Discussion of *Adaptive confidence intervals for the test error in classification* by Laber and Murphy

Richard J. Samworth
Statistical Laboratory
University of Cambridge
r.samworth@statslab.cam.ac.uk

February 20, 2011

1 Introduction

Let me begin by congratulating the authors on the substantial progress that they have made on an important and challenging problem in classification. They provide penetrating insight into the reasons why the most obvious methods for constructing confidence intervals for the test error have poor coverage properties, and propose an innovative and effective solution that can readily be adopted in practice.

The focus of the paper is on *linear* classifiers. While this is certainly an important family, it of course has limitations in terms of the complexity of the boundaries that can be handled. Here, we make a first attempt at extending the scope of the methodology in the paper to other types of classifiers.

2 Weighted nearest neighbor classifiers

In this section, we consider weighted nearest neighbor classifiers, which are especially attractive here because of their relative simplicity to compute (an important feature in the context of constructing computationally intensive confidence intervals). We will also see that they illustrate the important issues in attempting to generalize the Adaptive Confidence Intervals to other classifiers. We retain the notation used in the paper, but in addition define the following quantities. Let P_Y denote the marginal distribution of Y, and let $P_Y(Y = 1) = \pi = 1 - P_Y(Y = -1)$. Suppose further that the conditional distribution

of X given that $Y = r$ has a p-dimensional Lebesgue density f_r for $r = -1, 1$. Let $\|\cdot\|$ denote an arbitrary norm on $\mathbb{R}^p$, and for a fixed $x \in \mathbb{R}^p$, let $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ denote the permutation of the pairs in $\mathcal{T}$ such that $\|X_{(1)} - x\| \leq \ldots \leq \|X_{(n)} - x\|$. Note that the inequalities here are strict with probability 1. If $(w_{ni})_{i=1}^n$ denotes a set of weights satisfying $\sum_{i=1}^n w_{ni} = 1$ (we shall also assume throughout that $w_{ni} \geq 0$), then the *weighted nearest neighbor* classifier is

$$\hat{c}^{\mathrm{wnn}}(x) = \mathrm{sign}\biggl(\sum_{i=1}^n w_{ni} \mathbb{1}_{\{Y_{(i)}=1\}} - \frac{1}{2}\biggr). \tag{2.1}$$

This classifier was first studied by Royall (1966); see also Devroye, Györfi and Lugosi (1996). Of course, an important special case is the celebrated *k-nearest neighbor classifier* (Fix and Hodges, 1951; Cover and Hart, 1967; Hall, Park and Samworth, 2008), where $w_{ni} = \frac{1}{k}\mathbb{1}_{\{i \leq k\}}$. Recently, Samworth (2011) has shown that, under regularity conditions, the asymptotically optimal weighting scheme is to choose $k^* = \lfloor B^* n^{4/(p+4)} \rfloor$, and then set

$$w^*_{ni} = \begin{cases} \frac{1}{k^*}\bigl[1 + \frac{p}{2} - \frac{p}{2(k^*)^{2/p}}\{i^{1+2/p} - (i-1)^{1+2/p}\}\bigr] & \text{for } i = 1, \ldots, k^* \\ 0 & \text{for } i = k^*+1, \ldots, n. \end{cases} \tag{2.2}$$

An explicit expression for B^* is given in Samworth (2011). The discrete distribution on $\{1, \ldots, n\}$ defined by the asymptotically optimal weights decreases in a concave fashion when $p = 1$, in a linear fashion when $p = 2$ and in a convex fashion when $p \geq 3$.

Let $\mathcal{S} = \{x \in \mathbb{R}^p : \psi(x) = 0\}$, where $\psi = \pi f_1 - (1-\pi) f_{-1}$, so $\mathcal{S}$ represents the decision boundary of the Bayes classifier. Provided that the derivative of ψ does not vanish on $\mathcal{S}$, the set $\mathcal{S}$ is a $(p-1)$-dimensional sub-manifold of $\mathbb{R}^p$ (Guillemin and Pollack, 1974, p.21). Notice that this assumption ensures that the condition $P_X(X \in \mathcal{S}) = 0$ is satisfied.

The key insight of Laber and Murphy is that, due to the inability of asymptotic theory or the bootstrap to adequately capture the additional variability of the test error across training samples caused by its non-smoothness, one should bound the empirical process of the test error on the set of points close to the decision boundary of the Bayes classifier. To determine an appropriate partition of the domain of X into points close to, and far from, this decision boundary in our context, we seek a hypothesis test of $H_0 : x \in \mathcal{S}$ against $H_1 : x \notin \mathcal{S}$. Writing $S_n(x) = \sum_{i=1}^n w_{ni} \mathbb{1}_{\{Y_{(i)}=1\}}$ and $s_n^2 = \sum_{i=1}^n w_{ni}^2$, we have that $\mathrm{Var}\{S_n(x)\} \leq \frac{1}{4} s_n^2$, and the bound is good when $x \in \mathcal{S}$. We therefore propose to reject H_0

if

$$T_n(x) := \frac{\{S_n(x) - 1/2\}^2}{\frac{1}{4}s_n^2} > \frac{1}{a_n}.$$

An alternative, less conservative test, would reject H_0 if

$$\tilde{T}_n(x) := \frac{\{S_n(x) - 1/2\}^2}{S_n(x)\{1 - S_n(x)\}} > \frac{1}{a_n}.$$

The choice of a_n is entirely analogous to the corresponding choice in the paper since, under mild regularity conditions on the weights, $S_n(x)$ is asymptotically normal.

We now require appropriate upper and lower bounds for the empirical process of the test error on the set where we do not reject H_0. This is challenging because, unlike classifiers based on empirical risk minimization such as the linear classifiers of the paper or support vector machines (Cortes and Vapnik, 1995; Blanchard, Bousquet and Massart, 2008), the form of the classification boundary is not specified in advance. Nevertheless, the spirit of the bounds in the paper is that we should consider bounds over classifiers *of the same type*. We therefore propose to bound the empirical process over asymptotically optimally weighted nearest neighbor classifiers as the number of positive weights varies over a range of possible values. More precisely, we write $\hat{c}_k^{\mathrm{wnn}}$ for the weighted nearest neighbor classifier (2.1), with the weights given by (2.2) but with k replacing k^*. Then

$$\begin{aligned} n^{1/2}\{\hat{\tau}(\hat{c}^{\mathrm{wnn}}) - \tau(\hat{c}^{\mathrm{wnn}})\} &= \mathbb{G}_n \mathbb{1}_{\{Y \mathrm{sign}\{S_n(X) - 1/2\} < 0\}} \\ &\leq \sup_{k=k_0,\ldots,k_1} \mathbb{G}_n \mathbb{1}_{\{T_n(X) \leq 1/a_n\}} \mathbb{1}_{\{Y\hat{c}_k^{\mathrm{wnn}} < 0\}} + \mathbb{G}_n \mathbb{1}_{\{T_n(X) > 1/a_n\}} \mathbb{1}_{\{Y \mathrm{sign}\{S_n(X) - 1/2\} < 0\}} \\ &=: u(\mathbb{G}_n, \mathcal{T}, a_n). \end{aligned}$$

Similarly, for the lower bound,

$$\begin{aligned} n^{1/2}\{\hat{\tau}(\hat{c}^{\mathrm{wnn}}) - \tau(\hat{c}^{\mathrm{wnn}})\} \\ &\geq \inf_{k=k_0,\ldots,k_1} \mathbb{G}_n \mathbb{1}_{\{T_n(X) \leq 1/a_n\}} \mathbb{1}_{\{Y\hat{c}_k^{\mathrm{wnn}} < 0\}} + \mathbb{G}_n \mathbb{1}_{\{T_n(X) > 1/a_n\}} \mathbb{1}_{\{Y \mathrm{sign}\{S_n(X) - 1/2\} < 0\}} \\ &=: \ell(\mathbb{G}_n, \mathcal{T}, a_n). \end{aligned}$$

The minor problem with the standard bootstrap approximation caused by the repeated observations (which will typically lead to ties in computing distances) can be alleviated by subsampling – that is, sampling without replacement rather than with replacement. A subsample size of $\lfloor n/2 \rfloor$ mimics the bootstrap most closely (Freedman, 1977). Writing

$u_{1-\delta/2}$ for the $1-\delta/2$ quantile of the subsampling distribution of $u(\mathbb{G}_n, \mathcal{T}, a_n)$ and $\ell_{\delta/2}$ for the $\delta/2$ quantile of the subsampling distribution of $\ell(\mathbb{G}_n, \mathcal{T}, a_n)$, our analogue of the $100(1-\delta)\%$ Adaptive Confidence Interval is given by

$$\left[\mathbb{P}_n \mathbb{1}_{\{Y \text{sign}\{S_n(X)-1/2\}<0\}} - n^{-1/2} u_{1-\delta/2}\,,\, \mathbb{P}_n \mathbb{1}_{\{Y \text{sign}\{S_n(X)-1/2\}<0\}} - n^{-1/2} \ell_{\delta/2}\right].$$

3 Other classifiers and outlook

Generalizing our discussion from the previous section, it seems that the most important questions to be answered in attempting to provide an Adaptive Confidence Interval for the test error for other types of classifiers are the following:

(a) Can we devise a suitable hypothesis test to determine whether or not a point is close to the decision boundary of the Bayes classifier?

(b) Is there a suitable class of classifiers, of the same type as the original, which we can use to construct bounds on the empirical process of the test error on the set of points which are close to the Bayes decision boundary (in the sense that we do not reject the null hypothesis above)?

In order to be able to answer (a) in the affirmative, we need to have an understanding of the sampling properties of our classifier (at least asymptotically, and under the null hypothesis), since bootstrap tests do not seem obvious, and would add yet another layer of computational complexity. Such tests are available for the linear classifiers studied in the paper, the weighted nearest neighbor classifers in Section 2 above, or for classifiers based on kernel density estimates of the class conditional densities (Hall and Kang, 2005). However, for classifiers that require an iterative algorithm for computation, such as various empirical risk minimization methods or, for example, the smoothed log-concave classifiers of Chen and Samworth (2011), such a test seems less clear.

Regarding (b), again the answer appears to depend on the context. For instance, with the Hall and Kang (2005) kernel classifiers, if the bandwidth matrix is a scalar multiple $h^2 I$ of the identity matrix, then choosing a range of values of h to contruct the bounds on the empirical process would seem appropriate. For more complicated bandwidth matrices, though, this could lead to a very computationally expensive procedure. Typically, an affirmative answer to (b) seems more obvious for methods based on empirical risk minimization, where the class of decision boundaries is pre-specified.

In this ground-breaking paper, Laber and Murphy have established a paradigm for performing statistical inference in a difficult, non-regular problem of clear scientific importance. The desire to understand the extent to which the methodology can be extended and developed sets a clear research agenda for the future. I look forward to witnessing, and perhaps even contributing to, this development in the coming years.

References

Blanchard, G., Bousquet, O. and Massart, P. (2008) *Statistical performance of support vector machines. Ann. Statist.*, **36**, 489–531.

Chen, Y. and Samworth, R. J. (2011) Smoothed log-concave maximum likelihood estimation with applications. `http://arxiv.org/abs/1102.1191`

Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Mach. Learning*, **20**, 273–297.

Cover, T. M. and Hart, P. E. (1967) Nearest neighbor pattern classification. *IEEE Trans. Inf. Th.*, **13**, 21–27.

Devroye, L., Györfi, L. and Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition.* Springer-Verlag, New York.

Fix, E. and Hodges, J. L. (1951) Discriminatory analysis – nonparametric discrimination: Consistency properties. Tech. Rep. 4, Project no. 21-29-004, USAF School of Aviation Medicine, Randolph Field, Texas.

Freedman, D. (1977) A remark on the difference between sampling with and without replacement. *J. Am. Statist. Assoc.*, **72**, 681.

Guillemin, V. and Pollack, A. (1974) *Differential Topology* Prentice-Hall, New Jersey.

Hall, P. and Kang, K.-H. (2005) Bandwidth choice for nonparametric classification. *Ann. Statist.*, **33**, 284–306.

Hall, P., Park, B. U. and Samworth, R. J. (2008) Choice of neighbor order in nearest-neighbor classification. *Ann. Statist.*, **36**, 2135–2152.

Royall, R. (1966) *A class of Nonparametric Estimators of a Smooth Regression Function.* PhD Thesis, Stanford University, Stanford, CA.

Samworth, R. J. (2011) Optimal weighted nearest neighbour classifiers. http://arxiv.org/abs/1101.5783