Comment

Peter BÜHLMANN

I congratulate the authors for their excellent contribution covering practical and nonstandard 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: constructing accurate confidence intervals for the test set error. I very much agree with them that quantifying accuracy should be pursued with measures taking uncertainty into account.

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

1. THE LOCAL VIEW, BAGGING, AND SUBSAMPLING

One of the key issues that Laber and Murphy address is the need for careful analysis when the points are near the classification boundary, formalized as distinguishing whether or not $\mathbb{P}[X^t\beta^* = 0]$ is strictly positive. The idea is then to look more closely at what happens at the boundary $X^t\beta^* = 0$. The approach considering "local alternatives" (section 3.3) is instructive, and I follow up on it by reusing a toy example from Bühlmann and Yu (2002).

Consider a scenario where we have a general estimator $\hat{\theta}_n$ for an unknown parameter $\theta_n^* = \theta^* + \Gamma/\sqrt{n}$ that is "moving" as the sample size n changes [see the authors' formula (11)]. For simplicity, assume that the value of the parameter is one-dimensional ($p = 1$). Consider the indicator decision (or classification) function

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

Assume that we are in a nice situation where

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

for some asymptotic variance σ_∞^2. We can then rewrite the estimator as (see also section 2)

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

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

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

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

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

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

Thus the bootstrapped indicator decision function becomes

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

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

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

where $Z \sim \mathcal{N}(0, 1)$. The first approximation is due to bootstrap consistency in (2), whereas the second approximation holds because of (1). Then, for the variance,

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

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

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

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

where the first statement about expectations holds only 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 $\hat{d} = 1(\hat{\theta}_n < \theta^*)$, one can use bagging (Breiman 1996). Consider the bagged version, which is simply the bootstrap expectation,

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

see formula (3). Figure 1 shows the asymptotic behavior of the decision function $\hat{d}$ and the (substantial) smoothing effect when using $\hat{d}_{\text{bag}}$ 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 $\theta^* = 0$ without loss of generality. This figure may be compared with the authors' Figure 1.

Peter Bühlmann is Professor, Seminar for Statistics, ETH Zürich, CH-8092 Zürich, Switzerland (E-mail: *buhlmann@stat.math.ethz.ch*).

Journal of the American Statistical Association
September 2011, Vol. 106, No. 495, Theory and Methods
DOI: 10.1198/jasa.2011.tm11315

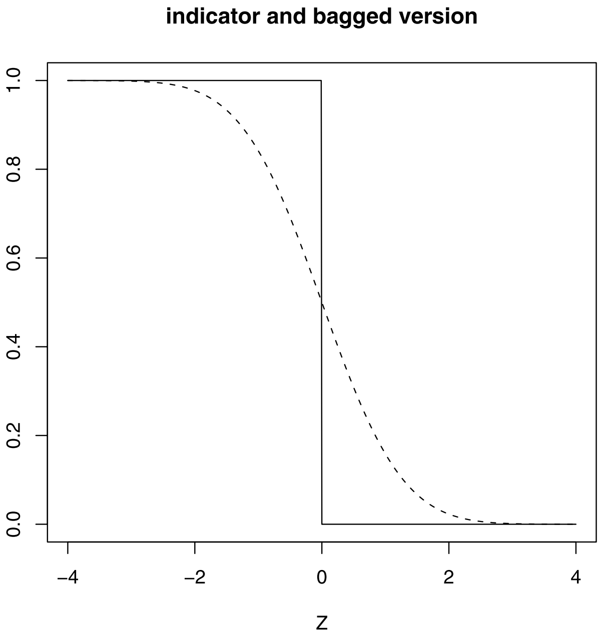

Figure 1. Asymptotic behavior of $\hat{d} \approx 1(Z < 0)$ and $\hat{d}_{\text{bag}} \approx \Phi(-Z)$ as a function of $Z \sim \mathcal{N}(0, 1)$, for $\Gamma = 0$ and, without loss of generality, $\theta^* = 0$.

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

$$\begin{aligned}
\mathbb{E}[\hat{d}_{\text{bag}}] &\approx \mathbb{E}[\Phi(-Z - \Gamma/\sigma_\infty)],\\
\mathbb{E}[\hat{d}] &\approx \Phi(-\Gamma/\sigma_\infty),\\
\text{Var}(\hat{d}_{\text{bag}}) &\approx \text{Var}(\Phi(Z - \Gamma/\sigma_\infty)),\\
\text{Var}(\hat{d}) &\approx \Phi(-\Gamma/\sigma_\infty)(1 - \Phi(-\Gamma/\sigma_\infty)).
\end{aligned}$$

The easiest comparison is for $\Gamma = 0$, which corresponds to the most "unstable" case where $\hat{d}$ has the greatest variance. Then, based on the simple fact that $\Phi(-Z)$ is a Uniform([0, 1]) random variable,

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

In words, this means that there is approximately no bias of the bagged decision $\hat{d}_{\text{bag}}$, whereas it enjoys a variance reduction of a factor 3. The mean squared error (MSE) can be computed for the target $\mathbb{E}[\hat{d}]$). The bagged procedure has smaller MSE than the nonbagged estimator for a large range where $|\Gamma| \le 2.3$, and the biggest gain (by a factor of 3) is at the most unstable value where $\Gamma = 0$. The entire 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 foregoing argument carries through. From a methodological standpoint, the bootstrap is still doing some sort of smoothing of the indicator (decision) function. Bühlmann and Yu (2002) have looked at subsampling, using subsample size $m < n$, instead of bootstrap resampling. The analog of $\hat{d}_{\text{bag}}$ is then

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

where the aggregation is 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 $\hat{d}_{\text{subag}(m)}$ instead of $\hat{d}$. A generic and good choice of the subsample size is $m = \lfloor n/2 \rfloor$. [For further details, see Bühlmann and Yu (2002).]

2. SAMPLE SPLITTING

As indicated earlier, 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 that discussed earlier.

2.1 *p*-Values and Confidence Intervals Based on Sample Splitting

Laber and Murphy 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, Berkhof, and van Wieringen (2009) and Meinshausen, Meier, and Bühlmann (2009) presents mathematical theory when using (multiple) sample splitting for constructing p-values.

The problem studied by van de Wiel, Berkhof, and van Wieringen (2009) is to test whether two methods exhibit a significant difference in terms of their misclassification error, a question closely related to Laber and Murphy's results (see their section 6). 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). This approach has the problem that the resulting p-value depends very heavily on the (random) sample split used, and thus the result is not really reproducible. Aggregating over multiple (random) sample splits is a useful idea (van de Wiel, Berkhof, and van Wieringen 2009; Meinshausen, Meier, and Bühlmann 2009), and I briefly outline in (5) how to aggregate p-values from such multiple sample splits.

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

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

Conditionally on the data from I_1, the expression in (4) has a Binomial$(|I_2|, \pi_{I_1})$ distribution where $\pi_{I_1} = \mathbb{P}[\text{sign}(\hat{f}_{I_1}(X)) \neq Y]$, 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 enables significance testing. The null and alternative hypotheses are formalized when conditioning on training data $I \subset \{1, \ldots, n\}$ with $|I| = \lfloor n/2 \rfloor$:

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

Usually, the interest lies in one-sided testing, and the alternative would be $H_A : \pi_I > \pi_0$ for some (training) set I. Note that

π_0 is a fixed value that does not depend 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 earlier, this p-value might be highly 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

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

where B is "large" such as $B = 100$–500. These (dependent) p-values can be aggregated using empirical quantiles. Write

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

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

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

corresponding to the rejection of H_0 if and only if $q^{\gamma}(\pi_0) \le \alpha$. We note that the aggregation of the p-values with the γ-quantile involves an additional factor, $1/\gamma$; for example, when using the median with $\gamma = 1/2$, the p-values $p_{I_1^{(j)}}(\pi_0)$ must be multiplied by the factor of 2 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 of Meinshausen, Meier, and Bühlmann (2009). The latter reference also provides a method for estimating a good value of γ while still providing error control. When making additional assumptions, the correction factor $1/\gamma$ can be dropped (see van de Wiel, Berkhof, and van Wieringen 2009).

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

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

Thus, a confidence interval is constructed for "some kind of" conditional misclassification error. The phrase "some kind of" refers to the issue of conditioning on all subsets $I \in \{I_1^{(1)}, \ldots, I_1^{(B)}\}$ that arise when performing B (random) sample splitting operations. This may be an unusual viewpoint, and this issue should be addressed in a more elegant and aesthetically pleasing 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 for example, high-dimensional problems with $p \gg n$, and it is very generic and easy to compute—as easy as bootstrapping or subsampling, which only requires programming an additional outer loop that repeats the same calculations B times. Moreover, such an approach enjoys the conceptual advantage of clearly separating training and test sets, whereas the bootstrap as used by Laber and Murphy involves the data that were used for training the classifier. Does this lead to overly optimistic results, especially in more complex problems? Would an out-of-bag bootstrap (Breiman 2001) be beneficial?

The sample splitting approach has two drawbacks and potential disadvantages. First, the approach operates on training sample size $\lfloor n/2 \rfloor$, which, despite aggregation afterward, is a potential loss of efficiency. Second, the p-value aggregation in (5) is conservative [note the additional factor $1/\gamma$ in (5)], a potential loss of power.

Of note, subsampling and sample splitting often lead to "stable" results. In various contexts of high-dimensional problems, subsampling and sample splitting can be tremendously useful for performing structure estimation (Meinshausen and Bühlmann 2010) or assigning (conservative) p-values in generalized regression (Meinshausen, Meier, and Bühlmann 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 indicating 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 interesting potential for stabilizing the indicator or decision function, as outlined in Section 1. In principle, the related concept of sample splitting can be used to construct confidence intervals for the misclassification error or for many other problems related to 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 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.

ADDITIONAL REFERENCES

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140. [916,918]

——— (2001), "Random Forests," *Machine Learning*, 45, 5–32. [918]

Bühlmann, P., and Yu, B. (2002), "Analyzing Bagging," *The Annals of Statistics*, 30, 927–961. [916,917]

Giné, E., and Zinn, J. (1990), "Bootstrapping General Empirical Measures," *The Annals of Probability*, 18, 851–869. [916]

Lin, Y., and Jeon, Y. (2006), "Random Forests and Adaptive Nearest Neighbors," *Journal of the American Statistical Association*, 101, 578–590. [918]

Meinshausen, N., and Bühlmann, P., (2010), "Stability Selection" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 72, 417–473. [918]

Meinshausen, N., Meier, L., and Bühlmann, P. (2009), "p-Values for High-Dimensional Regression," *Journal of the American Statistical Association*, 104, 1671–1681. [917,918]

van de Wiel, M., Berkhof, J., and van Wieringen, W. (2009), "Testing the Prediction Error Difference Between Two Predictors," *Biostatistics*, 10, 550–560. [917,918]