Author feedback: “An adaptive nearest neighbor rule for classification"

Our thanks to all the reviewers for their time and constructive comments.

Reviewer 1

We thank the reviewer for the careful and detailed comments. It is clear that the reviewer has understood our paper well.

In terms of the comparison to existing work: Thank you for pointing out some other papers that suggest methods for setting k locally. We will add a discussion of these to our paper; the most relevant is probably Kpotufe (2011).

In a nutshell: (1) Even though classification is often reduced to regression in nonparametric analysis, methods of setting k locally should be rather different in the two settings, and this is reflected in the large difference between Kpotufe's setting and ours (for instance, the physical value of the radius containing k points actually matters in his setting but plays no role in ours); (2) what's more, the benefit of this local adaptivity is likely to be more pronounced in the case of classification. Our analysis, for instance, shows that in classification, there is a radius $r(x)$ around each point x such that prediction based on $B(x, r)$ for any $r \leq r(x)$ will w.h.p. be perfect, provided enough points fall in this ball. (Once n is large enough that there are enough points within this radius, you're done!) This is not, of course, true for regression, where the target y is a real value and thus the radius needs to keep shrinking.

Reviewer 2

The reviewer asks for clarification on how the parameter δ (i.e. A, in practice) should be set. We will clarify this point further in the paper, and discuss theoretical considerations in the response to Reviewer 3. Briefly, we find in the paper that in practice, increasing A lowers coverage and raises performance on the predicted set, but increases the neighborhood size required to predict rather than abstain. On the other hand, practical considerations typically imply that not too many neighbors can be used. So a practical A should be as small as possible to achieve a desired coverage, with a given maximum neighborhood size per point.

Reviewer 3

We thank the reviewer for the detailed comments and address some of the questions raised.

1. *Setting the parameter δ.* This is the standard confidence parameter of statistics and learning theory: it provides an upper bound on the failure probability of the algorithm. It can be set to 0.05, for instance. So it is not the case that we have replaced one parameter (k) with another (δ). Rather, our algorithm automatically makes infinitely many parameter choices (we pick a different k for each point) and asks for just a single failure probability that lets it know how aggressively to set its confidence intervals.

2. *Rates of convergence on single points vs the entire test distribution.* We provide both types of bounds: Theorem 1 gives a bound for individual points, whereas Theorem 2 provides the sort of uniform convergence bound that is more standard in learning theory.

3. *Instance optimality*. The reviewer is correct that claims of instance-optimality in the introduction are not substantiated later in the paper. This is indeed an omission, but it is not terribly complicated and we will add it to the paper.

In a nutshell: For any point x, the "advantage" $\mathrm{adv}(x)$ is equal to $p\gamma^2$, where there is a ball centered at x that has probability mass p and has average y-value that is either $\frac{1}{2} + \gamma$ or $\frac{1}{2} - \gamma$. Given only this information about x, in order to predict x's label correctly with constant probability, we need γ^{-2} points in the ball; thus we need $\Omega(1/(p\gamma^2)) = \Omega(1/\mathrm{adv}(x))$ data points overall.