Discussion on "Adaptive confidence intervals for the test error in classification" by Eric B. Laber and Susan A. Murphy

Peter Bühlmann *

May 20, 2011

I congratulate the authors for their excellent contribution covering practical and non-standard mathematical aspects of inference. Quantifying uncertainty belongs to the core of statistics: the standard (and overly simplified) measure for classification accuracy is an estimated test set or generalization error. Laber and Murphy address a much more appropriate and more challenging task, namely to construct accurate confidence intervals for the test set error. I very much agree that quantifying accuracy should be pursued with measures taking uncertainty into account.

Laber and Murphy present thorough mathematical analysis and arguments showing that with low sample size, the asymptotic framework should be chosen carefully. I concur with their views and mathematical argumentation. In the following, I am trying to make a few selective cross-connections to related issues which have been worked out in the past.

1 The local view, bagging and subsampling

One of the key issues in L&M is a careful analysis when the points are near the classification boundary, formalized as distinguishing whether $\mathbb{P}[X^t\beta^* = 0]$ is strictly positive or not. The idea is then to look more closely at what happens at the boundary $X^t\beta^* = 0$. The approach considering "local alternatives" (L&M Section 3.3) is instructive and I am following up on it by re-using a toy example from Bühlmann and Yu (2002).

Consider a scenario where we have a general estimator $\hat{\theta}_n$ for an unknown parameter $\theta_n^* = \theta^* + \Gamma/\sqrt{n}$ which is "moving" as sample size n changes, see formula (11) in L&M. For simplicity, assume that the value of the parameter is

*Peter Bühlmann is Professor, Seminar for Statistics, ETH Zürich, CH-8092 Zürich, Switzerland.

1-dimensional ($p = 1$). Consider the indicator decision (or classification) function

$$\hat{d} = \hat{d}_n = 1(\hat{\theta}_n < \theta^*) = 1(\sqrt{n}(\hat{\theta}_n - \theta_n^*) < -\Gamma).$$

Assume that we are in a nice situation where

$$\sqrt{n}(\hat{\theta}_n - \theta_n^*) \Rightarrow \mathcal{N}(0, \sigma_\infty^2) \ (n \to \infty) \tag{1}$$

for some asymptotic variance σ_∞^2. We can then rewrite the estimator as (see also Section 2 in L&M)

$$\hat{d}_n = 1(\hat{\theta}_n < \theta^*) = 1(\sqrt{n}(\hat{\theta}_n - \theta_n^*)/\sigma_\infty < -\Gamma/\sigma_\infty) \approx 1(Z < -\Gamma/\sigma_\infty),$$

where $Z \sim \mathcal{N}(0, 1)$. For any Γ (including $\Gamma = 0$, which is slightly different from formula (11) in L&M), the indicator decision function does not converge to a constant since the variance is not converging to zero:

$$\begin{aligned}
&\mathbb{E}[\hat{d}_n] = \mathbb{E}[1(\hat{\theta}_n < \theta^*)] \to \Phi(-\Gamma/\sigma_\infty) \ (n \to \infty), \\
&\mathrm{Var}(\hat{d}_n) = \mathrm{Var}(1(\hat{\theta}_n < \theta^*)) \to \Phi(-\Gamma/\sigma_\infty)(1 - \Phi(-\Gamma/\sigma_\infty)) \ (n \to \infty),
\end{aligned}$$

where $\Phi(\cdot)$ is the cdf of $\mathcal{N}(0, 1)$. We simply recover here aspects of what Laber and Murphy have discussed in detail.

Next, let us look at the bootstrap. The bootstrap is typically consistent for asymptotic normally distributed estimators (Giné and Zinn, 1990): we assume

$$\sqrt{n}(\hat{\theta}_n^{(b)} - \hat{\theta}_n) \Rightarrow \mathcal{N}(0, \sigma_\infty^2) \ (n \to \infty) \text{ in probability.} \tag{2}$$

Thus, the bootstrapped indicator decision function becomes:

$$\begin{aligned}
1(\hat{\theta}_n^{(b)} < \theta^*) &= 1(\sqrt{n}(\hat{\theta}_n^{(b)} - \hat{\theta}_n)/\sigma_\infty < \sqrt{n}(\theta^* - \hat{\theta}_n)/\sigma_\infty) \\
&= 1(\sqrt{n}(\hat{\theta}_n^{(b)} - \hat{\theta}_n)/\sigma_\infty < -\sqrt{n}(\hat{\theta}_n - \theta_n^*)/\sigma_\infty - \Gamma/\sigma_\infty)
\end{aligned}$$

Now, let us look at the first two moments again, with respect to the bootstrap distribution:

$$\begin{aligned}
& \mathbb{E}^{(b)}[1(\hat{\theta}_n^{(b)} < \theta^*)] \\
= \ & \mathbb{P}^{(b)}[\sqrt{n}(\hat{\theta}_n^{(b)} - \hat{\theta}_n)/\sigma_\infty < -\sqrt{n}(\hat{\theta}_n - \theta_n^*)/\sigma_\infty - \Gamma/\sigma_\infty] \\
\approx \ & \Phi(-\sqrt{n}(\hat{\theta}_n - \theta_n^*)/\sigma_\infty - \Gamma/\sigma_\infty) \approx \Phi(-Z - \Gamma/\sigma_\infty),
\end{aligned} \tag{3}$$

where $Z \sim \mathcal{N}(0, 1)$. The first approximation is due to bootstrap consistency in (2) while the second approximation holds because of (1). For the variance, we then obtain

$$\mathrm{Var}^{(b)}(1(\hat{\theta}_n^{(b)} < \theta^*)) \approx \Phi(-Z - \Gamma/\sigma_\infty)(1 - \Phi(-Z - \Gamma/\sigma_\infty)).$$

The bootstrap is not picking up the first two moments in a consistent way, that is:

$$\frac{\mathbb{E}[1(\hat{\theta}_n < \theta^*)]}{\mathbb{E}^{(b)}[1(\hat{\theta}_n^{(b)} < \theta^*)]} - 1 \neq o_P(1), \quad \frac{\mathrm{Var}(1(\hat{\theta}_n < \theta^*))}{\mathrm{Var}^{(b)}(1(\hat{\theta}_n^{(b)} < \theta^*))} - 1 \neq o_P(1),$$

where the first statement about expectations only holds for $\Gamma \neq 0$. Thus, clearly, the bootstrap does not provide confidence intervals for $\mathbb{E}[\hat{d}_n]$ or similar quantities. However, the bootstrap can be used to stabilize.

Instead of using the estimator $\hat{d} = 1(\hat{\theta}_n < \theta^*)$, we can use bagging (Breiman, 1996). Consider the bagged version which is simply the bootstrap expectation:

$$\hat{d}_{\mathrm{bag}} = \mathbb{E}^{(b)}[1(\hat{\theta}_n^{(b)} < \theta^*)] \approx \Phi(-Z - \Gamma/\sigma_\infty),$$

see formula (3). Figure 1 shows the asymptotic behavior of the decision function $\hat{d}$ and the (substantial) smoothing effect when using $\hat{d}_{\mathrm{bag}}$, as a function of the random variable $Z \sim \mathcal{N}(0,1)$ for the value $\Gamma = 0$ (which corresponds to the most unstable point with maximal variance for $\hat{d}$) and using w.l.o.g. $\theta^* = 0$. The figure may be compared to Figure 1 in L&M.

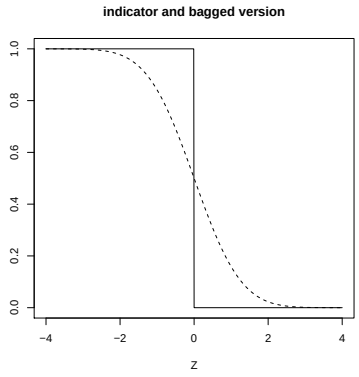

Figure 1: Asymptotic behavior of $\hat{d} \approx 1(Z < 0)$ and $\hat{d}_{\mathrm{bag}} \approx \Phi(-Z)$ as a function of $Z \sim \mathcal{N}(0,1)$, for $\Gamma = 0$ and w.l.o.g. $\theta^* = 0$.

The smoothing operation (see Figure 1) introduces some bias but reduces variance:

$$\mathbb{E}[\hat{d}_{\mathrm{bag}}] \approx \mathbb{E}[\Phi(-Z - \Gamma/\sigma_\infty)], \;\; \mathbb{E}[\hat{d}] \approx \Phi(-\Gamma/\sigma_\infty),$$
$$\mathrm{Var}(\hat{d}_{\mathrm{bag}}) \approx \mathrm{Var}(\Phi(Z - \Gamma/\sigma_\infty)), \;\; \mathrm{Var}(\hat{d}) \approx \Phi(-\Gamma/\sigma_\infty)(1 - \Phi(-\Gamma/\sigma_\infty)).$$

The easiest comparison is for $\Gamma = 0$ which corresponds to the most "unstable" case where $\hat{d}$ has highest variance. Then, we can use the simple fact that $\Phi(-Z)$ is a

Uniform($[0,1]$) random variable and hence:

$$\mathbb{E}[\hat{d}_{\mathrm{bag}}] \approx 1/2, \;\; \mathbb{E}[\hat{d}] \approx 1/2,$$
$$\mathrm{Var}(\hat{d}_{\mathrm{bag}}) \approx 1/12, \;\; \mathrm{Var}(\hat{d}) \approx 1/4.$$

In words, this means: there is approximately no bias of the bagged decision $\hat{d}_{\mathrm{bag}}$ while it enjoys a variance reduction of a factor 3. One can compute the mean squared error (for the target $\mathbb{E}[\hat{d}]$): the bagged procedure has lower MSE than the non-bagged estimator for a large range where $|\Gamma| \leq 2.3$ and the biggest gain (by a factor 3) is at the most unstable value where $\Gamma = 0$. The whole presented analysis hinges on asymptotic normality and bootstrap consistency in (1) and (2).

For more complicated estimators $\hat{\theta}_n$ where (1) and (2) do not hold, I do not know how the argument above carries through. From a methodological point of view, the bootstrap is still doing some sort of smoothing of the indicator (decision) function. Bühlmann and Yu (2002) have therefore looked at subsampling, using subsample size $m < n$, instead of bootstrap resampling. The analogue of $\hat{d}_{\mathrm{bag}}$ is then

$$\hat{d}_{\mathrm{subag}(m)} = \mathbb{E}^{(s)}[1(\hat{\theta}_n^{\mathrm{subs}(m)} < \theta^*)]$$

where we aggregate over subsampled estimators ($\mathbb{E}^{(s)}$ is with respect to subsampling, and in fact is a finite sum over all $\binom{n}{m}$ different subsamples of size m). One can then prove that there is again a substantial gain in terms of MSE when using $\hat{d}_{\mathrm{subag}(m)}$ instead of $\hat{d}$. A generic and good choice of the subsample size is $m = \lfloor n/2 \rfloor$. For further details we refer to Bühlmann and Yu (2002).

2 Sample splitting

As indicated above, subsampling with subsample size $m = \lfloor n/2 \rfloor$ has the potential to stabilize and improve the decision function $\hat{d}$. Subsampling with such a subsample size is very closely related to sample splitting with two half-samples indexed by $I_1 = \{1, \ldots, \lfloor n/2 \rfloor\}$ and $I_2 = \{1, \ldots, n\} \setminus I_1$. One can pursue a very different route with sample splitting than what we discussed before.

2.1 P-values and confidence intervals based on sample splitting

L&M make a connection to Yang (referenced by L&M) and raise the issue that Yang's approach lacks rigorous mathematical justification. Other work by van de Wiel et al. (2009) and Meinshausen et al. (2009) present mathematical theory when using (multiple) sample splitting for constructing p-values.

The problem studied by van de Wiel et al. (2009) is to test whether two methods exhibit a significant difference in terms of their misclassification error, a question closely related to the results in L&M (see also Section 6 in L&M). One can use the first half-sample I_1 to train two different classifiers and then use I_2 to test on $|I_2|$ sample points the difference in performance for misclassification leading to a p-value (conditional on training data from I_1). The approach suffers from the problem that the resulting p-value depends very heavily on the (random) sample split which is used and hence, the result is not really reproducible. Aggregating over multiple (random) sample splits is a useful idea (van de Wiel et al., 2009; Meinshausen et al., 2009), and we briefly outline in (5) below how to aggregate p-values from such multiple sample splits.

As an alternative to the approach by Laber and Murphy, we could use sample splitting as follows. On I_1, we train the method $\hat{f} = \hat{f}_{I_1}$ and build the classifier $\mathrm{sign}(\hat{f}_{I_1}(x))$ for a new covariate x. We can then look at the performance on the other (test) sample:

$$\sum_{i \in I_2} 1(\mathrm{sign}(\hat{f}_{I_1}(X_i)) \neq Y_i). \tag{4}$$

Conditionally on the data from I_1, the expression in (4) has a Binomial$(|I_2|, \pi_{I_1})$ distribution where $\pi_{I_1} = \mathbb{P}[\mathrm{sign}(\hat{f}_{I_1}(X)) \neq Y]$, where (X, Y) is a new test data point (e.g. a sample point from I_2) and the probability is conditional on the samples from I_1.

This then allows one to do significance testing. The null- and alternative hypotheses are formalized when conditioning on training data $I \subset \{1, \ldots, n\}$ with $|I| = \lfloor n/2 \rfloor$:

$$H_0 : \ \pi_I \leq \pi_0 \text{ for all } I \subset \{1, \ldots, n\} \text{ with } |I| = \lfloor n/2 \rfloor.$$

Usually, we are interested in one-sided testing and the alternative would be H_A : $\pi_I > \pi_0$ for some (training) set I. We note that π_0 is a fixed value, not depending on I. Using the summary statistics in (4), with its corresponding Binomial$(|I_2|, \pi_0)$ distribution under H_0, we obtain a p-value

$$p_{I_1}(\pi_0)$$

which is conditional on the training half-sample I_1.

As indicated above, this p-value might be very sensitive to the sample split and its corresponding sets I_1 and $I_2 = \{1, \ldots, n\} \setminus I_1$. The remedy is to use B (random) sample splits yielding p-values

$$p_{I_1^{(j)}}(\pi_0), \ j = 1, \ldots, B,$$

where B is "large" such as $B = 100 - 500$. We can aggregate these (dependent) p-values using empirical quantiles. Denote by

$$q^{\gamma}(\pi_0) = q^{\gamma}_{I_1^{(1)},\ldots,I_1^{(B)}}(\pi_0) = \gamma\text{-quantile of } \{p_{I_1^{(j)}}(\pi_0)/\gamma;\ j = 1,\ldots,B\}. \qquad (5)$$

Then, $q^{\gamma}(\pi_0)$ controls the type I error:

$$\mathbb{P}_{H_0}[q^{\gamma}(\pi_0) \leq \alpha] \leq \alpha \ \ (0 < \alpha < 1),$$

corresponding to the rejection of H_0 if and only if $q^{\gamma}(\pi_0) \leq \alpha$. We note that the aggregation of the p-values with the γ-quantile involves an additional factor $1/\gamma$; e.g., when using the median with $\gamma = 1/2$, we have to multiply the p-values $p_{I_1^{(j)}}(\pi_0)$ by the factor 2 in order to obtain error control. The proof of such p-value aggregation under no additional assumptions (other than that the p-value has a Uniform$([0,1])$ distribution under the null-hypothesis) can be adopted from Theorem 3.1 in Meinshausen et al. (2009). The latter reference also provides a method to estimate a good value of γ while still providing error control. When making additional assumptions, one can drop the correction factor $1/\gamma$, see van de Wiel et al. (2009).

From the p-values $q^{\gamma}(\pi_0)$ we can construct a confidence interval via duality:

$$I(1-\alpha) = \{\pi_0;\ q^{\gamma}(\pi_0) > \alpha\} \ \ (0 < \alpha < 1).$$

Thus, we have constructed a confidence interval for "some kind of" conditional misclassification error. The words "some kind of" refers to the issue that we are conditioning on all subsets $I \in \{I_1^{(1)},\ldots,I_1^{(B)}\}$ which arise when doing B (random) sample splitting operations. And this may be an unusual view point and an issue which should be addressed in a more elegant and aesthetical way.

2.2 Pros and cons, and some remarks

The confidence interval $I(1-\alpha)$ does not require any asymptotic approximations, it is applicable in e.g. high-dimensional problems with $p \gg n$, and it is very generic and easy to compute, i.e., as easy as bootstrapping or subsampling where we only need to program an additional outer loop which repeats the same calculations B times. Moreover, such an approach enjoys the conceptual advantage of separating clearly between training and test set, whereas the bootstrap as employed in L&M involves the data which has been used for training the classifier. Does this lead to overly optimistic results, especially in more complex problems? And would an out-of-bag bootstrap (Breiman, 2001) be beneficial?

The drawbacks and potential disadvantages of the sample splitting approach are: (i) it operates on training sample size $\lfloor n/2 \rfloor$ which, despite aggregation afterwards, is a potential loss of efficiency; (ii) the p-value aggregation in (5) is conservative (note the additional factor $1/\gamma$ in (5)) and hence again, a potential loss of power.

It is worth pointing out that subsampling and sample splitting often lead to "stable" results: in various contexts of high-dimensional problems, we have found that subsampling and sample splitting can be tremendously useful for the tasks of structure estimation (Meinshausen and Bühlmann, 2010) or assigning (conservative) p-values in generalized regression (Meinshausen et al., 2009). The gain in stability when randomizing over different subsamples (and/or subsets of the feature space) is only partially understood (Lin and Jeon, 2006; Meinshausen and Bühlmann, 2010): nevertheless, Leo Breiman, the "inventor" of this kind of thinking, has provided utterly convincing examples that these methods have the potential to provide very competitive answers and results (Breiman, 1996, 2001), perhaps in a much broader range than his fundamental contributions in "improving" regression or classification methods.

3 Conclusions

Subsampling has an interesting potential to stabilize the indicator or decision function as outlined in Section 1. The related concept of sample splitting can be used – in principle – to construct confidence intervals for the misclassification error or for many other problems about assigning uncertainty, see Section 2.

Laber and Murphy have presented an impressive path of ideas and results. My remarks do not diminish in any sense their beautiful contribution, and they should be interpreted as an attempt to provide some complementary thoughts about the issue of constructing uncertainty measures for the misclassification error or other quantities of interest.

References

BREIMAN, L. (1996). Bagging predictors. *Machine Learning* **24** 123–140.

BREIMAN, L. (2001). Random forests. *Machine Learning* **45** 5–32.

BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *Annals of Statistics* **30** 927–961.

GINÉ, E. and ZINN, J. (1990). Bootstrapping general empirical measures. *Annals of Probability* **18** 851–869.

LIN, Y. and JEON, Y. (2006). Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association* **101** 578–590.

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society Series B* **72** 417–473.

MEINSHAUSEN, N., MEIER, L. and BÜHLMANN, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association* **104** 1671–1681.

VAN DE WIEL, M., BERKHOF, J. and VAN WIERINGEN, W. (2009). Testing the prediction error difference between two predictors. *Biostatistics* **10** 550–560.