A Prediction Interval for the Misclassification Rate

E.B. Laber
&
S.A. Murphy
Outline

– Review
– Three challenges in constructing PIs
– Combining a statistical approach with a learning theory approach to constructing PIs
– Relevance to confidence measures for the value of a dynamic treatment regime.
Review

– $X$ is the vector of features in $R^q$, $Y$ is the binary label in $\{-1,1\}$
– Misclassification Rate: $\text{err}(f) = E[1\{Y \neq f(X)\}]$

– Data: $N$ iid observations of $(Y,X)$

– Given a space of classifiers, $\mathcal{F}$, and the data, use some method to construct a classifier, $\hat{f}$

– The goal is to provide a PI for $\text{err}(\hat{f})$
Review

- Since the loss function $1\{Y \neq f(X)\}$ is not smooth, one commonly uses a smooth surrogate loss to estimate the classifier

- Surrogate Loss: $L(Y, f(X))$

- $\hat{f} \in \min_{f \in \mathcal{F}} E_N[L(Y, f(X))]$

($E_N$ denotes expectation with respect to empirical distribution)
Review

General approach to providing a PI:

– We estimate $\hat{err}(\hat{f})$ using the data, resulting in $\hat{err}(\hat{f})$
– Derive approximate distribution for $\left(\hat{err}(\hat{f}) - err(\hat{f})\right)$
– Use this approximate distribution to construct a prediction interval for $err(\hat{f})$
A common choice for $\hat{err}(\hat{f})$ is the resubstitution error or training error:

$$\hat{err}_{rs}(f) = E_n[1\{Y \neq f(X)\}]$$

evaluated at $f = \hat{f}$ e.g. if $f(x) = \text{sign}(x^T\beta)$ then

$$\hat{err}(\hat{f}) = E_n[1\{Y X^T\hat{\beta} < 0\}]$$
Three challenges

1) $\mathcal{F}$ is too large leading to over-fitting and

$$E \left[ err(\hat{f}) - err(\tilde{f}) \right] < 0 \text{ (negative bias)}$$

2) $err(f) = E[1\{Y \neq f(X)\}]$ is a non-smooth function of $f$.

3) $err(\hat{f})$ may behave like an extreme quantity

No assumption that $\hat{f}$ is close to optimal.
2) \( err(f) = E[1\{Y \neq f(X)\}] \) is non-smooth.

Example: The unknown optimal classifier has quadratic decision boundary. We fit, by least squares, a linear decision boundary

\[ f(x) = \text{sign}(\beta_0 + \beta_1 x) \]

\( err(f) = E[1\{Y(\beta_0 + \beta_1 X) < 0)\}] \)
Density of $\text{err}(\hat{f})$

Three Point Dist. (n=30)

Three Point Dist. (n=100)
Bias of Common $\hat{err}(\hat{f})$ on Three Point Example
Coverage of Bootstrap PI in Three Point Example (goal = 95%)
Coverage of Correctly Centered Bootstrap PI (goal = 95%)
### Coverage of 95% PI (Three Point Example)

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Bootstrap Percentile</th>
<th>Yang CV</th>
<th>CUD-Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>.79</td>
<td>.75</td>
<td>.97</td>
</tr>
<tr>
<td>50</td>
<td>.79</td>
<td>.62</td>
<td>.97</td>
</tr>
<tr>
<td>100</td>
<td>.78</td>
<td>.46</td>
<td>.96</td>
</tr>
<tr>
<td>200</td>
<td>.78</td>
<td>.35</td>
<td>.96</td>
</tr>
</tbody>
</table>
Non-smooth

In general the distribution of

\[ \sqrt{N}(\hat{err}(\hat{f}) - err(\hat{f})) \]

may not converge as the training set increases (variance never settles down).
Intuition

Consider the large sample variance of

\[ \sqrt{N} \left( E_N[1\{Y X^T \beta < 0\}] - E[1\{Y X^T \beta < 0\}] \right) \]

Variance is \( p(1 - p), \quad p = P[Y X^T \beta < 0] \)

if in place of \( \beta \) we put \( \bar{\beta} \) where \( \bar{\beta} \) is close to 0

then due to the non-smoothness in

\[ p = P(Y X^T \beta < 0) \]

at \( \beta = 0 \) we can get jittering.
PIs from Learning Theory

Given a result of the form: for all $N$

$$P \left[ \sup_{f \in \mathcal{G}_N} |\text{err}_{rs}(f) - \text{err}(f)| < B_{N,\delta} \right] > 1 - \delta$$

where $\hat{f}$ is known to belong to $\mathcal{G}_N$ and

$$\text{err}_{rs}(f) = E_N[1\{Y \neq f(X)\}]$$

forms a conservative $1-\delta$ PI:

$$\text{err}_{rs}(\hat{f}) - B_{N,\delta} < \text{err}(\hat{f}) < \text{err}_{rs}(\hat{f}) + B_{N,\delta}$$
Combine statistical ideas with learning theory ideas

Construct a prediction interval for

$$\sup_{f \in \mathcal{G}_N} |\widehat{err}_{rs}(f) - err(f)|$$

where $\mathcal{G}_N$ is chosen to be small yet contain $\hat{f}$

---from this PI deduce a conservative PI for $err(\hat{f})$

---use the surrogate loss to perform estimation and to construct $\mathcal{G}_N$
Construct a prediction interval for
\[ \sup_{f \in \mathcal{G}_N} |\hat{err}_{rs}(f) - \hat{err}(f)| \]

--- \( \mathcal{G}_N \) should contain all \( f \) that are close to \( \hat{f} \)

--- all \( f \) for which

\[ E_N[L(Y, \hat{f}(X)) - E_N[L(Y, f(X)]] > 0 \]

--- \( \hat{f} \) is the “limiting value” of \( \hat{f} \);

\[ \hat{f} = \arg \max_{f \in \mathcal{F}} E[L(Y, f(X))] \]
Prediction Interval

Construct a prediction interval for

$$\sup_{f \in \mathcal{F}} \left\{ \left| \hat{\text{err}}_{\text{rs}}(f) - \text{err}(f) \right| \times g \left( N \left( E_N[L(Y, \tilde{f}(X))] - E_N[L(Y, f(X))] \right) \right) \right\}$$

$$g(u) = (1 + u)1\{-1 \leq u \leq 0\} + 1\{u > 0\}$$
Prediction Interval

\[ \left| \widehat{\text{err}}(\hat{f}) - \text{err}(\hat{f}) \right| \leq \left| \widehat{err}_r(\hat{f}) - \text{err}(\hat{f}) \right| \]

= 

\[ \left| \widehat{err}_r(\hat{f}) - \text{err}(\hat{f}) \right| \times \]
\[ g \left( N(E_N[L(Y, \hat{f}(X))] - E_N[L(Y, \hat{f}(X))] \right) \]

\leq 

\[ \sup_{f \in \mathcal{F}} \left\{ \left| \widehat{err}_r(f) - \text{err}(f) \right| \times \right. \]
\[ g \left( N(E_N[L(Y, \hat{f}(X))] - E_N[L(Y, f(X))] \right) \left. \right\} \]
Bootstrap

We use bootstrap to obtain an estimate of an upper percentile of the distribution of 

\[ \sup_{f \in \mathcal{F}} \left\{ \left( \widehat{err}_{rs}(f) - err(f) \right) \times \right. \]
\[ \left. g \left( N \left( E_N[L(Y, \tilde{f}(X)) - E_N[L(Y, f(X))] \right) \right) \right\} \]

to obtain \( b_U \). The PI is then

\[ \widehat{err}(\tilde{f}) - b_L \leq err(\tilde{f}) \leq \widehat{err}(\tilde{f}) + b_U \]
Implementation

• Approximation space for the classifier is linear:
  \[ \mathcal{F} = \{ f(x) = \text{sign}(x^T \beta) : \beta \in \mathbb{R}^p \} \]

• Surrogate loss is least squares:
  \[ L(y, f(x)) = (y - x^T \beta)^2 \]

• \( \hat{err}(\hat{f}) = err_{rs}(\hat{f}) \) (resubstitution error)
Implementation

\[
\sup_{f \in \mathcal{F}} \left\{ \left( \hat{err}_{\mathcal{S}}(f) - err(f) \right) \times g \left( N(E_N[L(Y, \bar{f}(X))] - E_N[L(Y, f(X))]) \right) \right\}
\]

becomes

\[
\sup_{\beta \in \mathbb{R}^q} \left\{ \left( E_N[1\{Y X^T \beta < 0\}] - E[1\{Y X^T \beta < 0\}] \right) \times g \left( N(E_N[(Y - X^T \bar{\beta})^2 - E_N[(Y - X^T \beta)^2]]) \right) \right\}
\]
Implementation

• Bootstrap version:

\[
\sup_{\beta \in \mathbb{R}^d} \left\{ \left[ E_N^* \left[ 1 \{Y X^T \beta < 0 \} \right] - E_N \left[ 1 \{Y X^T \beta < 0 \} \right] \times \left( N \left( E_N^* \left[ (Y - X^T \hat{\beta})^2 \right] - E_N^* \left[ (Y - X^T \hat{\beta})^2 \right] \right) \right) \right\}
\]

• \( E_N^* \) denotes the expectation for the bootstrap distribution
Cud-Bound Level Sets (n=30)

Three Point Dist.
Computational Issues

\[ \sup_{\beta \in R^q} \left\{ \left[ E_N^* \left[ 1 \{ Y X^T \beta < 0 \} \right] - E_N \left[ 1 \{ Y X^T \beta < 0 \} \right] \right] \times \right. \\
\left. g \left( N \left( E_N^* \left[ (Y - X^T \hat{\beta})^2 \right] - E_N^* \left[ (Y - X^T \beta)^2 \right] \right) \right) \right\} \]

- Partition \( R^q \) into equivalence classes defined by the 2N possible values of the first term.
- Each equivalence class, \( M_i \), can be written as a set of \( \beta \) satisfying linear constraints.
- The first term is constant on \( M_i \).
Computational Issues

\[
\begin{align*}
\sup_{\beta \in \mathcal{R}^q} & \left\{ \left[ E_N^* [1 \{ Y X^T \beta < 0 \}] - E_N [1 \{ Y X^T \beta < 0 \}] \right] \times \\
g \left( N \left( E_N^* [(Y - X^T \hat{\beta})^2] - E_N^* [(Y - X^T \beta)^2] \right) \right) \right\} \\
\end{align*}
\]

can be written as

\[
\max_i \left\{ C(\mathcal{M}_i) \times \\
g \left( N \left( E_N^* [(Y - X^T \hat{\beta})^2] - \inf_{\beta \in \mathcal{M}_i} E_N^* [(Y - X^T \beta)^2] \right) \right) \right\}
\]

since \( g \) is non-decreasing.
Computational Issues

• Reduced the problem to the computation of at most $2N$ mixed integer quadratic programming problems.

• Using commercial solvers (e.g. CPLEX) the CUD bound can be computed for moderately sized data sets in a few minutes on a standard desktop (2.8 GHz processor 2GB RAM).
## Comparisons, 95% PI

<table>
<thead>
<tr>
<th>Data</th>
<th>CUD</th>
<th>BS</th>
<th>M</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magic</td>
<td>1.0</td>
<td>.92</td>
<td>.98</td>
<td>.99</td>
</tr>
<tr>
<td>Mamm.</td>
<td>1.0</td>
<td>.68</td>
<td>.43</td>
<td>.98</td>
</tr>
<tr>
<td>Ion.</td>
<td>1.0</td>
<td>.61</td>
<td>.76</td>
<td>.99</td>
</tr>
<tr>
<td>Donut</td>
<td>1.0</td>
<td>.88</td>
<td>.63</td>
<td>.94</td>
</tr>
<tr>
<td>3-Pt</td>
<td>.97</td>
<td>.83</td>
<td>.90</td>
<td>.75</td>
</tr>
<tr>
<td>Balance</td>
<td>.95</td>
<td>.91</td>
<td>.61</td>
<td>.99</td>
</tr>
<tr>
<td>Liver</td>
<td>1.0</td>
<td>.96</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Sample size = 30  (1000 data sets)
## Comparisons, Length of PI

<table>
<thead>
<tr>
<th>Data</th>
<th>CUD</th>
<th>BS</th>
<th>M</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Magic</td>
<td>.60</td>
<td>.31</td>
<td>.28</td>
<td>.46</td>
</tr>
<tr>
<td>Mamm.</td>
<td>.46</td>
<td>.53</td>
<td>.32</td>
<td>.42</td>
</tr>
<tr>
<td>Ion.</td>
<td>.42</td>
<td>.43</td>
<td>.30</td>
<td>.50</td>
</tr>
<tr>
<td>Donut</td>
<td>.47</td>
<td>.59</td>
<td>.32</td>
<td>.41</td>
</tr>
<tr>
<td>3-Pt</td>
<td>.38</td>
<td>.48</td>
<td>.32</td>
<td>.46</td>
</tr>
<tr>
<td>Balance</td>
<td>.38</td>
<td>.09</td>
<td>.29</td>
<td>.48</td>
</tr>
<tr>
<td>Liver</td>
<td>.62</td>
<td>.37</td>
<td>.33</td>
<td>.49</td>
</tr>
</tbody>
</table>

Sample size=30 (1000 data sets)
Intuition

In large samples

\[
\sup_{\beta \in \mathbb{R}^q} \left\{ \sqrt{N} \left( E_N[1\{Y X^T \beta < 0 \}] - E[1\{Y X^T \beta < 0 \}] \right) \times g \left( N(E_N[(Y - X^T \tilde{\beta})^2] - E_N[(Y - X^T \beta)^2]) \right) \right\}
\]

behaves like

\[
\sup_{\gamma \in \mathbb{R}^q} \left[ X(\gamma) \right] g \left( Z^T \gamma - \frac{1}{2} \gamma^T \Sigma \gamma \right)
\]

\[\gamma = \sqrt{N}(\beta - \tilde{\beta})\]
Intuition

The large sample distribution is the same as the distribution of

$$\sup_{\gamma \in R^d} [X(\gamma)] g \left( Z^T \gamma - \frac{1}{2} \gamma^T \Sigma \gamma \right)$$

where

$$\Sigma = E \left[ XX^T \right], \ Z \sim N(0, \sigma^2 \Sigma), \ X(\gamma) \sim N(0, p_{\gamma}(1 - p_{\gamma}))$$

$$p_{\gamma} = P[X^T \tilde{\beta} < 0] + P[X^T \tilde{\beta} = 0, \ YX^T \gamma < 0]$$
Intuition

If \( P[X^T \tilde{\beta} \neq 0] = 1 \)

then the distribution is approximately that of a

\[
N(0, p(1 - p)), \quad p = P[X^T \tilde{\beta} < 0]
\]

(limiting distribution for binomial, as expected).
Intuition

If \( P[X^T \tilde{\beta} = 0] = 1 \)
the distribution is approximately that of

\[
\sup_{\gamma \in \mathcal{G}} N(0, P[Y X^T \gamma < 0] P[Y X^T \gamma \geq 0])
\]

where

\[
\mathcal{G} = \{ \gamma : (\gamma - \Sigma^{-1} Z)^T \Sigma (\gamma - \Sigma^{-1} Z) \leq B \}
\]

\[
\sqrt{N} (\hat{\beta} - \tilde{\beta}) = \Sigma_n^{-1} Z_n
\]
Discussion

• Further reduce the conservatism of the CUD-bound.
  – Replace $\tilde{\beta}$ by other quantities.
  – Other surrogates (exponential, logit)
• Construct a principle for minimizing the length of the conservative PI?
• The real goal is to produce PIs for the Value of a policy.
The simplest **Dynamic treatment regime** (e.g. policy) is a decision rule if there is only one stage of treatment

1 Stage for each individual

\[ X_1, A_1, X_2 \]

\( X_j \): Observation available at j\(^{th} \) stage

\( A_j \): Action at j\(^{th} \) stage (usually a treatment)

Primary Outcome:

\[ Y = r(X_1, X_2) \]
Goal:

Construct decision rules that input patient information and output a recommended action; these decision rules should lead to a maximal mean \( Y \).

In future one selects action:  \( a_1 = d(X_1) \)
Single Stage

• Find a confidence interval for the mean outcome if a particular estimated policy (here one decision rule) is employed.
• Treatment $A$ is randomized in $\{-1,1\}$.
• Suppose the decision rule is of form

$$\hat{d}(X_1) = \text{sign}(\hat{\beta}^T X_1)$$

• *We do not assume the optimal decision boundary is linear.*
Single Stage

Mean outcome following this policy is \( V(\hat{\beta}) \)

\[
V(\beta) = E \left[ E[Y | X_1, A = \text{sign}(X_1^T \beta)] \right] \\
= E \left[ \frac{Y}{p(A | X_1)} I\{AX_1^T \beta > 0\} \right]
\]

\( p(A_1 | X_1) \) is the randomization probability
STAR*D "Sequenced Treatment to Relieve Depression"

Preference Treatment Two Intermediate Outcome Preference Treatment Three

Follow-up

Remission

Augment

R

CIT + BUS

CIT + BUP-SR

Non-remission

Augment

R

CIT

Switch

R

Bup-SR

VEN

SER

Switch

R

L2-Tx +THY

L2-Tx +LI

MIRT

NTP
Oslin ExTENd

Early Trigger for Nonresponse

Random assignment:

Late Trigger for Nonresponse

Random assignment:

8 wks Response

Nonresponse

Random assignment:

Random assignment:

8 wks Response

Nonresponse

Random assignment:
This seminar can be found at:
http://www.stat.lsa.umich.edu/~samurphy/seminars/UFlorida01.09.09.ppt

Email Eric or me with questions or if you would like a copy of the associated paper:
laber@umich.edu  or samurphy@umich.edu
Bias of Common $\text{err}(\hat{f})$ on Three Point Example