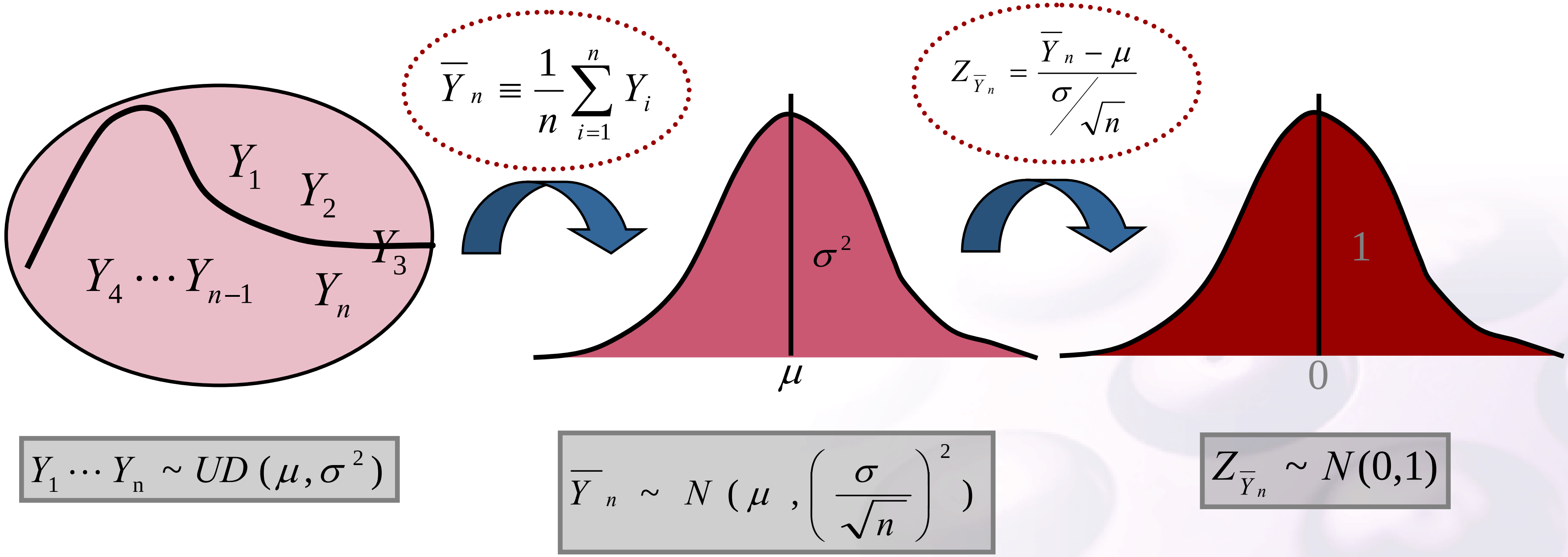
: LowerBound and UpperBound

$$\mu \pm z_n \cdot \sigma \xrightarrow{\text{For Discrete-valued Hypothesis}} error_S(h) \pm z_n \cdot \sqrt{\frac{error_s(h)(1-error_s(h))}{n}}$$

A General Approach for
Deriving Confidence Intervals (cont.)

- Central limit theorem

Consider a set of iid random variables $Y_1 \cdots Y_n$ governed by an arbitrary probability distribution with mean μ and finite variance σ^2. Define the sample mean, $\bar{Y}_n \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then as $\underline{n \to \infty}$, the distribution governing $\frac{\bar{Y}_n - \mu}{\sigma / \sqrt{n}}$ approaches a Normal Dist[n]. with zero mean and standard deviation equal to 1.



A General Approach for

Deriving Confidence Intervals (cont.)

- Why central limit theorem is useful ?
  - **We can know the dist^n. of sample mean $\overline{Y}$**
    (even when we do not know the dist^n. of Y_i)
  - **We can determine the mean(μ) and variance(σ^2) of Y_i .**
    (from the mean and variance of $\overline{Y}$)

$$Y_{ki} \sim UD(\mu, \sigma^2) \qquad\qquad \overline{Y}_k \sim N(\mu, S^2)$$

$$\begin{array}{lcl}
(1)\ Y_{11} \cdots Y_{1n} & \longrightarrow & \overline{Y}_1 \\
(2)\ Y_{21} \cdots Y_{2n} & \longrightarrow & \overline{Y}_2 \\
(3)\ Y_{31} \cdots Y_{3n} & \longrightarrow & \overline{Y}_3 \\
\vdots & \vdots & \vdots \\
(K)\ Y_{K1} \cdots Y_{Kn} & \longrightarrow & \overline{Y}_K
\end{array}$$

$$\text{mean}\ (\overline{Y}_k) = \mu$$

$$\text{variance}(\overline{Y}_k) = S^2 = \frac{\sigma^2}{n}$$

→ ***Then we can compute confidence interval !*** $\mu \pm z_n \cdot \sigma$

A Difference in Error of Two Hypotheses

- Parameter to be estimated

  : The difference between the true error of 2 hypotheses, h_1 & h_2.

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

- CASE 1 : Tested on <u>independent</u> test samples
  - Hypothesis h_1 : sample S_1 containing n_1 examples
  - Hypothesis h_2 : sample S_2 containing n_2 examples

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

  - $\hat{d}$ gives an unbiased estimate of d : $E(\hat{d}) = d$

$$
\begin{aligned}
E(\hat{d}) - d &= E\{error_{S1}(h_1) - error_{S2}(h_2)\} - \{error_D(h_1) - error_D(h_2)\} \\
&= E\{error_{S1}(h_1)\} - E\{error_{S2}(h_2)\} - \{error_D(h_1) - error_D(h_2)\} \\
&= [E\{error_{S1}(h_1)\} - error_D(h_1)] + [-E\{error_{S2}(h_2)\} + error_D(h_2)] \\
&\cong [error_D(h_1) - error_D(h_1)] + [-error_D(h_2) + error_D(h_2)] \\
&= 0
\end{aligned}
$$

A Difference in Error of Two Hypotheses (cont.)

- CASE 1 : Tested on <u>independent</u> test samples (continued)
  - **For large n_1, n_2 (>= 30), dist^n. of $\hat{d}$ is approximately Normal dist^n.**

    $$\because \quad error_{S1}(h_1) \sim N(\mu_1, \sigma_1), \quad error_{S2}(h_2) \sim N(\mu_2, \sigma_2)$$

    Difference of 2 normal distributions is also a normal distribution
  - **Mean of** $\hat{d}$

    $$E(\hat{d}) = \mathrm{E}\{error_{S1}(h_1) - error_{S2}(h_2)\} \cong \mu_1 - \mu_2$$

    recall : $E(aX - bY) = aE(X) - bE(Y)$ (if X and Y are independent R.V.)
  - **Variance of** $\hat{d}$

    $$\sigma_{\hat{d}}^2 \equiv \frac{error_{S1}(h_1)(1-error_{S1}(h_1))}{n_1} + \frac{error_{S2}(h_2)(1-error_{S2}(h_2))}{n_2}$$

    recall : $Var(aX - bY) = a^2 Var(X) + b^2 Var(Y)$ (if X and Y are independent R.V.)
  - **Confidence Interval of** $\hat{d}$ **(when n_1, n_2 are large enough).**

    $$\hat{d} \pm z_N \cdot \sqrt{\frac{error_{S1}(h_1)(1-error_{S1}(h_1))}{n_1} + \frac{error_{S2}(h_2)(1-error_{S2}(h_2))}{n_2}}$$

A Difference in Error of Two Hypotheses (cont.)

- CASE 2 : Tested on <u>a single</u> test sample

  : Hypothesis h_1 & Hypothesis h_2 are tested on a single test sample S.

  : Estimator $\hat{d} \equiv error_S(h_1) - error_S(h_2)$

– **Confidence interval of $\hat{d}$.**

$$\hat{d} \pm z_N \cdot \sqrt{\frac{error_S(h_1)(1-error_S(h_1)) + error_S(h_2)(1-error_S(h_2))}{n}}$$

– **Smaller variance comparing with CASE1.**

  : Single sample S eliminates the variance due to random differences in the S_1 and S_2.

A Difference in Error of Two Hypotheses (cont.)

- Hypothesis testing

: Testing for some specific conjecture (rather than in confidence intervals for some parameter)

– **Situation**

- Independent sample S_1 & S_2 ($|S_1| = |S_2| = 100$)
- $\textbf{error}_{S1}(h_1) = 0.30$
- $\textbf{error}_{S2}(h_2) = 0.20$
- $\hat{d} = 0.10$

"What is the probability the $\text{error}_D(h_1) > \text{error}_D(h_2)$ given $\hat{d} = 0.10$?"

"What is the probability that d>0 given $\hat{d} = 0.10$?"

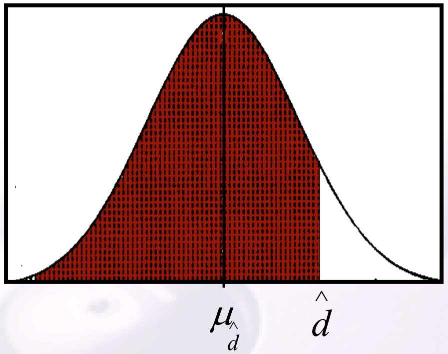
- $\hat{d}$ **falls into the one-sided interval** $\hat{d} < d + 0.10 \rightarrow \hat{d} < \mu_{\hat{d}} + 0.10$

$$\hat{d} < \mu_{\hat{d}} + Z_N \cdot \sigma_{\hat{d}}$$

$$Z_N \cdot \sigma_{\hat{d}} = 0.10 \quad , \quad \sigma_{\hat{d}} = \sqrt{\frac{0.3(1-0.3) + 0.2(1-0.2)}{100}} \approx 0.061$$

$$Z_N = 1.64$$

Two-sided constant for 90% confidence interval



– **Test result**

Therefore, the probability the $error_D(h_1) > error_D(h_2)$ is approximately 95% .

- **Accept H_0 with 95% confidence**
- **Reject H_0 with 5% significant level**

Comparing Learning Algorithms

Which of L_A and L_B is the better learning method on average for learning some particular target function f?

- Comparing the performance of two algorithms (L_A, L_B)

: Expected value of the difference in errors between L_A and L_B, where $L_A(S)$ is the hypothesis output by learning method, L_A, on the sample, S, of training data.

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

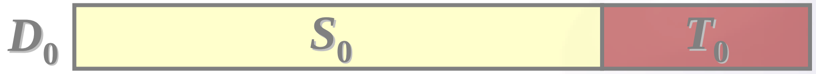
(S : Training Data sampled from underlying distribution D)

- Practical ways of algorithm comparison given limited sample, D_0, of data

(1) Partitioning data set into training set & test set

: A limited sample D_0 is divided into a training set S_0 and Test Set T_0

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



Comparing Learning Algorithms (cont.)

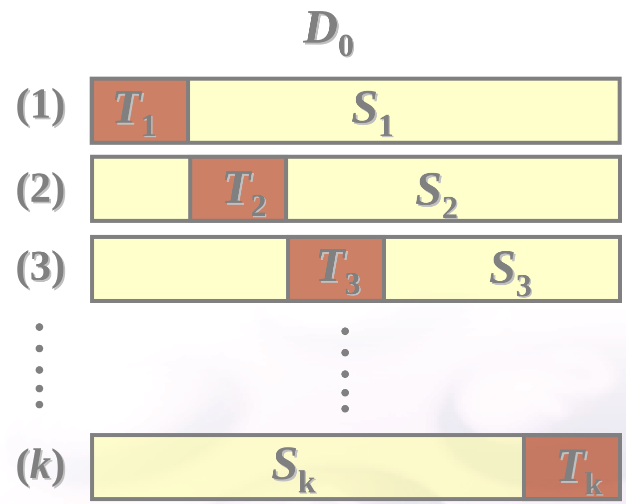
(2) Repeated partitioning and averaging : *k*-fold method

: D_0 is divided into disjoint training and test sets repeatedly and then the mean of the test set errors for these different experiment is calculated. $\underset{S \subset D_0}{E}[error_D(L_A(S)) - error_D(L_B(S))]$

1. Partition the available data D_0 into k disjoint subsets $T_1, T_2, \ldots, T_k$ 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
   - $S_i \leftarrow \{D_0 - T_i\}$
   - $h_A \leftarrow L_A(S_i)$
   - $h_B \leftarrow L_B(S_i)$
   - $\delta_i \leftarrow error_{T_i}(h_A) - error_{T_i}(h_B)$
3. Return the value $\bar{\delta}$, where

$$\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i \qquad \text{(T5.1)}$$



$\bar{\delta}$ returned from the above is the estimate of

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

$$|S_k| = \frac{k-1}{k}|D_0|, \qquad |T_k| \geq 30$$

which is again the approximation of $\underset{S \subset D_0}{E}[error_D(L_A(S)) - error_D(L_B(S))]$

Comparing Learning Algorithms (cont.)

(2) Repeated partitioning and averaging : *k*-fold method (continued)

- The approximate *N%* confidence interval

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

- *N* : Confidence level ,
- *k-1* : Degrees of freedom ν, number of independent random events producing the values for random variable $\bar{\delta}$
- If $k \rightarrow \infty$ $t_{N,k-1}$ approaches the constant z_N.

	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

Paired test : *Tests where the hypotheses are evaluated over identical samples.*

Paired Test generate tighter confidence interval than Test on Separate Data samples (Due to eliminate the difference of sample makeup)

Comparing Learning Algorithms (cont.)

- Paired *t*-test

: Statistical justification of the previous comparing algorithm procedure

– **Estimation procedure**

(1) Given i.i.d. random variables : $Y_1, \ldots, Y_k$

(2) Estimate the mean μ of distribution governing Y_i from estimator

(3) Estimator : $$\overline{Y} \equiv \frac{1}{k}\sum_{i=1}^{k} Y_i$$

Comparing Learning Algorithms (cont.)

- *t*-test, which is applicable to the special case of the estimator procedure where each Y_i follows a Normal distribution, provides

$$\overline{Y} - t_{N,k-1} \cdot s_{\overline{Y}} \leq \mu = E(Y_i) \leq \overline{Y} + t_{N,k-1} \cdot s_{\overline{Y}} \quad \text{, where} \quad s_{\overline{Y}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (Y_i - \overline{Y})^2}$$

where $t_{N,k\text{-}1}$ is a constant characterizing *t* distribution as z_n characterizes a Normal distribution.

- In the previous comparing learning algorithm, if on each iteration a new random training set S_i and new random test set T_i are drawn from the underlying instance distribution instead of the fixed sample D_0, then each $\delta_i = error_{T_i}(h_A) - error_{T_i}(h_B)$ with $|T_{\mathrm{i}}| \geq 30$ follows a normal distribution and thus from *t*-test result,

$$\mu = E(\delta_i) = \underset{S \subset D}{E}[error_D(L_A(S)) - error_D(L_B(S))] = \overline{\delta} \pm t_{N,K-1} \cdot s_{\overline{\delta}}$$

Comparing Learning Algorithms (cont.)

- Practical considerations

Paired t-test does not strictly justify the confidence interval previously discussed because it is evaluated on a limited data D_0 and partitioned method. Nevertheless, this confidence interval provides good basis for experimental comparisons of learning methods.

- When data is limited…

’ **(1) *k*-fold method**

- k is limited.
- Test set are drawn independently (examples are tested exactly once)

(2) Randomized method

: Randomly choose a test set at least 30 examples from D_0 and use remaining examples for training.

- Procedure can be repeated infinitely
  (k can be infinite number → narrower confidence interval)
- Test sets are not independent.