Evaluating Hypotheses

[Read Ch. 5]
[Recommended exercises: 5.2, 5.3, 5.4]

• 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, $\text{error}_S(h)$ is optimistically biased

   \[
   \text{bias} \equiv E[\text{error}_S(h)] - \text{error}_D(h)
   \]

   For unbiased estimate, $h$ and $S$ must be chosen independently

2. **Variance:** Even with unbiased $S$, $\text{error}_S(h)$ may still vary from $\text{error}_D(h)$
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 $\mathcal{D}$

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, $error_D(h)$ 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, $\text{error}_D(h)$
  lies in interval

\[
\text{error}_S(h) \pm z_N \sqrt{\frac{\text{error}_S(h)(1 - \text{error}_S(h))}{n}}
\]

where

<table>
<thead>
<tr>
<th>$N%$</th>
<th>50%</th>
<th>68%</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>98%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_N$</td>
<td>0.67</td>
<td>1.00</td>
<td>1.28</td>
<td>1.64</td>
<td>1.96</td>
<td>2.33</td>
<td>2.58</td>
</tr>
</tbody>
</table>
$error_S(h)$ is a Random Variable

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

Probability of observing $r$ misclassified examples:

$$P(r) = \frac{n!}{r!(n-r)!} \cdot error_D(h)^r \cdot (1 - error_D(h))^{n-r}$$
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 \( p = \text{Pr(heads)} \)

- Expected, or mean value of \( X \), \( E[X] \), is
  \[ E[X] = \sum_{i=0}^{n} i P(i) = np \]

- Variance of \( X \) is
  \[ \text{Var}(X) = E[(X - E[X])^2] = np(1-p) \]

- Standard deviation of \( X \), \( \sigma_X \), is
  \[ \sigma_X = \sqrt{E[(X - E[X])^2]} = \sqrt{np(1-p)} \]
Normal Distribution Approximates Binomial

$\text{error}_S(h)$ follows a *Binomial* distribution, with

- mean $\mu_{\text{error}_S(h)} = \text{error}_D(h)$
- standard deviation $\sigma_{\text{error}_S(h)}$

$$
\sigma_{\text{error}_S(h)} = \sqrt{\frac{\text{error}_D(h)(1 - \text{error}_D(h))}{n}}
$$

Approximate this by a *Normal* distribution with

- mean $\mu_{\text{error}_S(h)} = \text{error}_D(h)$
- standard deviation $\sigma_{\text{error}_S(h)}$

$$
\sigma_{\text{error}_S(h)} \approx \sqrt{\frac{\text{error}_S(h)(1 - \text{error}_S(h))}{n}}
$$
Normal Probability Distribution

\[
p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}
\]

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
  \[
  \sigma_X = \sigma
  \]
80% of area (probability) lies in $\mu \pm 1.28\sigma$

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

<table>
<thead>
<tr>
<th>$N%$:</th>
<th>50%</th>
<th>68%</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>98%</th>
<th>99%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z_N$:</td>
<td>0.67</td>
<td>1.00</td>
<td>1.28</td>
<td>1.64</td>
<td>1.96</td>
<td>2.33</td>
<td>2.58</td>
</tr>
</tbody>
</table>
Confidence Intervals, More Correctly

If

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

Then

- With approximately 95% probability, $\text{error}_S(h)$ lies in interval

$$\text{error}_D(h) \pm 1.96 \sqrt{\frac{\text{error}_D(h)(1 - \text{error}_D(h))}{n}}$$

equivalently, $\text{error}_D(h)$ lies in interval

$$\text{error}_S(h) \pm 1.96 \sqrt{\frac{\text{error}_D(h)(1 - \text{error}_D(h))}{n}}$$

which is approximately

$$\text{error}_S(h) \pm 1.96 \sqrt{\frac{\text{error}_S(h)(1 - \text{error}_S(h))}{n}}$$
Central Limit Theorem

Consider a set of independent, identically distributed random variables $Y_1 \ldots Y_n$, 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 $n \to \infty$, 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
   - \( \text{error}_D(h) \)
2. Choose an estimator
   - \( \text{error}_S(h) \)
3. Determine probability distribution that governs estimator
   - \( \text{error}_S(h) \) 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

$$d \equiv \text{error}_D(h_1) - \text{error}_D(h_2)$$

2. Choose an estimator

$$\hat{d} \equiv \text{error}_{S_1}(h_1) - \text{error}_{S_2}(h_2)$$

3. Determine probability distribution that governs estimator

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

4. Find interval $(L, U)$ such that $N\%$ of probability mass falls in the interval

$$\hat{d} \pm z_N \sqrt{\frac{\text{error}_{S_1}(h_1)(1 - \text{error}_{S_1}(h_1))}{n_1} + \frac{\text{error}_{S_2}(h_2)(1 - \text{error}_{S_2}(h_2))}{n_2}}$$
Paired t test to compare $h_A, h_B$

1. Partition data into $k$ disjoint test sets $T_1, T_2, \ldots, T_k$ 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 $d$:
   $$
   \bar{\delta} \pm t_{N,k-1} \ s_{\bar{\delta}}
   $$
   $$
   s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}
   $$

*Note $\delta_i$ approximately Normally distributed*
Comparing learning algorithms $L_A$ and $L_B$

What we’d like to estimate:

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

where $L(S)$ is the hypothesis output by learner $L$

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 $T_1, T_2, \ldots, T_k$ 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

$$\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i$$
Comparing learning algorithms $L_A$ and $L_B$

Notice we’d like to use the paired $t$ test on $\tilde{\delta}$ 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

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

instead of

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

but even this approximation is better than no comparison