Evaluating Hypotheses

Henrik Andreasson
AASS, Learning Systems Lab
April 2005
Main problem to solve

- Given two (or more) hypotheses (classifiers) and a set of data: determine which one that works best

\[ \text{error}(h1) < \text{error}(h2) ? \]
error(h1) < error(h2)?

- How to select the LIMITED data to do the evaluation in the best way? (paired test etc.).
- How GOOD will (probably) future sample be classified? (confidence interval).
error(h) measure

- The "correct" error: the TRUE error (ground truth) can't be measured without going true all data D ... (impossible if future data is to be considered)

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

- SAMPLE error, randomly draw some samples n from the data set S and evaluate the performance

\[
error_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x))
\]
error(h) measure

- How well does $\text{error}_S(h)$ fits $\text{error}_D(h)$?

  $\text{error}_D(h)$ is *UNKNOWN*, but how to estimate it???

"Problems":

- bias (or an offset)

  $\text{bias} = \mathbb{E}[\text{error}_S(h)] - \text{error}_D(h)$

- variance (we'll have this even without bias)

  *we sample the distribution:*

    *sometimes we're lucky (get better results than average)*

    *sometimes we're not (get worse results than average)*
example: flipping a coin

- What's the probability that the coin will show heads / tails?
- 12 heads of 40 gives
  \[ p_{\text{getting a head}} = \frac{40}{12} = 0.3 \]
- What about the variance?

Could be seen as the probability a hypotheses (classifier) is correct.
binomial distribution

- If you have true / false statements this is the distribution you will have. \( \text{variance} = np(1-p) \)
- Convert to normal distribution (its easier to ”calculate” confidence intervals) with the Central Limit Theorem.
normal distribution

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

Normal distribution:
80% confidence interval

- Confidence interval: an area relative to a probability

<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>
central limit theorem

• How do we convert from binomial to normal distribution?

• By trowing the coin ”many” times it will automatically converge to a normal distribution (in the book 30 times).

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

\[
\bar{Y} \equiv \frac{1}{n} \sum_{i=1}^{n} Y_i
\]
error_S(h) follows a Binomial distribution, with

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

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

standard deviation for error_S(h) is calculated from the variance equation for the binomial case: $\text{variance} = pn(1-p)$

Approximate this by a Normal distribution with

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

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

Select the confidence interval (normal distribution)

<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>

- 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

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

which is approximately

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

We have an estimation, it's variance and confidence interval!
evaluation of hypotheses

\[ \text{error}(h_1) < \text{error}(h_2)? \]

• Which is most likely the best classifier: 1 or 2?

\[ d = \text{error}(h_1) - \text{error}(h_2) \]

Same approach as before. To estimate the new variances, add the previous ones (\textit{conservative}).

\[
\sigma_d \approx \sqrt{\frac{\text{errors}_1(h_1)(1 - \text{errors}_1(h_1))}{n_1} + \frac{\text{errors}_2(h_2)(1 - \text{errors}_2(h_2))}{n_2}}
\]

Use confidence interval to get the estimation.

\textit{(Area right of "0" is voting for classifier 1.)}

Note that it's different samples for each hypotheses!
evaluation of hypotheses paired

\[ \text{error}(h_1) < \text{error}(h_2)? \]

Same sample S is evaluated for both algorithms

Create k different subsets of the data (at least 30)

Bigger k = better (less variance)

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

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

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

error(L1) < error(L2)?

Same sample S is evaluated for both algorithms

Limited data -> reuse the data (bad, what else to do?)

Divide subset of data into training and testing

(same size, disjoint)

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

\[ \bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^{k} \delta_i \]
finito