Outline

- Sample error vs. true error
- Confidence intervals for observed hypothesis error
- Estimators
- Binomial distribution, Normal distribution, Central Limit Theorem
- Paired $t$ tests
- Comparing learning methods
Notation & Two Questions

A Hypothesis is an approximation of \( f(x) \), the target function

Notation
- \( X \): the space of instances
- \( D \): the probability distribution of encountering instances from \( X \)
- \( f \): the target function
- \( H \): the hypothesis space
- \( h \): a particular hypothesis in \( H \)
- \( (x, f(x)) \): a training instance
- \( S \): all training instances

Two Questions
- Given \( h \) constructed from \( n \) examples drawn randomly from \( D \), what is the best estimate of \( h \) over future instances drawn from \( D \)?
- What is the probable error in this accuracy estimate?
True error vs. sample 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.

Q: How well does $error_{S}(h)$ estimate $error_{\mathcal{D}}(h)$?
Example

Hypothesis $h$ misclassifies 12 of the 40 examples in $S$

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

Q: What is $error_D(h)$?

- It is like estimating $Pr(tail)$ from the results of a series of coin-tossing experiments, i.e., $Pr(tail) = \frac{n_T}{N}$, where $n_T$ is the number of tail events.

- That is, $error_S(h)$ is a natural estimator for $error_D(h)$, but, $error_S(h)$ will be different for different choice of $S$ just as $\frac{n_T}{N}$ has experimental variation.

- We need a statistical measure of the confidence about the estimator $error_S(h)$. 
Bias and Variance

• *Bias*: If $S$ is training set, $\text{error}_S(h)$ is optimistically biased

\[ \text{bias} \equiv E[\text{error}_S(h)] - \text{error}_\mathcal{D}(h) \]

If there is no bias, then

\[ E[\text{error}_S(h)] = \text{error}_\mathcal{D}(h) \]

It will be the case when $S$ is chosen independent of $h$.

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

- **Experiment:**
  1. choose sample $S$ of size $n$ according to distribution $\mathcal{D}$
  2. measure $\text{error}_S(h)$

- $\text{error}_S(h)$ is a random variable (i.e., result of an experiment)
- $\text{error}_S(h)$ is an unbiased estimator for $\text{error}_\mathcal{D}(h)$
- Given observed $\text{error}_S(h)$, what can we conclude about $\text{error}_\mathcal{D}(h)$?
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 $X$: $E[X] \equiv \sum_{i=0}^{n} iP(i) = np$

- Variance of $X$: $Var(X) \equiv E[(X - E[X])^2] = np(1-p)$

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

Example: Initializing a Genetic Algorithm genome by randomly setting the bits is producing a binomially distributed initialization. This will bias the GA.
Binomial Probability Distribution

Binomial distribution for $n = 40, p = 0.3$
error_{S}(h) is a Random Variable

- PDF:
  \[ P(error_{S} = r/n) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}, p = error_{D}(h) \]

- \( E[error_{S}] = p \)

- \( Var(error_{S}) = \ldots = \frac{p(1-p)}{n} \)

- \( \sigma_{error_{S}} = \sqrt{\frac{p(1-p)}{n}} \)
Normal Probability Distribution
Normal Probability Distribution

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

- The probability that $X$ will fall into the interval $(a, b)$:
  
  $$\int_a^b p(x) \, dx$$

- Expected, or mean value of $X$: $E[X] = \mu$

- Variance of $X$: $Var(X) = \sigma^2$

- Standard deviation of $X$: $\sigma_X = \sigma$
• 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>
**error}_S(h) is a Random Variable

Approximate \( P\{error}_S \) by a Normal distribution with

- mean \( \mu_{error}_S(h) = error}_D(h) \)
- standard deviation \( \sigma_{error}_S(h) \approx \sqrt{\frac{error}_S(h)(1-error}_S(h))}{n} \)
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>
Confidence Intervals, More Correctly

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

$$error_S(h) \pm 1.96 \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$
Estimating Confidence Intervals In General

1. Pick parameter $p$ to estimate
   - $error_D(h)$

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
Why are we doing all of this Math???

Because it’s important for evaluating learning algorithms against one another

How would you compare two decision trees against each other? A decision tree and a neural network? How do you know if you are testing the performance of learner enough?

Work Exercises 5.2, 5.3, 5.4